Proceedings of 15^{th} Gökova
Geometry-Topology Conference
pp. 1 – 30

An introduction to reduced volume

Peng Lu

Abstract. In these lecture notes we give an introduction to Perelman's theory of reduced distance and reduced volume. We begin with a quick introduction to some classical results about Riemannian manifolds with nonnegative Ricci curvature, and we end with two applications of the monotonicity of reduced volume.

1. Introduction

In Riemannian geometry we study the geometric properties of Riemannian manifolds, such as the relation between curvature and topology, the existence of metrics satisfying certain curvature conditions, and properties of various other geometric quantities. Such a study often requires us to consider collections of Riemannian manifolds, one such example is the Cheeger-Gromov compactness theorem and its applications. A special case of collections of Riemannian manifolds is a one-parameter family of Riemannian metrics $g(s)$ on a given manifold M, this is used by S.T. Yau in his celebrated proof of the existence of Ricci flat Kähler metrics using the continuity method. Ricci flow, introduced by R. Hamilton, is a one-parameter family of metrics $\tilde{g}(t)$, $t \in (\alpha, T)$, which satisfies the flow equation $\frac{\partial}{\partial t}\tilde{g}(t) = -2\,\mathrm{Rc}(\tilde{g}(t))$ where $\mathrm{Rc}(\tilde{g}(t))$ stands for the Ricci curvature of $\tilde{g}(t)$. We hasten to add that Ricci flow is a very special one-parameter family of metrics. Through the work of G. Perelman on smooth Ricci flow we clearly see that there is an organic relation between the time variable t and the space variables in Ricci flow, which suggests some kind of space-time geometry. In these notes of a mini-course we give an introduction to two key notions which support the space-time viewpoint: reduced distance and reduced volume.

Because of the analogy between Ricci flow and the geometry of Riemannian manifolds with nonnegative Ricci curvature, we begin the lectures with a quick review of some classical global results which hold on any Riemannian manifolds with nonnegative Ricci curvature. The purpose of §2 is to help the reader to understand the results and calculations that appear later in §3 and §4 about the reduced distance and the reduced volume, respectively. In §5 we give two applications of reduced volume, in particular, the no local collapsing theorem.

Key words and phrases. Ricci flow, reduced distance, reduced volume, no local collapsing.

As always going back to the source (original article(s)) fundamentally helps one's understanding of math, for reduced distance and reduced volume we strongly suggest that the reader consults Perelman's paper [Pe02], in particular, §7, §8, and §6.

We assume that the reader has some exposure to Riemannian geometry and Ricci flow before.

We end this introduction with a **warning** and a list. In these notes we have swept under the rug the issue about how to do calculus at cut-locus points where distance function $r(\cdot)$ is not smooth. There are standard procedures to handle it: Calabi trick or barrier functions. In the calculation below we pretend that $r(\cdot)$ is smooth. Similar issue exists for the reduced distance (e.g., §3.2A) and the issue can be solved using barrier functions, again in the calculation below we pretend that the reduced distance is smooth.

Conventions and notations:
We adopt Einstein summation convention
$g(x,t)$: the inner product on the tangent space at point x defined by metric $g(t)$
R: scalar curvature
$R(x,t)$: scalar curvature of metric $g(t)$ at x
Rc: Ricci curvature. $R_{ij}=R_{ikkj}=R^k_{kij}$
$\mathrm{Rc}(x,t)$: Ricci curvature of metric $g(t)$ at x
Rm: Riemann curvature tensor; $\mathrm{Rm}\left(\frac{\partial}{\partial x^i},\frac{\partial}{\partial x^j}\right)\frac{\partial}{\partial x^k}=R^l_{ijk}\frac{\partial}{\partial x^l}$ and $R_{ijkl}=g_{lp}R^p_{ijk}$
$\mathrm{Rm}(x,t)$: Riemann curvature tensor of metric $g(t)$ at x
Distance function: $r(x)=d(p,x)$ for some point p
Lapalace-Beltrami operator: $\Delta=\frac{1}{\sqrt{|g|}}\frac{\partial}{\partial x^i}\left(\sqrt{|g|}g^{ij}\frac{\partial}{\partial x^j}\right)$ where $|g|=\det(g_{ij})$
Volume form: $d\mu_g=\sqrt{|g|}dx^1\cdots dx^n$

2. Riemannian manifolds with nonnegative Ricci curvature

Throughout this section (M^n,g) denotes a smooth n dimensional complete connected oriented Riemannian manifold with $\mathrm{Rc}\geq 0$. Mainly this section is about the distance function.

2.1. First variation of length

2.1A. Length. Given a smooth path $\gamma:[a,b]\to M$, its **length** is defined by

$$\mathrm{L}(\gamma)=\int_a^b|\dot\gamma(u)|du \tag{2.1}$$

where $|\dot\gamma(u)|$ is the length of the tangent vector $\dot\gamma(u)=\frac{d\gamma}{du}$. Given two points $x,y\in M$ the **distance** between them is defined by

$$d(x,y)=\inf_{\{\gamma:\gamma(a)=x,\gamma(b)=y\}}\mathrm{L}(\gamma)$$

where γ runs through smooth paths.

2.1B. First variation of length. Now we compute the **first variation of length**. Let $\gamma_s(u)$, $s \in (-\epsilon, \epsilon)$ be a smooth family of smooth paths with $\gamma_0 = \gamma$ parametrized by arc-length parameter u (i.e., $|\dot\gamma(u)| = 1$). Denote $Y = \frac{\partial \gamma_s}{\partial s}\Big|_{s=0}$ to be the variation vector field. Then using $\nabla_Y \dot\gamma_s - \nabla_{\dot\gamma_s} Y = [Y, \dot\gamma_s] = 0$ and integrating by parts we get

$$\begin{aligned}\frac{d}{ds}\Big|_{s=0} \mathrm{L}(\gamma_s) &= \int_a^b \frac{1}{2} \left\langle \dot\gamma_s(u), \dot\gamma_s(u)\right\rangle^{-1/2} \cdot Y \left\langle \dot\gamma_s, \dot\gamma_s\right\rangle \Big|_{s=0} du \\ &= \int_a^b \left\langle \dot\gamma, \nabla_{\dot\gamma} Y\right\rangle du = \int_a^b \left(\dot\gamma(\left\langle \dot\gamma, Y\right\rangle) - \left\langle \nabla_{\dot\gamma}\dot\gamma, Y\right\rangle\right) du \\ &= \left\langle \dot\gamma, Y\right\rangle (b) - \left\langle \dot\gamma, Y\right\rangle (a) - \int_a^b \left\langle \nabla_{\dot\gamma}\dot\gamma, Y\right\rangle du.\end{aligned}$$

Hence the critical point equation of the length functional L on the space of smooth paths with fixed end points is

$$\nabla_{\dot\gamma}\dot\gamma = 0. \tag{2.2}$$

This is the **equation of geodesics**. Any path that satisfies (2.2) is called a **geodesic**. In local coordinates $x = (x^i)$ the equation is the following system of ordinary differential equations (ode)

$$\frac{d^2x^i}{du^2} + \Gamma^i_{jk}(x(u))\frac{dx^j}{du}\cdot\frac{dx^k}{du} = 0, \tag{2.3}$$

where $\Gamma^i_{jk} = \frac{1}{2}g^{il}\left(\partial_j g_{lk} + \partial_k g_{jl} - \partial_l g_{jk}\right)$ is the **Christoffel symbol**.

2.1C. Jacobi field. Consider the set of all geodesics defined on $[a, b]$ and call it the moduli space of geodesics, the tangent directions of this space satisfy a second order linear ode. Let $\gamma_s(u)$, $s \in (-\epsilon, \epsilon)$ be a smooth family of geodesics with $\gamma_0 = \gamma$. Let $Y = \frac{\partial \gamma_s}{\partial s}\Big|_{s=0}$ be the variation vector field of γ_s. Taking the ∇_Y-derivative of the geodesic equation $\nabla_{\dot\gamma_s}\dot\gamma_s \equiv 0$, we have

$$0 = \nabla_Y \nabla_{\dot\gamma_s}\dot\gamma_s = \nabla_{\dot\gamma_s}\nabla_Y\dot\gamma_s + \mathrm{Rm}\left(Y, \dot\gamma_s\right)\dot\gamma_s.$$

Using $[Y, \dot\gamma_s] = 0$ and evaluating at $s = 0$, we get the **Jacobi equation**

$$\nabla_{\dot\gamma}\nabla_{\dot\gamma} Y + \mathrm{Rm}\left(Y, \dot\gamma\right)\dot\gamma = 0. \tag{2.4}$$

Any vector field Y along a geodesic γ which satisfies (2.4) is called a **Jacobi field**.

2.1D. Exponential map and Jacobian under geodesic spherical coordinates. Consider the initial value problem of second order ode (2.2) by choosing a point $p \in M$ (corresponding to the value of $\{x^i(0)\}$ in (2.3)) and a unit vector $V \in T_pM$ (corresponding to the value of $\{\frac{dx^i}{du}(0)\}$ in (2.3)). This defines the so-called **exponential map** $\exp_p(uV) = \gamma(u)$. There is a star-shaped open connected subset $0 \in \Omega_p \subset T_pM$ such that $\exp_p : \Omega_p \to M \setminus \mathrm{Cut}_p$ is a diffeomorphism. Here Cut_p is the so-called **cut locus** of p in M.

Let $r(x) = d(x,p)$ and let $(\theta^1, \cdots, \theta^{n-1}, r)$ be local spherical coordinates on $T_pM - \{0\}$. Then the inverse map $\exp_p^{-1} : M \setminus (\{p\} \cup \mathrm{Cut}_p) \to \Omega_p \setminus \{0\}$ defines the **geodesic spherical coordinates** of M. By Gauss lemma $g(\frac{\partial}{\partial r}, \frac{\partial}{\partial r}) = 1$ and $g(\frac{\partial}{\partial \theta^a}, \frac{\partial}{\partial r}) = 0$, $a = 1, \cdots, n-1$, and hence the metric can be written as

$$g = dr^2 + g_{ab} d\theta^a d\theta^b. \tag{2.5}$$

The volume form can be written as

$$d\mu = \sqrt{\det(g_{ab})}\, d\theta^1 \cdots d\theta^{n-1} dr.$$

$J = \sqrt{\det(g_{ab})}$ is called the **Jacobian of the exponential map**. The **area form** on the sphere $S(p, r_0)$ of radius r_0 and centered at p can be written as

$$d\sigma = J d\theta^1 \wedge \cdots d\theta^{n-1}. \tag{2.6}$$

Recall **Hessian** $\mathrm{Hess}\, f$ of a smooth function f on M is defined by $(\mathrm{Hess}\, f)(X,Y) = X(Yf) - (\nabla_X Y) f$ for any $X, Y \in T_xM$. It follows from (2.5) that $|\frac{\partial}{\partial r}| = 1$, $\nabla_{\frac{\partial}{\partial r}} \frac{\partial}{\partial r} = 0$, and $(\mathrm{Hess}\, r)(\frac{\partial}{\partial r}, \frac{\partial}{\partial r}) = 0$. Hence $|\mathrm{Hess}\, r|^2 \geq \frac{1}{n-1} (\Delta r)^2$.

2.1E. Mean curvature of spheres. We continue to use geodesic spherical coordinates $(\theta^1, \cdots, \theta^{n-1}, r)$. Let h denote the **second fundamental form** of $S(p, r_0)$. Note $\frac{\partial}{\partial r}$ is the unit outward normal vector field to $S(p, r_0)$. From equation (2.5) we have

$$\begin{aligned} h_{ab} &= h\left(\frac{\partial}{\partial \theta^a}, \frac{\partial}{\partial \theta^b}\right) = \left\langle \nabla_{\frac{\partial}{\partial \theta^a}} \frac{\partial}{\partial r}, \frac{\partial}{\partial \theta^b} \right\rangle \\ &= -\left\langle \frac{\partial}{\partial r}, \nabla_{\frac{\partial}{\partial \theta^a}} \frac{\partial}{\partial \theta^b} \right\rangle = -\Gamma^n_{ab} = \frac{1}{2} \frac{\partial}{\partial r} g_{ab}. \end{aligned}$$

Hence the **mean curvature** H of $S(p, r_0)$ is given by

$$H = g^{ab} h_{ab} = \frac{1}{2} g^{ab} \frac{\partial}{\partial r} g_{ab} = \frac{\partial}{\partial r} \log \sqrt{\det(g_{ab})} = \frac{\partial}{\partial r} \log \mathrm{J}. \tag{2.7}$$

On the other hand the mean curvature H can be computed as

$$H = \left\langle \nabla_{\frac{\partial}{\partial \theta_a}} \frac{\partial}{\partial r}, \frac{\partial}{\partial \theta_a} \right\rangle = \left\langle \nabla_{\frac{\partial}{\partial \theta_a}} \frac{\partial}{\partial r}, \frac{\partial}{\partial \theta_a} \right\rangle + \left\langle \nabla_{\frac{\partial}{\partial r}} \frac{\partial}{\partial r}, \frac{\partial}{\partial r} \right\rangle = \Delta r. \tag{2.8}$$

2.2. Bochner formula for functions and its consequence

2.2A. Bochner formula. Let f be a smooth function on M. We compute the commutator $[\Delta, \nabla_j] f$ as

$$\Delta \nabla_j f = \nabla_i \nabla_i \nabla_j f = \nabla_i \nabla_j \nabla_i f = \nabla_j \nabla_i \nabla_i f - R_{ijil} \nabla_l f = \nabla_j \Delta f + R_{jl} \nabla_l f. \tag{2.9}$$

We also compute

$$\Delta |\nabla f|^2 = 2 \nabla_i (\nabla_j f \cdot \nabla_i \nabla_j f) = 2 \nabla_i \nabla_j f \cdot \nabla_i \nabla_j f + 2 \nabla_j f \cdot \nabla_i \nabla_i \nabla_j f.$$

Hence we get the **Bochner formula** for f

$$\Delta |\nabla f|^2 = 2 |\nabla \nabla f|^2 + 2 R_{ij} \nabla_i f \nabla_j f + 2 \nabla_i f \nabla_i (\Delta f). \tag{2.10}$$

2.2B. Laplacian comparison theorem. In (2.10) let $f(x) = r(x)$ be the distance function. Using $\mathrm{Rc} \geq 0$, $|\nabla r| = 1$ and $|\mathrm{Hess}\, r|^2 \geq \frac{1}{n-1}(\Delta r)^2$ we get

$$\frac{\partial}{\partial r}(\Delta r) + \frac{1}{n-1}(\Delta r)^2 \leq 0.$$

Since $\Delta r \to \frac{n-1}{r}$ as $r \to 0^+$ and $\frac{\partial}{\partial r}(\frac{n-1}{r}) + \frac{1}{n-1}(\frac{n-1}{r})^2 = 0$, from the differential inequality above and the ode comparison property we conclude the **Laplacian comparison theorem** on complete manifolds with $\mathrm{Rc} \geq 0$

$$\Delta r \leq \frac{n-1}{r}. \tag{2.11}$$

2.2C. Bishop-Gromov volume comparison theorem. Let $A(r_0)$ be the **area** of sphere $S(p, r_0)$. Then by (2.6), (2.7), (2.8), and (2.11) we have

$$\frac{d}{dr}A(r) = \int_{S^{n-1}} \frac{dJ}{dr} d\theta^1 \wedge \cdots d\theta^{n-1} = \int_{S(p,r)} HJ d\theta^1 \wedge \cdots d\theta^{n-1} \leq \frac{n-1}{r}A(r).$$

Integrating this differential inequality, we see that

$$A(r_2)\, r_1^{n-1} \leq A(r_1)\, r_2^{n-1} \qquad \text{for } r_2 \geq r_1 > 0.$$

Integrating again we get

$$n\omega_n \int_0^{r_2} dr_2 \int_0^{r_1} A(r_2)\, r_1^{n-1} dr_1 \leq n\omega_n \int_0^{r_2} dr_2 \int_0^{r_1} A(r_1)\, r_2^{n-1} dr_1,$$

where ω_n is the volume of unit Euclidean ball.

Let $B(p, r_0)$ be the ball of radius r_0 and centered at p, and let $\mathrm{Vol}\, B(p, r_0)$ denote the volume of the ball. Using $\mathrm{Vol}\, B(p, r_0) = \int_0^{r_0} A(r)\, dr$ the inequality above can be written as

$$\frac{\mathrm{Vol}\, B(p, r_2)}{\omega_n r_2^n} \leq \frac{\mathrm{Vol}\, B(p, r_1)}{\omega_n r_1^n} \qquad \text{for } r_2 \geq r_1 > 0, \tag{2.12}$$

i.e., $\frac{\mathrm{Vol}\, B(p,r)}{\omega_n r^n}$ is a monotone decreasing function (compare with (4.7)). (2.12) is the **Bishop-Gromov volume comparison theorem** for complete manifolds with $\mathrm{Rc} \geq 0$.

2.3. Differential Harnack inequality

2.3A. Li-Yau Harnack estimate. Now we discuss Li-Yau differential Harnack estimate for positive solutions of heat equation $\frac{\partial u}{\partial t} = \Delta u$ defined on (M^n, g). Let $L = \log u$ and $Q = \Delta L$. We compute the evolution equation of L as

$$\frac{\partial L}{\partial t} = \frac{1}{u} \triangle u = \triangle L + |\nabla L|^2.$$

We compute the evolution equation of Q as

$$\begin{aligned}\frac{\partial Q}{\partial t} &= \triangle\left(\frac{\partial}{\partial t}L\right) = \triangle\left(\triangle L + |\nabla L|^2\right)\\ &= \triangle Q + 2\left\langle \triangle\nabla L, \nabla L\right\rangle + 2\left|\nabla\nabla L\right|^2\\ &= \triangle Q + 2\left\langle \nabla\triangle L, \nabla L\right\rangle + 2\,\mathrm{Rc}\left(\nabla L, \nabla L\right) + 2\left|\nabla\nabla L\right|^2,\end{aligned}$$

where we have used (2.9) to get the last equality. Hence, from $\mathrm{Rc} \geq 0$ and $|\nabla\nabla L|^2 \geq \frac{1}{n}(\Delta L)^2$ we have

$$\frac{\partial Q}{\partial t} \geq \triangle Q + 2\left\langle \nabla L, \nabla Q\right\rangle + \frac{2}{n}Q^2.$$

From this inequality and the maximum principle we may deduce the following **Li-Yau differential Harnack estimate** for positive solutions of heat equation on complete manifolds with $\mathrm{Rc} \geq 0$

$$Q = \Delta \log u \geq -\frac{n}{2t}. \tag{2.13}$$

Note that the equality above is satisfied by the heat kernel $u = \frac{1}{(4\pi t)^{n/2}} e^{-\frac{|x|^2}{4t}}$ on Euclidean space $\mathbb{R}^n$.

2.3B. Harnack estimate for heat kernel. The following estimate is proved by Lei Ni in 2004 (J. Geom. Anal. **14**, 87–100). Let $H = \frac{1}{(4\pi t)^{n/2}} e^{-f}$ be a heat kernel of (M^n, g) with $\mathrm{Rc} \geq 0$. Then for $t > 0$

$$t(2\Delta f - |\nabla f|^2) + f - n \leq 0. \tag{2.14}$$

This estimate is closely related to Perelman's Harnack inequality ([Pe02], Corollary 9.3).

2.4. Second variation of length

2.4A. Second variation formula. Let $\gamma_s(u)$, $u \in [a, b]$, $s \in (-\epsilon, \epsilon)$ be a smooth family of paths. Assume that $\gamma_0(u) = \gamma(u)$ is a geodesic of unit speed (i.e., $|\dot\gamma(u)| = 1$). Denote $Y = \left.\frac{\partial \gamma_s}{\partial s}\right|_{s=0}$ to be the variation field. We compute

$$\begin{aligned}&\left.\frac{d^2}{ds^2}\right|_{s=0} \mathrm{L}(\gamma_s) = \left.\frac{d}{ds}\right|_{s=0} \int_a^b |\dot\gamma_s(u)|^{-1} \cdot \left\langle \dot\gamma_s, \nabla_Y \dot\gamma_s\right\rangle du\\ &= \int_a^b \Big(-\left\langle \dot\gamma, \nabla_Y \dot\gamma\right\rangle^2 + \left\langle \nabla_Y \dot\gamma, \nabla_{\dot\gamma} Y\right\rangle + \left\langle \dot\gamma, \nabla_Y \nabla_{\dot\gamma} Y\right\rangle \Big)\, du\\ &= \int_a^b \Big(\left|\nabla_{\dot\gamma} Y\right|^2 - \left\langle \nabla_{\dot\gamma} Y, \dot\gamma\right\rangle^2 + \left\langle \dot\gamma, \mathrm{Rm}\left(Y, \dot\gamma\right) Y\right\rangle + \left\langle \dot\gamma, \nabla_{\dot\gamma}\nabla_Y Y\right\rangle \Big)\, du.\end{aligned}$$

Hence the **second variation of length** at a unit speed geodesic γ is given by

$$\begin{aligned}&\frac{d^2}{ds^2}\Big|_{s=0} \mathrm{L}(\gamma_s)\\ &= \int_a^b \left(|\nabla_{\dot\gamma} Y|^2 - \langle \nabla_{\dot\gamma} Y, \dot\gamma\rangle^2 - \langle \mathrm{Rm}(Y,\dot\gamma)\,\dot\gamma, Y\rangle \right) ds + \langle \nabla_Y Y, \dot\gamma\rangle\big|_a^b . \end{aligned} \tag{2.15}$$

2.4B. Index form. Let $\gamma : [a,b] \to M$ be a geodesic. The **index form** is defined by

$$I(V,W) = \int_a^b \left(\langle \nabla_{\dot\gamma} V, \nabla_{\dot\gamma} W\rangle - \langle \mathrm{Rm}(V,\dot\gamma)\dot\gamma, W\rangle \right) du, \tag{2.16}$$

where V and W are vector fields along γ vanishing at $\gamma(a)$, $\gamma(b)$ and perpendicular to $\dot\gamma$. Note that $\langle \nabla_{\dot\gamma} Y, \dot\gamma\rangle = 0$ when Y in (2.15) is perpendicular to $\dot\gamma$, hence as a quadratic form of Y the right side of (2.15) gives rise to the bilinear form (2.16).

The **Index lemma** says the following. Suppose there is not any pair of conjugate points on geodesic $\gamma(u)$, $u \in [a,b]$. Given $A \in T_{\gamma(a)}M, B \in T_{\gamma(b)}M$ with $\langle A, \dot\gamma(a)\rangle = \langle B, \dot\gamma(b)\rangle = 0$, the unique Jacobi field J along γ with $J(a) = A$, $J(b) = B$, satisfies

$$I(J,J) \le I(W,W). \tag{2.17}$$

Here W is any the vector field along γ which is perpendicular to $\dot\gamma$ and satisfies $W(a) = A$, $W(b) = B$.

2.5. Some other results

There are other theorems that hold for all complete manifolds with $\mathrm{Rc} \ge 0$. Here we give one example. Recall that a geodesic $\gamma : \mathbb{R} \to (M,g)$ is called a **line** if $\mathrm{L}(\gamma|_{[a,b]}) = d(\gamma(a),\gamma(b))$. The **Cheeger-Gromoll splitting theorem** says the following. Let (M^n, g) be a complete Riemannian manifold with $\mathrm{Rc} \ge 0$. Suppose there is a line in M, then (M,g) is isometric to $\mathbb{R} \times \left(N^{n-1}, h\right)$ with the product metric. Here (N,h) is a complete Riemannian manifold with $\mathrm{Rc} \ge 0$. The proof uses **Busemann** function, an important function in the study of noncompact Riemannian manifolds.

3. The reduced distance

The reduced distance and reduced volume are formally introduced by Perelman in [Pe02], §7. The motivation he gives in §6 is both interesting and mysterious. From the work of Perelman and others it is evident that the reduced distance and reduced volume are fundamental tools in Ricci flow. It is desirable to find more applications of them.

In this section we establish various identities and inequalities about reduced distance. Some of these inequalities will be used in next section to show the finiteness and monotonicity of reduced volume. Since formulae in this section come out of relatively lengthy calculation, here we only provide glimpses of these calculation, readers can either figure out the detail themselves or find the detailed calculation in the literature.

Throughout this section N^n is a n-dimensional connected oriented manifold, and $(N^n, g(\tau))$, $\tau \in [0,T]$, is a solution to the **backward Ricci flow** $\frac{\partial}{\partial \tau} g(\tau) = 2\,\mathrm{Rc}(g(\tau))$ with bounded Riemann curvature $\sup_{M\times[0,T]} |\,\mathrm{Rm}(x,\tau)| < \infty$. We assume that $g(\tau)$ is complete for each $\tau \in [0,T]$ (called **complete solution**). Below, the notation ∇, Δ, R, Rc, Rm stand for connection, Lapalace-Beltrami operator and curvatures defined by $g(\tau)$.

3.1. First variation formula of $\mathcal{L}$-length

3.1A. Definition of reduced distance. Let $\gamma : [0,\tau] \to N$ be a smooth path with $\tau \le T$. The **$\mathcal{L}$-length** of γ is defined by

$$\mathcal{L}(\gamma) = \int_0^\tau \sqrt{\tilde{\tau}} \left(R(\gamma(\tilde{\tau}), \tilde{\tau}) + \left| \frac{d\gamma}{d\tilde{\tau}}(\tilde{\tau}) \right|^2_{g(\tilde{\tau})} \right) d\tilde{\tau}. \tag{3.1}$$

Fix a point $p \in N$, the **L-distance** from $(p,0)$ to $(x,\tau) \in N \times (0,T]$ is defined by

$$L(x,\tau) = \inf_{\{\gamma:\gamma(0)=p,\gamma(\tau)=x\}} \mathcal{L}(\gamma).$$

We call $(p,0)$ the **basepoint**. A minimizing path in the definition of L-distance is called a **minimal L-geodesic**. The **reduced distance** is defined by

$$\ell(x,\tau) = \frac{1}{2\sqrt{\tau}} L(x,\tau).$$

3.1B. First variation of $\mathcal{L}$-length. Let $\gamma_s(\tilde{\tau})$, $\tilde{\tau} \in [0,\tau]$, $s \in (-\epsilon,\epsilon)$, be a smooth family of smooth paths with $\gamma_0 = \gamma$. Let $X = \dot{\gamma}$, and let $Y = \left.\frac{\partial \gamma_s}{\partial s}\right|_{s=0}$ be the variation field. The **first variation formula** for $\mathcal{L}$-length is given by

$$\begin{aligned} &\frac{1}{2}\frac{d}{ds}\bigg|_{s=0} \mathcal{L}(\gamma_s) \\ &= \sqrt{\tilde{\tau}} \langle Y, X\rangle \Big|_0^\tau + \int_0^\tau \sqrt{\tilde{\tau}} Y \cdot \left(\frac{1}{2}\nabla R - \frac{1}{2\tilde{\tau}} X - \nabla_X X - 2\,\mathrm{Rc}(X) \right) d\tilde{\tau}, \end{aligned} \tag{3.2}$$

where the covariant derivative $\nabla = \nabla_{g(\tilde{\tau})}$.

If γ is a critical point of the $\mathcal{L}$-length functional (3.1) among smooth paths with fixed endpoints, then γ is called an **$\mathcal{L}$-geodesic**. From (3.2) we get the **$\mathcal{L}$-geodesic equation**:

$$\nabla_X X - \frac{1}{2}\nabla R + 2\,\mathrm{Rc}(X) + \frac{1}{2\tilde{\tau}} X = 0. \tag{3.3}$$

3.1C. Calculating (3.2). We compute in a way similar to the deduction of the first variation formula for length (§2.1B).

$$\begin{aligned} \frac{d}{ds}\mathcal{L}(\gamma_s) &= \int_0^\tau \sqrt{\tilde{\tau}} \left(\frac{\partial}{\partial s} R(\gamma_s(\tilde{\tau}), \tilde{\tau}) + \frac{\partial}{\partial s} \left| \frac{\partial \gamma_s}{\partial \tilde{\tau}}(\tilde{\tau}) \right|^2_{g(\tilde{\tau})} \right) d\tilde{\tau} \\ &= \int_0^\tau \sqrt{\tilde{\tau}} \left(\langle \nabla R, Y\rangle + 2\langle \nabla_Y X, X\rangle \right) d\tilde{\tau}. \end{aligned} \tag{3.4}$$

Using

$$\langle \nabla_Y X, X\rangle = \langle \nabla_X Y, X\rangle = \frac{d}{d\tilde{\tau}}\left[g\left(Y,X\right)\right] - \langle Y, \nabla_X X\rangle - 2\,\mathrm{Rc}\left(Y,X\right),$$

we get

$$\frac{1}{2}\frac{d}{ds}\mathcal{L}\left(\gamma_s\right) = \int_0^\tau \sqrt{\tilde{\tau}}\left(\frac{1}{2}\left\langle \nabla R, Y\right\rangle + \frac{d}{d\tilde{\tau}}\left\langle Y, X\right\rangle - \left\langle Y, \nabla_X X\right\rangle - 2\,\mathrm{Rc}\left(Y,X\right)\right) d\tilde{\tau}.$$

(3.2) follows from integration by parts

$$\int_0^\tau \sqrt{\tilde{\tau}}\frac{d}{d\tilde{\tau}}\left\langle Y, X\right\rangle d\tilde{\tau} = -\frac{1}{2}\int_0^\tau \frac{1}{\sqrt{\tilde{\tau}}}\left\langle Y, X\right\rangle d\tilde{\tau} + \sqrt{\tilde{\tau}}\left\langle Y, X\right\rangle\Big|_0^\tau.$$

3.1D. Other form of (3.1) and (3.3). Sometimes we need to use the following parametrization:

$$\tilde{\sigma} = 2\sqrt{\tilde{\tau}} \ \text{ and } \ \beta\left(\tilde{\sigma}\right) = \gamma\left(\tilde{\sigma}^2/4\right). \tag{3.5}$$

Then we may rewrite $\mathcal{L}$-length as

$$\mathcal{L}\left(\gamma\right) = \int_0^{\sigma=2\sqrt{\tau}}\left(\frac{\tilde{\sigma}^2}{4}R\left(\beta\left(\tilde{\sigma}\right), \tilde{\sigma}^2/4\right) + \left|\frac{d\beta}{d\tilde{\sigma}}\left(\tilde{\sigma}\right)\right|^2_{g(\tilde{\sigma}^2/4)}\right) d\tilde{\sigma}, \tag{3.6}$$

and $\mathcal{L}$-geodesic equation as

$$\nabla_Z Z - \frac{\tilde{\sigma}^2}{8}\nabla R + \tilde{\sigma}\,\mathrm{Rc}\left(Z\right) = 0, \tag{3.7}$$

where $Z\left(\tilde{\sigma}\right) = \frac{d\beta(\tilde{\sigma})}{d\tilde{\sigma}} = \sqrt{\tilde{\tau}}X(\tilde{\tau})$. A simple consequence of (3.7) is that solutions to the initial value problem for β exist. Consequently the minimal $\mathcal{L}$-geodesic between any two points exists.

3.2. First order derivatives of L-distance

In this subsection we give some consequences of the first variation formula of $\mathcal{L}$-length. Below L-distance is defined using the basepoint $(p,0)$.

3.2A. Spatial derivative ∇L. Given $Y \in T_xN$, by choosing a family of minimal $\mathcal{L}$-geodesics $\gamma_s(\tilde{\tau})$, $\tilde{\tau} \in [0,\tau]$, $s \in (-\epsilon,\epsilon)$ such that $\gamma_s(0) = p$, $\frac{d}{ds}\big|_{s=0}\gamma_s(\tau) = Y$, and $\mathcal{L}(\gamma_s) = L(\gamma_s(\tau),\tau)$, we get from the first variation formula (3.2)

$$\left\langle \nabla L\left(x,\tau\right), Y\right\rangle = \left.\frac{d}{ds}\right|_{s=0}\mathcal{L}(\gamma_s) = \left\langle 2\sqrt{\tau}X\left(\tau\right), Y\right\rangle.$$

Hence the spatial derivative of the L-distance function

$$\nabla L\left(x,\tau\right) = 2\sqrt{\tau}X\left(\tau\right), \tag{3.8}$$

where X is the tangent vector field of the minimal $\mathcal{L}$-geodesic $\gamma = \gamma_0$ from $(p,0)$ to (x,τ).

3.2B. Time derivative $\frac{\partial L}{\partial \tau}$. We compute, using the chain rule and (3.8),

$$\begin{aligned}\frac{\partial L}{\partial \tau}(x,\tau) &= \frac{\partial L(\gamma(\tau),\tau)}{\partial \tau} = \frac{d}{d\tau}\left(L(\gamma(\tau),\tau)\right) - \nabla L(x,\tau)\cdot X(\tau)\\ &= \frac{d}{d\tau}\left[\int_0^\tau \sqrt{\tilde{\tau}}\left(R(\gamma(\tilde{\tau}),\tilde{\tau}) + \left|\frac{d\gamma}{d\tilde{\tau}}(\tilde{\tau})\right|^2\right)d\tilde{\tau}\right] - 2\sqrt{\tau}\,|X(\tau)|^2\\ &= \sqrt{\tau}\left(R(x,\tau) + |X(\tau)|^2\right) - 2\sqrt{\tau}\,|X(\tau)|^2.\end{aligned}$$

Hence the time-derivative of the L-distance function

$$\frac{\partial L}{\partial \tau}(x,\tau) = -\sqrt{\tau}\left(R(x,\tau) + |X(\tau)|^2\right) + 2\sqrt{\tau}R(x,\tau), \tag{3.9}$$

where X is the tangent vector field of the minimal $\mathcal{L}$-geodesic γ from $(p,0)$ to (x,τ).

3.3. Second variation formula of $\mathcal{L}$-length

For the purpose of getting information about the second order derivatives of L-distance, in this subsection we compute the second variation of $\mathcal{L}$-length.

3.3A. Second variation formula of $\mathcal{L}$-length. Using (3.4) and the notation in calculating the first variation formula we have

$$\left.\frac{d^2\mathcal{L}(\gamma_s)}{ds^2}\right|_{s=0} = \int_0^\tau \sqrt{\tilde{\tau}}\left(Y(Y(R)) + 2\langle\nabla_Y\nabla_Y X, X\rangle + 2|\nabla_Y X|^2\right)d\tilde{\tau}.$$

From

$$\langle\nabla_Y\nabla_Y X, X\rangle = \langle\nabla_Y\nabla_X Y, X\rangle = \langle\mathrm{Rm}(Y,X)Y, X\rangle + \langle\nabla_X\nabla_Y Y, X\rangle,$$

it follows

$$\begin{aligned}&\left.\frac{d^2\mathcal{L}(\gamma_s)}{ds^2}\right|_{s=0}\\ &= \int_0^\tau \sqrt{\tilde{\tau}}\left(Y(Y(R)) + 2\langle\mathrm{Rm}(Y,X)Y, X\rangle + 2\langle\nabla_X\nabla_Y Y, X\rangle + 2|\nabla_Y X|^2\right)d\tilde{\tau}.\end{aligned}$$

To get a better form for the term $2\langle\nabla_X\nabla_Y Y, X\rangle$, we have

$$\frac{d}{d\tilde{\tau}}\langle\nabla_Y Y, X\rangle = \langle\nabla_X\nabla_Y Y, X\rangle + \langle\nabla_Y Y, \nabla_X X\rangle + \frac{\partial g}{\partial\tilde{\tau}}(\nabla_Y Y, X) + \left\langle\left(\frac{\partial}{\partial\tilde{\tau}}\nabla\right)_Y Y, X\right\rangle,$$

this is because both the inner product and the connection in $\langle\nabla_Y Y, X\rangle$ depend on $\tilde{\tau}$. Using

$$\left\langle\left(\frac{\partial}{\partial\tilde{\tau}}\nabla\right)_Y Y, X\right\rangle = 2(\nabla_Y\,\mathrm{Rc})(Y,X) - (\nabla_X\,\mathrm{Rc})(Y,Y)$$

(derived from the formula for Christoffel symbols), we get

$$\begin{aligned}\frac{d}{d\tilde{\tau}}\langle\nabla_Y Y, X\rangle =& \langle\nabla_X\nabla_Y Y, X\rangle + \langle\nabla_Y Y, \nabla_X X\rangle + 2\,\mathrm{Rc}(\nabla_Y Y, X)\\ &+ 2(\nabla_Y\,\mathrm{Rc})(Y,X) - (\nabla_X\,\mathrm{Rc})(Y,Y).\end{aligned} \tag{3.10}$$

Hence

$$\begin{aligned}\left.\frac{d^2\mathcal{L}(\gamma_s)}{ds^2}\right|_{s=0} &= \int_0^\tau \sqrt{\tilde{\tau}}\left(Y(Y(R)) + 2\langle \mathrm{Rm}(Y,X)Y,X\rangle + 2|\nabla_Y X|^2\right)d\tilde{\tau} \\ &\quad + 2\int_0^\tau \sqrt{\tilde{\tau}}\begin{pmatrix} \frac{d}{d\tilde{\tau}}\langle \nabla_Y Y, X\rangle - \langle \nabla_Y Y, \nabla_X X\rangle - 2\,\mathrm{Rc}(\nabla_Y Y, X) \\ -2(\nabla_Y \mathrm{Rc})(Y,X) + (\nabla_X \mathrm{Rc})(Y,Y)\end{pmatrix} d\tilde{\tau} \\ &= \int_0^\tau \sqrt{\tilde{\tau}}\left(Y(Y(R)) + 2\langle \mathrm{Rm}(Y,X)Y,X\rangle + 2|\nabla_Y X|^2\right)d\tilde{\tau} \\ &\quad + 2\int_0^\tau \sqrt{\tilde{\tau}}\begin{pmatrix} -\langle \nabla_Y Y, \nabla_X X\rangle - 2\,\mathrm{Rc}(\nabla_Y Y, X) \\ -2(\nabla_Y \mathrm{Rc})(Y,X) + (\nabla_X \mathrm{Rc})(Y,Y)\end{pmatrix} d\tilde{\tau} \\ &\quad + 2\sqrt{\tilde{\tau}}\langle \nabla_Y Y, X\rangle\Big|_0^\tau - \int_0^\tau \frac{1}{\sqrt{\tilde{\tau}}}\langle \nabla_Y Y, X\rangle\, d\tilde{\tau}.\end{aligned}$$

Assume γ_0 is an $\mathcal{L}$-geodesic and

$$Y(0) = 0, \tag{3.11}$$

we compute using integration by parts,

$$\begin{aligned}\left.\frac{d^2\mathcal{L}(\gamma_s)}{ds^2}\right|_{s=0} &= 2\sqrt{\tau}\langle \nabla_Y Y, X\rangle + \int_0^\tau \sqrt{\tilde{\tau}}\begin{pmatrix} Y(Y(R)) - \nabla_Y Y\cdot\nabla R \\ +2\langle \mathrm{Rm}(Y,X)Y,X\rangle + 2|\nabla_Y X|^2\end{pmatrix} d\tilde{\tau} \\ &\quad + 2\int_0^\tau \sqrt{\tilde{\tau}}\begin{pmatrix} -\left\langle \nabla_Y Y, \left[\nabla_X X + 2\,\mathrm{Rc}(X) - \frac{1}{2}\nabla R + \frac{1}{2\tilde{\tau}}X\right]\right\rangle \\ -2(\nabla_Y \mathrm{Rc})(Y,X) + (\nabla_X \mathrm{Rc})(Y,Y)\end{pmatrix} d\tilde{\tau} \\ &= 2\sqrt{\tau}\langle \nabla_Y Y, X\rangle + \int_0^\tau \sqrt{\tilde{\tau}}\left(\nabla^2_{Y,Y}R + 2\langle \mathrm{Rm}(Y,X)Y,X\rangle + 2|\nabla_Y X|^2\right)d\tilde{\tau} \\ &\quad + \int_0^\tau \sqrt{\tilde{\tau}}\left(-4(\nabla_Y \mathrm{Rc})(Y,X) + 2(\nabla_X \mathrm{Rc})(Y,Y)\right)d\tilde{\tau},\end{aligned}$$

where $\nabla^2_{Y,Y}R$ denotes the Hessian $(\mathrm{Hess}\, R)(Y,Y)$.

We have derived formula (7.7) in [Pe02].

Lemma 3.1. *Let $\gamma : [0,\tau] \to N$ be an $\mathcal{L}$-geodesic and let γ_s be a smooth variation of $\gamma = \gamma_0$. Assume the variation field $Y = \frac{\partial}{\partial s}\gamma_s$ satisfies $Y(0) = 0$. The second variation of $\mathcal{L}$-length is given by*

$$\begin{aligned}\left.\frac{d^2\mathcal{L}(\gamma_s)}{ds^2}\right|_{s=0} &= 2\sqrt{\tau}\langle \nabla_Y Y, X\rangle(\tau) \\ &+ \int_0^\tau \sqrt{\tilde{\tau}}\begin{pmatrix} \nabla^2_{Y,Y}R + 2\langle \mathrm{Rm}(Y,X)Y,X\rangle + 2|\nabla_Y X|^2 \\ -4(\nabla_Y \mathrm{Rc})(Y,X) + 2(\nabla_X \mathrm{Rc})(Y,Y)\end{pmatrix} d\tilde{\tau}.\end{aligned} \tag{3.12}$$

3.3B. Another form of the second variation formula of $\mathcal{L}$-length. To write the second variation formula of $\mathcal{L}$-length in a better form (relating to Hamilton's matrix

Harnack quantity), we need to introduce a term $\left(\frac{\partial}{\partial\tilde{\tau}}\operatorname{Rc}\right)(Y,Y)$ into (3.12). Note that $\operatorname{Rc}(Y,Y)=\operatorname{Rc}(\gamma(\tilde{\tau}),\tilde{\tau})(Y(\tilde{\tau}),Y(\tilde{\tau}))$, we have

$$\frac{d}{d\tilde{\tau}}\left[\operatorname{Rc}(Y,Y)\right]=\left(\frac{\partial}{\partial\tilde{\tau}}\operatorname{Rc}\right)(Y,Y)+\left(\nabla_X\operatorname{Rc}\right)(Y,Y)+2\operatorname{Rc}\left(\nabla_X Y,Y\right).$$

It follows

$$\begin{aligned}&-\int_0^\tau\sqrt{\tilde{\tau}}\left(\frac{\partial}{\partial\tilde{\tau}}\operatorname{Rc}\right)(Y,Y)\,d\tilde{\tau}=-\sqrt{\tilde{\tau}}\operatorname{Rc}(Y,Y)\Big|_0^\tau+\\&+\int_0^\tau\sqrt{\tilde{\tau}}\left(\frac{1}{2\tilde{\tau}}\operatorname{Rc}(Y,Y)+\left(\nabla_X\operatorname{Rc}\right)(Y,Y)+2\operatorname{Rc}\left(\nabla_X Y,Y\right)\right)d\tilde{\tau}.\end{aligned}$$

Hence we can rewrite (3.12) as

$$\begin{aligned}&\left.\frac{d^2\mathcal{L}(\gamma_s)}{ds^2}\right|_{s=0}-2\sqrt{\tau}\left\langle\nabla_Y Y,X\right\rangle+2\sqrt{\tau}\operatorname{Rc}(Y,Y)\\=&\int_0^\tau\sqrt{\tilde{\tau}}\left(\left(2\frac{\partial}{\partial\tilde{\tau}}\operatorname{Rc}+\frac{1}{\tilde{\tau}}\operatorname{Rc}\right)(Y,Y)+\nabla^2_{Y,Y}R-2\left|\operatorname{Rc}(Y)\right|^2\right)d\tilde{\tau}\\&+\int_0^\tau\sqrt{\tilde{\tau}}\left(2\left\langle\operatorname{Rm}(Y,X)Y,X\right\rangle-4\left(\nabla_Y\operatorname{Rc}\right)(Y,X)+4\left(\nabla_X\operatorname{Rc}\right)(Y,Y)\right)d\tilde{\tau}\\&+\int_0^\tau 2\sqrt{\tilde{\tau}}\left|\nabla_X Y+\operatorname{Rc}(Y)\right|^2d\tilde{\tau}.\end{aligned}$$

In the above formula by substituting Hamilton's **matrix Harnack quantity**

$$\begin{aligned}H(X,Y)=&-2\left(\frac{\partial}{\partial\tilde{\tau}}\operatorname{Rc}\right)(Y,Y)-\nabla^2_{Y,Y}R+2\left|\operatorname{Rc}(Y)\right|^2-\frac{1}{\tilde{\tau}}\operatorname{Rc}(Y,Y)\\&-2\left\langle\operatorname{Rm}(Y,X)Y,X\right\rangle-4\left(\nabla_X\operatorname{Rc}\right)(Y,Y)+4\left(\nabla_Y\operatorname{Rc}\right)(Y,X),\end{aligned}\tag{3.13}$$

we obtain

$$\begin{aligned}&\left.\frac{d^2\mathcal{L}(\gamma_s)}{ds^2}\right|_{s=0}-2\sqrt{\tau}\left\langle\nabla_Y Y,X\right\rangle+2\sqrt{\tau}\operatorname{Rc}(Y,Y)\\=&-\int_0^\tau\sqrt{\tilde{\tau}}H(X,Y)\,d\tilde{\tau}+\int_0^\tau 2\sqrt{\tilde{\tau}}\left|\nabla_X Y+\operatorname{Rc}(Y)\right|^2d\tilde{\tau}.\end{aligned}\tag{3.14}$$

By a little calculation we have

$$\int_0^\tau 2\sqrt{\tilde{\tau}}\left|\nabla_X Y+\operatorname{Rc}(Y)\right|^2d\tilde{\tau}=\int_0^\tau 2\sqrt{\tilde{\tau}}\left|\nabla_X Y+\operatorname{Rc}(Y)-\frac{1}{2\tilde{\tau}}Y\right|^2d\tilde{\tau}+\frac{|Y(\tau)|^2}{\sqrt{\tau}}.$$

Hence we have proved

Lemma 3.2. *Let $\gamma : [0,\tau] \to N$ be an $\mathcal{L}$-geodesic and let γ_s be a smooth variation of $\gamma = \gamma_0$. Assume the variation field $Y = \frac{\partial}{\partial s}\gamma_s$ satisfies that $Y(0) = 0$ when $s = 0$. Then*

$$\begin{aligned}&\frac{d^2\mathcal{L}(\gamma_s)}{ds^2}\bigg|_{s=0} - 2\sqrt{\tau}\langle\nabla_Y Y, X\rangle(\tau) + 2\sqrt{\tau}\,\mathrm{Rc}(Y,Y)(\tau)\\ =&-\int_0^\tau \sqrt{\tilde\tau}H(X,Y)\,d\tilde\tau + \int_0^\tau 2\sqrt{\tilde\tau}\left|\nabla_X Y + \mathrm{Rc}(Y) - \frac{1}{2\tilde\tau}Y\right|^2 d\tilde\tau + \frac{|Y(\tau)|^2}{\sqrt{\tau}},\end{aligned} \tag{3.15}$$

where $H(X,Y)$ is defined by (3.13).

3.4. Estimate of second derivatives of L-distance

In this subsection we give some consequences of the second variation formula of $\mathcal{L}$-length. Below L-distance is defined using the basepoint $(p,0)$.

3.4A. Estimate of Hessian of L-distance. Given $(x,\tau) \in N\times(0,T]$, let $\gamma : [0,\tau] \to N$ be a minimal $\mathcal{L}$-geodesic from p to x. Fix a vector $Y \in T_xN$ and define a vector field $\tilde Y(\tilde\tau)$ along γ by solving the following ode along γ:

$$\nabla_X \tilde Y = -\mathrm{Rc}\left(\tilde Y\right) + \frac{1}{2\tilde\tau}\tilde Y, \quad \tilde\tau \in [0,\tau], \tag{3.16a}$$

$$\tilde Y(\tau) = Y. \tag{3.16b}$$

A direct computation gives $\frac{d}{d\tilde\tau}|\tilde Y|^2 = \frac{1}{\tilde\tau}|\tilde Y|^2$, which implies that

$$|\tilde Y(\tilde\tau)|^2_{g(\tilde\tau)} = |\tilde Y|^2 = \frac{\tilde\tau}{\tau}|Y|^2. \tag{3.17}$$

Hence from (3.15) we have

$$\begin{aligned}&\frac{d^2\mathcal{L}(\gamma_s)}{ds^2}\bigg|_{s=0} - 2\sqrt{\tau}\langle\nabla_Y Y, X\rangle + 2\sqrt{\tau}\,\mathrm{Rc}(Y,Y)\\ =&-\int_0^\tau \sqrt{\tilde\tau}H\left(X,\tilde Y\right)d\tilde\tau + \frac{|Y|^2}{\sqrt{\tau}}\end{aligned} \tag{3.18}$$

for any γ_s with variation field being $\tilde Y(\tilde\tau)$.

Let $\gamma_s : [0,\tau] \to N$, $s \in (-\varepsilon,\varepsilon)$, be a smooth family of paths with

$$\frac{\partial\gamma_s}{\partial s}\bigg|_{s=0}(\tilde\tau) = \tilde Y(\tilde\tau) \quad \text{and} \quad \left(\nabla_{\frac{\partial\gamma_s}{\partial s}}\frac{\partial\gamma_s}{\partial s}\right)\bigg|_{s=0}(\tau) = 0.$$

Since $\mathcal{L}(\gamma_s)$ is an upper barrier function for the L-distance function $L(\gamma_s(\tau),\tau)$ at $s=0$, we have

$$\left(\mathrm{Hess}_{(x,\tau)} L\right)(Y,Y) \le \frac{d^2}{ds^2}\bigg|_{s=0}\mathcal{L}(\gamma_s).$$

From $(\nabla_Y Y)(\tau) = 0$ and (3.18) we get

Theorem 3.3 (**Hessian Comparison for L-distance**). *Let γ be a minimal $\mathcal{L}$-geodesic from $(p,0)$ to (x,τ). Given $Y \in T_xN$, let $\tilde{Y}(\tilde{\tau})$ be a solution of (3.16a) and (3.16b). Then*

$$\left(\mathrm{Hess}_{(x,\tau)} L\right)(Y,Y) \leq -\int_0^\tau \sqrt{\tilde{\tau}} H\left(X,\tilde{Y}\right) d\tilde{\tau} + \frac{|Y|^2}{\sqrt{\tau}} - 2\sqrt{\tau}\,\mathrm{Rc}\,(Y,Y)\,, \tag{3.19}$$

where X is the tangent vector field of γ and $H(X,\tilde{Y})$ is defined by (3.13). Equality in (3.19) *holds when $L(\cdot,\tau)$ is C^2 at x and $\tilde{Y}(\tilde{\tau})$ is the variation vector field of a family of minimal $\mathcal{L}$-geodesics γ_s satisfying $\left(\nabla_{\frac{\partial \gamma_s}{\partial s}} \frac{\partial \gamma_s}{\partial s}\right)\Big|_{s=0}(\tau) = 0$.*

3.4B. Laplacian comparison theorem for L-distance. Given $(x,\tau) \in N \times (0,T]$, let $\gamma : [0,\tau] \to N$ be a minimal $\mathcal{L}$-geodesic from p to x and let $X = \dot{\gamma}$. Fix an orthonormal basis $\{E_i\}_{i=1}^n$ of $(T_xN, g(x,\tau))$. For each i we define the vector field $\tilde{E}_i(\tilde{\tau})$ along γ to be the solution of (3.16a) with $\tilde{E}_i(\tau) = E_i$. A direct computation gives $\frac{d}{d\tilde{\tau}}\left\langle \tilde{E}_i, \tilde{E}_j \right\rangle = \frac{1}{\tilde{\tau}}\left\langle \tilde{E}_i, \tilde{E}_j \right\rangle$, which implies that

$$\left\langle \tilde{E}_i, \tilde{E}_j \right\rangle(\tilde{\tau}) = \frac{\tilde{\tau}}{\tau}\langle E_i, E_j\rangle = \frac{\tilde{\tau}}{\tau}\delta_{ij}. \tag{3.20}$$

Hence the matrix Harnack expression

$$\sum_{i=1}^n H\left(X(\tilde{\tau}), \tilde{E}_i(\tilde{\tau})\right) = \frac{\tilde{\tau}}{\tau}\sum_{i=1}^n H\left(X(\tilde{\tau}), \sqrt{\frac{\tau}{\tilde{\tau}}}\tilde{E}_i(\tilde{\tau})\right) = \frac{\tilde{\tau}}{\tau} H(X)\,, \tag{3.21}$$

where $H(X)$ is the **trace Harnack quantity**

$$H(X) = -\frac{\partial R}{\partial \tilde{\tau}} - 2\nabla R \cdot X + 2\,\mathrm{Rc}\,(X,X) - \frac{R}{\tilde{\tau}}. \tag{3.22}$$

Taking $Y = E_i$ in (3.19) and summing over i, we have

$$\begin{aligned}
\Delta L(x,\tau) &= \sum_{i=1}^n \left(\mathrm{Hess}_{(x,\tau)}\right) L(E_i,E_i) \\
&\leq -\int_0^\tau \sqrt{\tilde{\tau}} \sum_{i=1}^n H\left(X(\tilde{\tau}), \tilde{E}_i(\tilde{\tau})\right) d\tilde{\tau} + \frac{n}{\sqrt{\tau}} - 2\sqrt{\tau}\sum_{i=1}^n \mathrm{Rc}\,(E_i,E_i) \\
&= -\int_0^\tau \frac{\tilde{\tau}^{3/2}}{\tau} H(X)\, d\tilde{\tau} + \frac{n}{\sqrt{\tau}} - 2\sqrt{\tau} R(x,\tau) \\
&= -\frac{1}{\tau}K + \frac{n}{\sqrt{\tau}} - 2\sqrt{\tau} R(x,\tau)\,,
\end{aligned}$$

where K is the **trace Harnack integral**

$$K = K(\gamma,\tau) = \int_0^\tau \tilde{\tau}^{3/2} H(X)\, d\tilde{\tau}. \tag{3.23}$$

We have proven

Theorem 3.4 (**Laplacian comparison for L-distance**). *For any $(x,\tau)\in N\times(0,T]$ the L-distance satisfies*

$$\Delta L(x,\tau)\leq -\frac{1}{\tau}K+\frac{n}{\sqrt{\tau}}-2\sqrt{\tau}R(x,\tau), \tag{3.24}$$

where K is defined by (3.23) using a minimal $\mathcal{L}$-geodesic $\gamma:[0,\tau]\to N$ from p to x.

3.5. Equalities and inequalities satisfied by L and ℓ

In this subsection L-distance and reduced distance ℓ are defined using basepoint $(p,0)$.

3.5A. A formula for K in (3.23). We need a better formula for K in (3.23). Let $\gamma:[0,\tau]\to N$ be an $\mathcal{L}$-geodesic from p to x and let $X=\dot{\gamma}$. Using the $\mathcal{L}$-geodesic equation (3.3), we compute the evolution of the $\mathcal{L}$-length integrand for γ

$$\begin{aligned}
&\frac{d}{d\tilde{\tau}}\left(R(\gamma(\tilde{\tau}),\tilde{\tau})+|X(\tilde{\tau})|^2_{g(\tilde{\tau})}\right)\\
&=\frac{\partial R}{\partial\tilde{\tau}}+\nabla R\cdot X+2\operatorname{Rc}(X,X)+2\langle\nabla_X X,X\rangle\\
&=\frac{\partial R}{\partial\tilde{\tau}}+\nabla R\cdot X+2\operatorname{Rc}(X,X)+\left\langle\nabla R-4\operatorname{Rc}(X)-\frac{1}{\tilde{\tau}}X,X\right\rangle\\
&=\frac{\partial R}{\partial\tilde{\tau}}+2\nabla R\cdot X-2\operatorname{Rc}(X,X)-\frac{1}{\tilde{\tau}}|X|^2.
\end{aligned}$$

Hence

$$\frac{d}{d\tilde{\tau}}\left(R+|X|^2\right)=-H(X)-\frac{1}{\tilde{\tau}}\left(R+|X|^2\right).$$

Multiplying the above equation by $\tilde{\tau}^{3/2}$ and integrating by parts, we get

$$\begin{aligned}
-K(\gamma,\tau)&=\int_0^\tau\left[\tilde{\tau}^{3/2}\frac{d}{d\tilde{\tau}}\left(R+|X|^2\right)+\tilde{\tau}^{1/2}\left(R+|X|^2\right)\right]d\tilde{\tau}\\
&=\tau^{3/2}\left(R(\gamma(\tau),\tau)+|X(\tau)|^2\right)-\frac{1}{2}\int_0^\tau\tilde{\tau}^{1/2}\left(R+|X|^2\right)d\tilde{\tau}\\
&=\tau^{3/2}\left(R(x,\tau)+|X(\tau)|^2\right)-\frac{1}{2}\mathcal{L}(\gamma).
\end{aligned}$$

We have proved that for any minimal $\mathcal{L}$-geodesic $\gamma:[0,\tau]\to N$ from p to x, we have

$$\tau^{3/2}\left(R(x,\tau)+|X(\tau)|^2\right)=-K(\gamma,\tau)+\frac{1}{2}L(x,\tau). \tag{3.25}$$

3.5B. Equalities and inequalities satisfied by L-distance. Using (3.25), we can rewrite (3.9) and (3.8), and (3.24) as the following: At (x,τ),

$$\frac{\partial L}{\partial \tau} = \frac{1}{\tau}K - \frac{1}{2\tau}L + 2\sqrt{\tau}R, \tag{3.26a}$$

$$|\nabla L|^2 = -4\tau R - \frac{4}{\sqrt{\tau}}K + \frac{2}{\sqrt{\tau}}L, \tag{3.26b}$$

$$\Delta L \leq -\frac{1}{\tau}K + \frac{n}{\sqrt{\tau}} - 2\sqrt{\tau}R, \tag{3.26c}$$

where $K = K(\gamma,\tau)$ is given by *(3.23)* and $\gamma : [0,\tau] \to N$ is a minimal $\mathcal{L}$-geodesic from p to x.

3.5C. Equalities and inequalities satisfied by reduced distance ℓ. Recall that the reduced distance $\ell(x,\tau) = \frac{1}{2\sqrt{\tau}}L(x,\tau)$. We have at (x,τ),

$$\frac{\partial \ell}{\partial \tau} = \frac{1}{2\tau^{3/2}}K - \frac{\ell}{\tau} + R, \tag{3.27a}$$

$$|\nabla \ell|^2 = -R - \frac{1}{\tau^{3/2}}K + \frac{\ell}{\tau}, \tag{3.27b}$$

$$\Delta \ell \leq -\frac{1}{2\tau^{3/2}}K + \frac{n}{2\tau} - R, \tag{3.27c}$$

where $K = K(\gamma,\tau)$ is given by *(3.23)* and $\gamma : [0,\tau] \to N$ is a minimal $\mathcal{L}$-geodesic from p to x.

Note that in (3.27a), (3.27b) and (3.27c) the trace Harnack integral K depends on the path γ which is not favorable. However, from (3.27a), (3.27b) and (3.27c) we have the following four partial differential inequalities or equality which do not involve K.

Lemma 3.5. *At (x,τ) the reduced distance ℓ satisfies*

$$\frac{\partial \ell}{\partial \tau} - \Delta \ell + |\nabla \ell|^2 - R + \frac{n}{2\tau} \geq 0, \tag{3.28a}$$

$$2\Delta \ell - |\nabla \ell|^2 + R + \frac{\ell - n}{\tau} \leq 0, \tag{3.28b}$$

$$\frac{\partial \ell}{\partial \tau} + \Delta \ell + \frac{\ell}{\tau} - \frac{n}{2\tau} \leq 0, \tag{3.28c}$$

$$2\frac{\partial \ell}{\partial \tau} + |\nabla \ell|^2 - R + \frac{\ell}{\tau} = 0, \tag{3.28d}$$

$$\lim_{\tau \to 0_+} \frac{\ell(x,\tau)}{\left(d_{g(0)}(p,x)\right)^2 \Big/ 4\tau} = 1, \tag{3.28e}$$

$$\inf_{x \in N} \ell(x,\tau) \leq \frac{n}{2}. \tag{3.28f}$$

We skip the proof of (3.28e) but use it to give a proof of (3.28f). It follows from (3.28c) that

$$\left(\frac{\partial}{\partial(-\tau)}-\Delta\right)(4\tau\ell(x,\tau)-2n\tau)\geq 0.$$

It follows from the maximum principle that $\inf_{x\in N}(4\tau\ell(x,\tau)-2n\tau)$ is a nondecreasing function of $-\tau$. Note that (3.28e) implies $\lim_{\tau\to 0_+}4\tau\ell(x,\tau)=\left(d_{g(0)}(p,x)\right)^2$ and hence $\lim_{\tau\to 0_+}\inf_{x\in N}(4\tau\ell(x,\tau)-2n\tau)=0$. We get $\inf_{x\in N}(4\tau\ell(x,\tau)-2n\tau)\leq 0$ and (3.28f) follows.

3.6. $\mathcal{L}$-Jacobi field

In this subsection we discuss the $\mathcal{L}$-Jacobi field associated with $\mathcal{L}$-length, which is analogous to the Jacobi field associated with length.

3.6A. $\mathcal{L}$-Jacobi field. Now we consider the moduli space of $\mathcal{L}$-geodesics, the tangent direction to this space satisfies a second order linear ode. Let $\gamma_s:[0,\tau]\to N$, $s\in(-\varepsilon,\varepsilon)$, be a smooth family of $\mathcal{L}$-geodesics. Denote $\gamma_0=\gamma$, $X_s=\dot{\gamma}_s$, and $Y_s=\frac{d}{ds}\gamma_s$. Taking ∇_{Y_s}-derivative of the $\mathcal{L}$-geodesic equation (3.3) for γ_s, we compute

$$\begin{aligned}\nabla_{X_s}(\nabla_{X_s}Y_s)&=\nabla_{X_s}(\nabla_{Y_s}X_s)=\mathrm{Rm}(X_s,Y_s)X_s+\nabla_{Y_s}(\nabla_{X_s}X_s)\\&=\mathrm{Rm}(X_s,Y_s)X_s+\nabla_{Y_s}\left(\frac{1}{2}\nabla R-2\,\mathrm{Rc}(X_s)-\frac{1}{2\tilde{\tau}}X_s\right).\end{aligned}$$

Set $s=0$, then $Y(\tilde{\tau})=Y_0(\tilde{\tau})$ satisfies the following ode called the **$\mathcal{L}$-Jacobi equation**:

$$\begin{aligned}&\nabla_X(\nabla_X Y)\\&=-2\,\mathrm{Rc}(\nabla_X Y)-\frac{1}{2\tilde{\tau}}\nabla_X Y+\mathrm{Rm}(X,Y)X+\frac{1}{2}\nabla_Y(\nabla R)-2(\nabla_Y\,\mathrm{Rc})(X).\end{aligned}\tag{3.29}$$

We call any solution of the above equation an **$\mathcal{L}$-Jacobi field**.

Using the parametrization defined in (3.5) and $Z(\tilde{\sigma})=\frac{d\beta}{d\tilde{\sigma}}=\sqrt{\tilde{\tau}}X(\tilde{\tau})$, we can rewrite the $\mathcal{L}$-Jacobi equation of $\hat{Y}(\tilde{\sigma})=Y(\tilde{\tau})$ as

$$\begin{aligned}&\nabla_Z\left(\nabla_Z\hat{Y}\right)\\&=-2\tilde{\sigma}\,\mathrm{Rc}\left(\nabla_Z\hat{Y}\right)+\mathrm{Rm}\left(Z,\hat{Y}\right)Z+\frac{\tilde{\sigma}^2}{2}\nabla_{\hat{Y}}(\nabla R)-2\tilde{\sigma}\left(\nabla_{\hat{Y}}\,\mathrm{Rc}\right)(Z).\end{aligned}\tag{3.30}$$

Hence the initial value problem of (3.30) is solvable.

3.6B. Estimate of $\mathcal{L}$-Jacobi field. Let γ_s be as in §3.6A. By the first variation formula for the $\mathcal{L}$-length,

$$\left.\frac{d}{ds}\right|_{s=0}\mathcal{L}(\gamma_s)=2\sqrt{\tau}\,\langle X_s,Y_s\rangle(\tau).$$

We differentiate this again to get

$$\left.\frac{d^2}{ds^2}\right|_{s=0}\mathcal{L}(\gamma_s)=2\sqrt{\tau}\,\langle\nabla_X Y,Y\rangle(\tau)+2\sqrt{\tau}\,\langle X,\nabla_{Y_s}Y_s|_{s=0}\rangle(\tau).$$

Now we compute the derivative of the norm squared of the $\mathcal{L}$-Jacobi field

$$\begin{aligned}&\left.\frac{d}{d\tilde{\tau}}\right|_{\tilde{\tau}=\tau}|Y(\tilde{\tau})|^2_{g(\tilde{\tau})} = 2\,\mathrm{Rc}\,(Y,Y)(\tau) + 2\,\langle \nabla_X Y, Y\rangle(\tau)\\ &= 2\,\mathrm{Rc}\,(Y,Y)(\tau) + \frac{1}{\sqrt{\tau}}\left(\left.\frac{d^2}{ds^2}\right|_{s=0}\mathcal{L}(\gamma_s)\right) - 2\,\langle X,\, \nabla_{Y_s}Y_s|_{s=0}\rangle(\tau),\end{aligned} \tag{3.31}$$

Let $\tilde{Y}$ be a vector field along γ which satisfies (3.16a) with $\tilde{Y}(\tau) = Y(\tau)$, here we need to assume $Y(0) = 0$. Let $\tilde{\gamma}_s : [0,\tau] \to N$ be a variation of γ with

$$\left.\frac{\partial}{\partial s}\right|_{s=0}\tilde{\gamma}_s = \tilde{Y}, \quad \tilde{\gamma}_s(\tau) = \gamma_s(\tau), \quad \text{and} \quad \tilde{\gamma}_s(0) = \gamma_s(0).$$

Note that the choice of $\tilde{\gamma}_s$ implies $\nabla_{Y_s}Y_s|_{s=0}(\tau) = \left(\nabla_{\frac{\partial\tilde{\gamma}_s}{\partial s}}\frac{\partial\tilde{\gamma}_s}{\partial s}\right)\Big|_{s=0}(\tau)$. If we assume that the γ_s are all minimal $\mathcal{L}$-geodesics, then $\mathcal{L}(\gamma_s) \le \mathcal{L}(\tilde{\gamma}_s)$ for all s, and equality holds at $s = 0$. Hence

$$\left.\frac{d^2}{ds^2}\right|_{s=0}\mathcal{L}(\gamma_s) \le \left.\frac{d^2}{ds^2}\right|_{s=0}\mathcal{L}(\tilde{\gamma}_s),$$

where equality holds if $\tilde{Y}$ is an $\mathcal{L}$-Jacobi field. Combining this with (3.31), we get

$$\left.\frac{d}{d\tilde{\tau}}\right|_{\tilde{\tau}=\tau}|Y|^2 \le 2\,\mathrm{Rc}\,(Y,Y)(\tau) + \frac{1}{\sqrt{\tau}}\left(\left.\frac{d^2}{ds^2}\right|_{s=0}\mathcal{L}(\tilde{\gamma}_s)\right) - 2\,\langle X, \nabla_Y Y\rangle(\tau).$$

By (3.18), we have

$$\begin{aligned}&\left(\left.\frac{d^2}{ds^2}\right|_{s=0}\mathcal{L}(\tilde{\gamma}_s)\right) - 2\sqrt{\tau}\left\langle X,\, \left(\nabla_{\frac{\partial\tilde{\gamma}_s}{\partial s}}\frac{\partial\tilde{\gamma}_s}{\partial s}\right)\Big|_{s=0}\right\rangle(\tau)\\ &= -\int_0^\tau \sqrt{\tilde{\tau}}H\left(X,\tilde{Y}\right)d\tilde{\tau} + \frac{|Y(\tau)|^2}{\sqrt{\tau}} - 2\sqrt{\tau}\,\mathrm{Rc}\,(Y,Y)(\tau).\end{aligned}$$

Hence we have proved

Lemma 3.6. *Let $\gamma_s : [0,\bar{\tau}] \to N$ be a smooth family of minimal $\mathcal{L}$-geodesics with $\gamma_s(0) = p$. Then for any $\tau \in (0,\bar{\tau}]$ the $\mathcal{L}$-Jacobi field $Y(\tau) = \left.\frac{d\gamma_s}{ds}\right|_{s=0}(\tau)$ satisfies the estimate*

$$\frac{d}{d\tau}|Y(\tau)|^2_{g(\tau)} \le -\frac{1}{\sqrt{\tau}}\int_0^\tau \sqrt{\tilde{\tau}}H\left(X,\tilde{Y}\right)d\tilde{\tau} + \frac{|Y(\tau)|^2}{\tau}, \tag{3.32}$$

where X is the tangent vector field of $\gamma = \gamma_0$, $\tilde{Y}(\tilde{\tau})$ is a solution of (3.16a) on $[0,\tau]$ with $\tilde{Y}(\tau) = Y(\tau)$, and $H(X,\tilde{Y})$ is defined by (3.13).

3.7. $\mathcal{L}$-exponential map and $\mathcal{L}$-Jacobian

In §2 we have seen the role played by the exponential map in considering the volume of balls, in this subsection we consider the $\mathcal{L}$-exponential map, in next section we will see that it plays a similar role in reduced volume. In this subsection we use the basepoint $(p,0)$ to define the L-distance.

3.7A. $\mathcal{L}$-exponential map. Given τ, define the **$\mathcal{L}$-exponential map at time τ**

$$\mathcal{L}_\tau \exp : T_pN \to N, \qquad \mathcal{L}_\tau \exp(V) = \gamma_V(\tau),$$

where $\gamma_V(\tilde{\tau}) = \beta(\tilde{\sigma})$ is the $\mathcal{L}$-geodesic obtained by solving (3.7) with $\frac{d\beta}{d\tilde{\sigma}}(0) = V \in T_pN$ and $\beta(0) = p$.

3.7B. $\mathcal{L}$-Jacobian. If we want to compute the tangent map $D(\mathcal{L}_\tau \exp(V))$ of $\mathcal{L}_\tau \exp$ at $V \in T_pN$, we need to consider a family of $\mathcal{L}$-geodesics $\gamma_{V_s}(\tilde{\tau})$ where V_s is a variation of V, and hence we need to consider the corresponding $\mathcal{L}$-Jacobi field.

Given an orthonormal basis $\{E_i^0\}_{i=1}^n$ of $(T_pN, g(p,0))$, let $J_i^V(\tilde{\tau}) = \hat{J}_i^V(\tilde{\sigma})$, $i = 1, \dots, n$, be $\mathcal{L}$-Jacobi fields along γ_V where $\tilde{\sigma} = 2\sqrt{\tilde{\tau}}$ and $\hat{J}_i^V(\tilde{\sigma})$ is defined by solving (3.30) with initial value

$$\hat{J}_i^V(0) = 0 \quad \text{and} \quad \left(\nabla_Z \hat{J}_i^V\right)(0) = E_i^0.$$

Then $D(\mathcal{L}_\tau \exp(V))(E_i^0) = J_i^V(\tau)$, and the Jacobian of the $\mathcal{L}$-exponential map $\mathcal{L}_\tau \mathrm{J}_V \in \mathbb{R}$ (called the **$\mathcal{L}$-Jacobian**) is given by

$$\mathcal{L}_\tau \mathrm{J}_V = \sqrt{\det\left(\left\langle J_i^V(\tau), J_j^V(\tau)\right\rangle_{g(\mathcal{L}_\tau \exp(V),\tau)}\right)_{n\times n}}. \tag{3.33}$$

Note that the pull-back volume form is

$$(\mathcal{L}_\tau \exp(V))^* \, d\mu_{g(\mathcal{L}_\tau \exp(V),\tau)} = \mathcal{L}_\tau \mathrm{J}_V \, dy$$

where dy is the standard Euclidean volume form on $(T_pN, g(p,0))$.

Let $\bar{E}_i(\tau)$ be the parallel translation of E_i^0 along $\gamma_V(\tilde{\tau})$ with respect to $g(0)$. From the definition of $\hat{J}_i^V(\tilde{\sigma})$ we have $|J_i^V(\tau) - 2\sqrt{\tau}\bar{E}_i(\tau)|_{g(0)} = o(2\sqrt{\tau})$ and hence we get the following asymptotic behavior of the $\mathcal{L}$-Jacobian

$$\lim_{\tau\to 0_+} \frac{\mathcal{L}_\tau \mathrm{J}_V}{\tau^{n/2}} = \lim_{\tau\to 0_+} \tau^{-n/2}\sqrt{\det\left(\langle 2\sqrt{\tau}\bar{E}_i(\tau), 2\sqrt{\tau}\bar{E}_j(\tau)\rangle_{g(0)}\right)} = 2^n. \tag{3.34}$$

3.7C. Estimate of $\mathcal{L}$-Jacobian. We have the following estimate of the time-derivative of the $\mathcal{L}$-Jacobian, which follows from the estimate of Jacobi fields.

Proposition 3.7. *Fix a $V \in T_pN$, let $\gamma_V(\tilde{\tau})$, $\tilde{\tau} \in [0, \bar{\tau}]$, be a minimal $\mathcal{L}$-geodesic with $\gamma_V(0) = p$ and $\lim_{\tilde{\tau}\to 0_+} \sqrt{\tilde{\tau}}\frac{d\gamma_V}{d\tilde{\tau}} = V$. For any $\tau \in (0, \bar{\tau})$ the $\mathcal{L}$-Jacobian $\mathcal{L}_\tau \mathrm{J}_V$ satisfies*

$$\left(\frac{d}{d\tau}\log \mathcal{L}_\tau \mathrm{J}_V\right) \le \frac{n}{2\tau} - \frac{1}{2\tau^{\frac{3}{2}}}K, \tag{3.35}$$

where $K = K(\gamma_V, \tau)$ is defined by (3.23).

Proof. Choose an orthonormal basis $\{E_i(\tau)\}$ of $(T_{\gamma_V(\tau)}N, g(\gamma_V(\tau),\tau))$. We can extend $E_i(\tau)$ to an $\mathcal{L}$-Jacobi field $E_i(\tilde{\tau})$ along γ_V for $\tilde{\tau}\in[0,\tau]$ with $E_i(0)=0$. Then there is a matrix $\left(A_i^j\right)\in \mathrm{GL}(n,\mathbb{R})$, such that

$$J_i^V(\tilde{\tau}) = \sum_{j=1}^n A_i^j E_j(\tilde{\tau})$$

for all $\tilde{\tau}\in[0,\tau]$. The reason for the existence of $E_i(\tilde{\tau})$ and $\left(A_i^j\right)$ is that there is no nontrivial $\mathcal{L}$-Jacobi field along $\gamma_V(\tilde{\tau})$ which vanishes at the endpoints $\tilde{\tau}=0,\ \tau$.

Now we compute the evolution of the $\mathcal{L}$-Jacobian along γ_V using (3.33) and (3.32):

$$\begin{aligned}
&\frac{d}{d\tilde{\tau}}\Big|_{\tilde{\tau}=\tau}\log\mathcal{L}_{\tilde{\tau}}\,\mathrm{J}_V = \frac{d}{d\tilde{\tau}}\Big|_{\tilde{\tau}=\tau}\log\sqrt{\det\left(\left\langle\sum_{k=1}^n A_i^k E_k(\tilde{\tau}),\sum_{\ell=1}^n A_j^\ell E_\ell(\tilde{\tau})\right\rangle_{g(\gamma_V(\tilde{\tau}),\tilde{\tau})}\right)}\\
&=\frac{1}{2}\,\frac{d}{d\tilde{\tau}}\Big|_{\tilde{\tau}=\tau}\log\det\left(\langle E_k,E_l\rangle(\tilde{\tau})\right)+\frac{1}{2}\,\frac{d}{d\tilde{\tau}}\Big|_{\tilde{\tau}=\tau}\log\det\left(A_i^k\right)+\frac{1}{2}\,\frac{d}{d\tilde{\tau}}\Big|_{\tilde{\tau}=\tau}\log\det\left(A_j^\ell\right)\\
&=\frac{1}{2}\sum_{i=1}^n\frac{d}{d\tilde{\tau}}\Big|_{\tilde{\tau}=\tau}\langle E_i,E_i\rangle(\tilde{\tau})\\
&\le -\frac{1}{2}\frac{1}{\sqrt{\tau}}\int_0^\tau\sqrt{\tilde{\tau}}\sum_{i=1}^n H\left(X,\tilde{E}_i\right)d\tilde{\tau}+\frac{1}{2}\sum_{i=1}^n\frac{|E_i(\tau)|^2}{\tau}\\
&=-\frac{1}{2\tau^{3/2}}\int_0^\tau\tilde{\tau}^{3/2}H(X)\,d\tilde{\tau}+\frac{n}{2\tau}.
\end{aligned}$$

Here $\det(\langle E_k,E_l\rangle(\tilde{\tau}))$ denotes the determinant of $n\times n$ matrix $(\langle E_k,E_l\rangle(\tilde{\tau}))$, the $\tilde{E}_i(\tilde{\tau})$ are the vector fields along γ_V satisfying (3.16a) and $\tilde{E}_i(\tau)=E_i(\tau)$, and in the last equality above we have used (3.21). □

4. Reduced volume

In this section we use the properties of reduced distance to prove the monotonicity of reduced volume, in the next section we will see that this monotonicity has fundamental consequences about the properties of Ricci flow.

Throughout this section N^n is an n-dimensional connected oriented manifold, and $(N^n,g(\tau))$, $\tau\in[0,T]$, is a complete solution to the **backward Ricci flow** $\frac{\partial}{\partial\tau}g(\tau)=2\,\mathrm{Rc}(g(\tau))$ with bounded Riemann curvature $\sup_{M\times[0,T]}|\mathrm{Rm}(x,\tau)|<\infty$. The reduced distance ℓ is defined using the basepoint $(p,0)$ for some $p\in N$. Several claims in this section will not be proved.

4.1. Definition of reduced volume

The **reduced volume** function $\tilde{V} : (0, T] \to \mathbb{R}_{>0}$ is defined by

$$\tilde{V}(\tau) = \int_N (4\pi\tau)^{-n/2} \exp\left[-\ell(x,\tau)\right] d\mu_{g(\tau)}(x), \tag{4.1}$$

and $\tilde{V}(0)$ is defined to be 1. We will see that the integral in (4.1) is finite in §4.2A.

Comment. 1) A simple exercise shows that on the Euclidean space where $g(\tau) = g_{\text{Euc}}$ we have $\ell(x,\tau) = \frac{(d_{\text{Euc}}(x,p))^2}{4\tau}$ and $\tilde{V}(\tau) = 1$ for all τ.

2) There is some similarity between the reduced volume and the volume ratio $\frac{\text{Vol}\, B(p,r)}{\omega_n r^n}$ which appeared in the Bishop-Gromov volume comparison theorem for complete manifolds with $\text{Rc} \geq 0$.

In the rest of this subsection we give another formula for the integral in (4.1). The **$\mathcal{L}$-cut-locus** $\mathcal{L}_\tau \,\text{Cut}_p$ of the map $\mathcal{L}_\tau \exp$ consists of points $\mathcal{L}_\tau \exp(V)$ where either $V \in T_pN$ is a critical point of $\mathcal{L}_\tau \exp$ or there is $\hat{V} \neq V$ such that $\gamma_{\hat{V}}$ (as defined above (3.35)) is a minimal $\mathcal{L}$-geodesic over $[0,\tau]$ joins p and $\mathcal{L}_\tau \exp(V)$. Note that $\mathcal{L}_\tau \,\text{Cut}_p$ has measure zero in $(N, g(\tau))$ (claim).

Given $V \in T_pN$ there is a unique $\tau_V \in (0,T]$ such that the $\mathcal{L}$-geodesic $\gamma_V|_{[0,\tau]}$ is minimal when $\tau < \tau_V$ and is not minimal when $\tau > \tau_V$ (claim). We define $\Omega_p(\tau) = \{V \in T_pN, \tau < \tau_V\}$. Then

$$\mathcal{L}_\tau \exp : \Omega_p(\tau) \to N \setminus \mathcal{L}_\tau \,\text{Cut}_p$$

is a diffeomorphism (claim).

Let dy be the standard Euclidean volume form on $(T_pN, g(p,0))$. Using $\mathcal{L}_\tau \exp$ and $\Omega_p(\tau)$ we can rewrite the reduced volume as

$$\begin{aligned}
\tilde{V}(\tau) &= \int_{N \setminus \mathcal{L}_\tau \,\text{Cut}_p} (4\pi\tau)^{-n/2} \exp\left[-\ell(x,\tau)\right] d\mu_{g(\tau)}(x) \\
&= \int_{\Omega_p(\tau)} (4\pi\tau)^{-n/2} e^{-\ell(\mathcal{L}_\tau \exp(V),\tau)} \mathcal{L}_\tau \,\text{J}_V \, dy\,(V), \qquad (4.2) \\
&= \int_{T_pN} (4\pi\tau)^{-n/2} e^{-\ell(\gamma_V(\tau),\tau)} \mathcal{L}_\tau \,\text{J}_V \, dy\,(V), \qquad (4.3)
\end{aligned}$$

where we have used the formula after (3.33) to get (4.2) and the convention $\mathcal{L}_\tau \,\text{J}_V = 0$ for $V \notin \Omega_p(\tau)$ to get (4.3).

4.2. The monotonicity of reduced volume

In this subsection we give two proofs of the monotonicity.

4.2A. Monotonicity of reduced volume using $\mathcal{L}$-Jacobian. For any $V \in \Omega_p(\tau)$, from (3.8) we have $\dot{\gamma}_V(\tau) = (\nabla\ell)(\gamma_V(\tau), \tau)$. We compute

$$\begin{aligned}
&\frac{d}{d\tau}\left[(4\pi\tau)^{-n/2} e^{-\ell(\gamma_V(\tau),\tau)} \mathcal{L}_\tau \mathrm{J}_V\right] \\
=&\left[(4\pi\tau)^{-n/2} e^{-\ell(\gamma_V(\tau),\tau)} \mathcal{L}_\tau \mathrm{J}_V\right]\left(-\frac{n}{2\tau} - \nabla\ell\cdot\dot{\gamma}_V - \frac{\partial\ell}{\partial\tau} + \frac{d}{d\tau}\log \mathcal{L}_\tau \mathrm{J}_V\right) \\
=&\left[(4\pi\tau)^{-n/2} e^{-\ell(\gamma_V(\tau),\tau)} \mathcal{L}_\tau \mathrm{J}_V\right]\left(-\frac{n}{2\tau} - |\nabla\ell|^2 - \frac{\partial\ell}{\partial\tau} + \frac{d}{d\tau}\log \mathcal{L}_\tau \mathrm{J}_V\right) \\
\leq& 0,
\end{aligned}$$

where the last inequality follows from (3.27a), (3.27b) and (3.35). Hence we have proved (i) of the following.

Lemma 4.1. *Using the notation from §4.1 we have*
(i) for any $V \in \Omega_p(\tau)$

$$\frac{d}{d\tau}\left[(4\pi\tau)^{-n/2} e^{-\ell(\gamma_V(\tau),\tau)} \mathcal{L}_\tau \mathrm{J}_V\right] \leq 0. \tag{4.4}$$

(ii) For any $V \in T_p\mathcal{M}$ and $0 \leq \tau \leq T$,

$$(4\pi\tau)^{-n/2} e^{-\ell(\gamma_V(\tau),\tau)} \mathcal{L}_\tau \mathrm{J}_V \leq \pi^{-n/2} e^{-|V|^2_{g(p,0)}}. \tag{4.5}$$

(iii) The reduced volume is well defined and takes values in $(0, 1]$.

Proof. (ii) Similar to (3.28e) we have (claim)

$$\lim_{\tau\to 0_+} \ell(\gamma_V(\tau), \tau) = |V|^2_{g(p,0)}.$$

Hence from (3.34) we have

$$\lim_{\tau\to 0_+} (4\pi\tau)^{-n/2} e^{-\ell(\gamma_V(\tau),\tau)} \mathcal{L}_\tau \mathrm{J}_V = \pi^{-n/2} e^{-|V|^2_{g(p,0)}}. \tag{4.6}$$

(ii) then follows from (i).

(iii) This follows from (4.3), (ii) and

$$\int_{T_pN} \pi^{-n/2} e^{-|V|^2_{g(p,0)}} dy(V) = 1$$

where dy is the standard Euclidean volume form on $(T_pN, g(p,0))$. The above equality and (4.6) indicates that defining $\tilde{V}(0) = 1$ is reasonable. □

The next theorem follows from (4.3) and (4.4).

Theorem 4.2 (**Monotonicity of reduced volume**). *We have $\tilde{V}(\tau_2) \leq \tilde{V}(\tau_1)$ for $\tau_2 \geq \tau_1$, i.e.,*

$$\frac{d}{d\tau}\tilde{V}(\tau) \leq 0. \tag{4.7}$$

4.2B. Monotonicity of reduced volume without using $\mathcal{L}$-Jacobian. We compute formally

$$\begin{aligned}\frac{d}{d\tau}\tilde{V}(\tau) &= \int_N \frac{\partial}{\partial\tau}\left((4\pi\tau)^{-n/2} e^{-\ell(\cdot,\tau)} d\mu_{g(\tau)}\right) && (4.8)\\ &= \int_N \left(-\frac{n}{2\tau} - \frac{\partial\ell}{\partial\tau} + R\right)(4\pi\tau)^{-n/2} e^{-\ell(\cdot,\tau)} d\mu_{g(\tau)} \\ &\le \int_N \left(|\nabla\ell|^2 - \Delta\ell\right)(4\pi\tau)^{-n/2} e^{-\ell(\cdot,\tau)} d\mu_{g(\tau)} && (4.9)\\ &\le 0. && (4.10)\end{aligned}$$

Here (4.9) follows from (3.28a). To justify the switch of the order of differentiation and integration to get (4.8) and the integration by parts to get (4.10), one need certain estimates on the growth of reduced distance $\ell(x,\tau)$ and its derivatives. These estimates are not established in §3, we encourage the reader to find the details of proving (4.8) and (4.10) in the literature.

5. Applications of monotonicity of reduced volume

In this section we give two applications of reduced distance and reduced volume to justify their importance in the study of Ricci flow. The reader can find their other applications in the literature.

5.1. No local collapsing theorem

Recall that if a Ricci flow solution does not exist up to time $+\infty$, we say that it develops a **singularity** in finite time. The no local collapsing theorem is used in the singularity analysis of Ricci flow. When combined with the **Hamilton's Cheeger-Gromov-type compactness theorem** it implies the existence of **singularity models** for Ricci flow developing singularities in finite time.

Throughout this subsection M^n is a n-dimensional connected oriented manifold, and $(M^n, \tilde{g}(t))$, $t \in [0,T)$, is a complete solution to the Ricci flow with $T < \infty$, and we assume $\sup_{M\times[0,T_0]} |\mathrm{Rm}_{\tilde{g}}(x,t)| < \infty$ for all $T_0 < T$.

5.1A. The statement of no local collapsing theorem. Before we state the theorem we need a

Definition 5.1 (**Strongly κ-collapsed**)**.** Let $\kappa > 0$ be a constant. We say that Ricci flow $(M^n, \tilde{g}(t))$, $t \in [0,T)$, is **strongly κ-collapsed at** $(x_0, T_0) \in M \times (0,T)$ **at scale** $r > 0$ if

(i) *(curvature bound in a parabolic cylinder)* $|\mathrm{Rm}_{\tilde{g}}(x,t)| \le \frac{1}{r^2}$ for all $x \in B_{\tilde{g}(T_0)}(x_0, r)$ and $t \in [\max\{T_0 - r^2, 0\}, T_0]$, and

(ii) *(volume of ball is κ-collapsed)*

$$\frac{\mathrm{Vol}_{\tilde{g}(T_0)} B_{\tilde{g}(T_0)}(x_0, r)}{r^n} < \kappa.$$

Given an $r>0$, if for any $T_0 \in [r^2, T)$ and any $x_0 \in M$ the solution $\tilde{g}(t)$ is not strongly κ-collapsed at (x_0, T_0) at scale r, then we say that $(M, \tilde{g}(t))$ is **weakly κ-noncollapsed at scale r**.

The following is the so-called **weakened no local collapsing theorem** ([Pe02], §7.3).

Theorem 5.1. *Let $(M^n, \tilde{g}(t))$, $t \in [0,T)$, be a complete solution to the Ricci flow with $T<\infty$. Suppose there exist $r_1>0$ and $v_1>0$ such that*

$$\mathrm{Vol}_{\tilde{g}(0)} B_{\tilde{g}(0)}(x, r_1) \geq v_1 \text{ for all } x \in M.$$

Then there exists $\kappa>0$ depending only on r_1, v_1, n, T, and $\sup_{M\times[0,T/2]} \mathrm{Rc}_{\tilde{g}}(x,t)$ such that $\tilde{g}(t)$ is weakly κ-noncollapsed at any point $(p, T_0) \in M \times (T/2, T)$ at any scale $r < \sqrt{T/2}$. Here $\sup \mathrm{Rc}_{\tilde{g}}(x,t)$ stands for the largest eigenvalue of $\mathrm{Rc}_{\tilde{g}}(x,t)$.

5.1B. Sketch of the proof of Theorem 5.1. Given a time $T_0 \in \left(\frac{T}{2}, T\right)$, let $g(\tau) = \tilde{g}(T_0 - \tau)$. Then $(M^n, g(\tau))$, $\tau \in [0, T_0]$, is a complete solution to the backward Ricci flow with initial metric $g(0) = \tilde{g}(T_0)$ and with bounded Riemann curvature tensor. Given a point $p \in M$, then we can define reduced distance $\ell(x,\tau)$ and reduced volume $\tilde{V}(\tau)$ using the basepoint $(p,0)$. The theorem follows easily from the following two lemmas.

On one hand, we have

Lemma 5.2. *There exist $c_1(n)>0$ depending only on n and a function $\phi(\epsilon, n)$ satisfying $\lim_{\epsilon\to 0_+} \phi(\varepsilon, n) = 0$ such that if for some κ satisfying $\kappa^{1/n} \leq c_1(n)$, the solution $\tilde{g}(t)$ is strongly κ-collapsed at some (p, T_0) at scale r, where $T_0 \in (\frac{T}{2}, T)$ and $r < \sqrt{T_0}$, then the reduced volume $\tilde{V}$ as defined above has the upper bound*

$$\tilde{V}\left(\varepsilon r^2\right) \leq \phi(\varepsilon, n),$$

where $\varepsilon = \kappa^{1/n}$.

Sketch of the proof of Lemma 5.2. From (4.3) we can write the reduced volume integral over T_pM as $\tilde{V}\left(\varepsilon r^2\right) = \tilde{V}_1\left(\varepsilon r^2\right) + \tilde{V}_2\left(\varepsilon r^2\right)$ where $\tilde{V}_1\left(\varepsilon r^2\right)$ and $\tilde{V}_2\left(\varepsilon r^2\right)$ are the integrals over $\{V \in T_pM, |V|_{g(p,0)} \leq \epsilon^{-1/4}\}$ and $\{V \in T_pM, |V|_{g(p,0)} > \epsilon^{-1/4}\}$, respectively. The lemma is proved by bounding $\tilde{V}_1\left(\varepsilon r^2\right)$ and $\tilde{V}_2\left(\varepsilon r^2\right)$ from above separately.

We have

$$\tilde{V}_1\left(\varepsilon r^2\right) \leq C_1(n)\epsilon^{n/2} \qquad \text{for } \epsilon < c_1(n)$$

where $C_1(n)$ and $c_1(n)$ are positive constants depending only on n. Using the assumption that the solution is strongly κ-collapsed at (p, T_0) at scale r, the upper bound estimate is proved by showing the following two estimates: the $\mathcal{L}$-geodesic $\gamma_V(\tau)$ (as defined in §3.7A) is contained in $B_{\tilde{g}(T_0)}(p, r/2)$ for some choice of $c_1(n)$; and when $|V|_{g(p,0)} \leq \epsilon^{-1/4}$ $\ell(\gamma_V(\varepsilon r^2), \varepsilon r^2)$ is bounded from below by a constant independent of ϵ.

We have

$$\tilde{V}_2\left(\varepsilon r^2\right) \leq C_2(n) e^{-\frac{1}{2\sqrt{\epsilon}}}$$

where $C_2(n)$ is a positive constant depending only on n. To see this estimate, by (4.5) we have

$$\left(4\pi\varepsilon r^2\right)^{-n/2} e^{-\ell\left(\gamma_V\left(\varepsilon r^2\right),\varepsilon r^2\right)}\mathcal{L}_{\varepsilon r^2}\,\mathrm{J}_V \le \pi^{-n/2}e^{-|V|^2_{g(p,0)}}.$$

Then

$$\tilde{V}_2\left(\varepsilon r^2\right) \le \int_{|V|_{g(p,0)}>\epsilon^{-1/4}} \pi^{-n/2}e^{-|V|^2_{g(p,0)}}dy$$

where dy is the standard Euclidean volume form on $(T_pN, g(p,0))$. The estimate follows.

On the other hand, we have

Lemma 5.3. *(i) Fix an arbitrary $r_0 > 0$. There exists a constant $C_3 > 0$, depending only on r_0, n, T, and $\sup_{M\times[0,T/2]}\mathrm{Rc}_{\tilde{g}}(x,t)$, and there exists $x_0 \in M$ such that reduced distance*

$$\ell(x,T_0) \le C_3 \quad \text{for all } x \in B_{\tilde{g}(0)}(x_0,r_0).$$

(ii) Suppose there exist $r_1 > 0$ and $v_1 > 0$ such that

$$\mathrm{Vol}_{\tilde{g}(0)}\, B_{\tilde{g}(0)}(x,r_1) \ge v_1$$

for all $x \in M$. Then there exists a constant $C_4 > 0$, depending only on r_1, v_1, n, T, and $\sup_{M\times[0,T/2]}\mathrm{Rc}_{\tilde{g}}(x,t)$, such that reduced volume

$$\tilde{V}(T_0) \ge C_4.$$

Sketch of the proof of Lemma 5.3. (i) By (3.28f), there is $x_0 \in M$ and a minimal $\mathcal{L}$-geodesic $\gamma_1 : [0, T_0 - \frac{T}{2}] \to M$ joining p and x_0 such that $\frac{1}{2\sqrt{T_0-\frac{T}{2}}}\mathcal{L}(\gamma_1) = \ell(x_0, T_0 - \frac{T}{2}) \le \frac{n}{2}$. Let $\beta : [T_0 - \frac{T}{2}, T_0] \to (M, \tilde{g}(0))$ be the constant speed path joining x_0 and $x \in B_{\tilde{g}(0)}(x_0, r_0)$. Since γ_1 followed by β is a path joining $(p,0)$ and (x,T_0),

$$\ell(x,T_0) \le \frac{1}{2\sqrt{T_0}}\left(\mathcal{L}(\gamma_1) + \int_{T_0-\frac{T}{2}}^{T_0}\sqrt{\tilde{\tau}}\left(R(\beta(\tilde{\tau}),\tilde{\tau}) + \left|\frac{d\beta}{d\tilde{\tau}}(\tilde{\tau})\right|^2_{g(\tilde{\tau})}\right)d\tilde{\tau}\right).$$

Note that the metric $g(\tilde{\tau})$, $\tilde{\tau} \in [T_0 - \frac{T}{2}, T_0]$, corresponds to the metric $\tilde{g}(t)$, $t \in [0, \frac{T}{2}]$, we can estimate the integral above to get (i).

(ii) We compute using x_0 in (i)

$$\begin{aligned}\tilde{V}(T_0) &\ge \int_{B_{\tilde{g}(0)}(x_0,r_1)} (4\pi T_0)^{-\frac{n}{2}} e^{-\ell(x,T_0)} d\mu_{\tilde{g}(0)}(x)\\ &\ge (4\pi T)^{-\frac{n}{2}} e^{-C_3} v_1.\end{aligned}$$

Now we finish the proof of Theorem 5.1. Suppose the solution $\tilde{g}(t)$ is strongly κ-collapsed at some (p,T_0) at scale r, where $\kappa^{1/n} \le c_1(n)$, $T_0 \in (\frac{T}{2}, T)$ and $r < \sqrt{T_0}$. Combining the two lemmas above about the upper and lower bound of reduced volume and the monotonicity of reduced volume, we have

$$C_4 \le \tilde{V}(T_0) \le \tilde{V}\left(\varepsilon r^2\right) \le \phi(\varepsilon, n).$$

This forces $\varepsilon = \kappa^{1/n}$ not going to zero. Hence the theorem is proved.

5.1C. Type I solution and Lemma 5.3. Let $(M^n, \tilde{g}(t))$, $t \in [0,T)$, be a complete solution to the Ricci flow with $T < \infty$. Recall that $\tilde{g}(t)$ is called a **type-I** solution if there is a constant C_0 such that $(T-t)|\operatorname{Rm}_{\tilde{g}}(x,t)| \le C_0$ for all $(x,t) \in M \times [0,T)$. For type-I solutions we have the following modification of Lemma 5.3.

Lemma 5.4. *Let $(M^n, \tilde{g}(t))$, $t \in [0,T)$, be a complete solution to the Ricci flow with $T < \infty$. Suppose for some constants C_0 and $\alpha \in [1, \frac{3}{2})$ we have*

$$(T-t)^{\alpha}|\operatorname{Rc}_{\tilde{g}}(x,t)| \le C_0 \qquad \text{for all } (x,t) \in M \times [0,T),$$

and suppose there exist $x_1 \in M$, $r_1 > 0$, and $v_1 > 0$ such that

$$\operatorname{Vol}_{\tilde{g}(0)} B_{\tilde{g}(0)}(x_1, r_1) \ge v_1.$$

Then for any $A > 0$ there exist two positive constants C_5 and C_6, both depending only on A, α, r_1, v_1, n, T, and C_0, such that for any $p \in B_{\tilde{g}(0)}(x_1, A)$ and $T_0 \in (T/2, T)$ the reduced distance, defined by backward solution $g(\tau) = \tilde{g}(T_0 - \tau)$ and basepoint $(p,0)$, satisfies

$$\ell(x, T_0) \le C_5 \quad \text{for all } x \in B_{\tilde{g}(0)}(x_1, r_1),$$

and the reduced volume $\tilde{V}(T_0) \ge C_6$.

Sketch of the proof of Lemma 5.4. Define $\gamma : [0, T_0] \to M$ to be a path joining p and $x \in B_{\tilde{g}(0)}(x_1, r_1)$ such that $\gamma(\tilde{\tau}) = x$ for $\tilde{\tau} \in [0, T_0 - T/2]$ and $\gamma|_{[T_0 - T/2, T_0]}$ is a constant speed minimal geodesic with respect to metric $\tilde{g}(0)$. By the Ricci curvature bound assumption we have for any $\tilde{\tau} \in [T_0 - T/2, T_0]$

$$|\dot{\gamma}(\tilde{\tau})|^2_{g(\tilde{\tau})} \le e^{2^{\alpha} C_0 T^{1-\alpha}} |\dot{\gamma}(\tilde{\tau})|^2_{\tilde{g}(0)} = e^{2^{\alpha} C_0 T^{1-\alpha}} \cdot \frac{4(A + r_1)^2}{T^2}.$$

Since $|\operatorname{Rc}_g(x,\tau)| \le C_0 \tau^{-\alpha}$, we compute

$$\begin{aligned} \ell(x, T_0) &\le \frac{1}{2\sqrt{T_0}} \mathcal{L}(\gamma) \\ &= \frac{1}{2\sqrt{T_0}} \left(\int_0^{T_0} 2\sqrt{\tilde{\tau}} R_g(\gamma(\tilde{\tau}), \tilde{\tau}) d\tilde{\tau} + \int_{T_0 - T/2}^{T_0} 2\sqrt{\tilde{\tau}} |\dot{\gamma}(\tilde{\tau})|^2_{g(\tilde{\tau})} d\tilde{\tau} \right) \le C_5. \end{aligned}$$

Hence

$$\tilde{V}(T_0) \ge \int_{B_{\tilde{g}(0)}(x_1, r_1)} (4\pi T_0)^{-\frac{n}{2}} e^{-\ell(x, T_0)} d\mu_{\tilde{g}(0)}(x) \ge (4\pi T)^{-\frac{n}{2}} e^{-C_5} v_1.$$

Following the same idea of the proof of Theorem 5.1 (see the paragraph above §5.1C), we have

Lemma 5.5. *Let $(M^n, \tilde{g}(t))$, $t \in [0,T)$, be a complete solution to the Ricci flow with $T < \infty$. Suppose for some constants C_0 and $\alpha \in [1, \frac{3}{2})$ we have*

$$(T-t)^{\alpha}|\operatorname{Rc}_{\tilde{g}}(x,t)| \le C_0 \qquad \text{for all } (x,t) \in M \times [0,T),$$

and suppose there exist $x_1 \in M$, $r_1 > 0$, *and* $v_1 > 0$ *such that* $\mathrm{Vol}_{\tilde{g}(0)} B_{\tilde{g}(0)}(x_1, r_1) \geq v_1$. *Then for any* $A > 0$ *there exists a positive constant* κ, *depending only on* A, α, r_1, v_1, n, T, *and* C_0, *such that at any scale* $r < \sqrt{T/2}$, $\tilde{g}(t)$ *is weakly* κ*-noncollapsed at any point* $(p, T_0) \in B_{\tilde{g}(0)}(x_1, A) \times (T/2, T)$.

Note that this lemma has a weaker volume assumption than that of Theorem 5.1.

5.2. Backward limits of κ-solutions are shrinkers

A special family of singularity models is the so-called κ-**solutions**. The main theorem of this subsection shows that some blow-down limits of the solutions are even more special: shrinking gradient Ricci solitons. This opens the door for possible classification of singularity models in lower dimensions. A near classification knowledge about κ-solutions in dimension 3 enables us to perform surgeries on Ricci flow and eventually leads to the longtime existence of the so-called surgical Ricci flow.

5.2A. κ**-solutions and the theorem**. First we give a

Definition 5.2. Let κ be a positive constant. A complete ancient solution $(M^n, \tilde{g}(t))$, $t \in (-\infty, 0]$, of the Ricci flow is called a κ-**solution** if it satisfies

(i) $\tilde{g}(t)$ is nonflat and has nonnegative curvature operator for each $t \in (-\infty, 0]$.

(ii) Scalar curvature satisfies $\sup_{M \times (-\infty, 0]} R_{\tilde{g}}(x, t) < \infty$.

(iii) $\tilde{g}(t)$ is κ-**noncollapsed on all scales** for all $t \in (-\infty, 0]$; i.e., for any $r > 0$ and for any $(p, t) \in M \times (-\infty, 0]$, if $|\mathrm{Rm}_{\tilde{g}}(x, t)| \leq r^{-2}$ for all $x \in B_{\tilde{g}(t)}(p, r)$, then

$$\frac{\mathrm{Vol}_{\tilde{g}(t)} B_{\tilde{g}(t)}(p, r)}{r^n} \geq \kappa.$$

Given a κ-solution $(\mathcal{M}^n, \tilde{g}(t))$, $t \in (-\infty, 0]$, we define a solution to the backward Ricci flow $(M^n, g(\tau))$, $\tau \in [0, \infty)$, by

$$g(\tau) = \tilde{g}(-\tau).$$

Given a point $p \in M$, we can define the reduced distance $\ell(x, \tau)$ and reduced volume $\tilde{V}(\tau)$ using basepoint $(p, 0)$. Let $q_\tau \in M$ be a point such that $\ell(q_\tau, \tau) \leq \frac{n}{2}$. The existence of q_τ is guaranteed by (3.28f). For any $\tau > 0$, we define solutions to the backward Ricci flow by parabolic scaling:

$$g_\tau(\theta) = \tau^{-1} \cdot g(\tau\theta), \quad \text{for } \theta \in [0, \infty). \tag{5.1}$$

The following is Proposition 11.2 in [Pe02].

Theorem 5.6. *For any sequence* $\tau_i \to \infty$, *there exists a subsequence still denoted by* τ_i, *such that* $(M^n, g_{\tau_i}(\theta), (q_{\tau_i}, 1))$, $\theta \in (0, \infty)$, *converges in the Cheeger-Gromov sense to a complete nonflat shrinking gradient Ricci soliton* $(M^n_\infty, g_\infty(\theta), (q_\infty, 1))$.

Below we give a **sketch** of the proof of Theorem 5.6.

5.2B. Estimating reduced distance associated to κ-solutions. The **Hamilton's trace Harnack inequality** says that for a backward Ricci flow solution on $[0,T)$ with nonegative curvature operator, the trace Harnack quantity as defined in (3.22) satisfies

$$H(X)(x,\tilde{\tau}) \geq -\left(\frac{1}{\tilde{\tau}}+\frac{1}{T-\tilde{\tau}}\right)R(x,\tilde{\tau}). \tag{5.2}$$

Here for $g(\tau)$ we can take $T=\infty$. By (3.23) we have

$$K \geq -\int_0^{\tau} \tilde{\tau}^{3/2}\cdot\tilde{\tau}^{-1}R(\gamma(\tilde{\tau}),\tilde{\tau})d\tilde{\tau} \geq -\mathcal{L}(\gamma) = -2\sqrt{\tau}\ell(x,\tau).$$

Plug this into (3.27b), we have the following estimate

$$|\nabla\ell(x,\tau)|^2 + R(x,\tau) \leq \frac{3\ell(x,\tau)}{\tau} \tag{5.3}$$

for $g(\tau)$ coming out of the κ-solution $\tilde{g}(t)$. From (3.27a) and (3.27b) we have

$$\frac{\partial\ell}{\partial\tau} = -\frac{1}{2}|\nabla\ell|^2 - \frac{\ell}{2\tau} + \frac{1}{2}R,$$

hence from (5.3) it follows

$$\left|\frac{\partial\ell}{\partial\tau}\right| \leq \frac{2\ell}{\tau}. \tag{5.4}$$

Lemma 5.7. *Given any $\varepsilon>0$ and $A>1$, there exists $\delta(n,\varepsilon,A)>0$ such that for any $\tau>0$,*

$$\ell(x,\tilde{\tau}) \leq \delta(n,\varepsilon,A)^{-1} \quad \textit{and} \quad \tilde{\tau}R(x,\tilde{\tau}) \leq \delta(n,\varepsilon,A)^{-1}$$

for all $(x,\tilde{\tau}) \in B_{g(\tau)}\left(q_\tau, \sqrt{\varepsilon^{-1}\tau}\right)\times\left[A^{-1}\tau, A\tau\right]$.

Sketch of the proof of Lemma 5.7. Note that the estimate of $\tilde{\tau}R(x,\tilde{\tau})$ follows from the estimate of $\ell(x,\tilde{\tau})$ and (5.3). To see the estimate of $\ell(x,\tilde{\tau})$, by (5.3) we have $|\nabla\sqrt{\ell(x,\tau)}|_{g(\tau)} \leq \frac{\sqrt{3}}{2}\tau^{-1/2}$, combining with $\ell(q_\tau,\tau)\leq\frac{n}{2}$ we get an estimate of $\ell(x,\tau)$. The estimate of $\ell(x,\tilde{\tau})$ then follows from (5.4).

5.2C. The existence of the limit in Theorem 5.6. Fix an $A>1$, for the sequence $\tau_i\to\infty$ in Theorem 5.6, we consider the sequence of pointed backward Ricci flow solutions

$$\left(M^n, g_{\tau_i}(\theta), (q_{\tau_i},1)\right), \quad \theta\in\left[A^{-1},A\right].$$

For any $\varepsilon>0$, after parabolic scaling of $g(\tau)$ by τ_i, Lemma 5.7 yields the curvature bound

$$\left|\mathrm{Rm}_{g_{\tau_i}}(x,\theta)\right| \leq R_{g_{\tau_i}}(x,\theta) \leq \theta^{-1}\delta(n,\varepsilon,A)^{-1} \leq A\delta(n,\varepsilon,A)^{-1} \tag{5.5}$$

on $B_{g_{\tau_i}(1)}\left(q_{\tau_i},\sqrt{\varepsilon^{-1}}\right)\times\left[A^{-1},A\right]$, here we have used that $g(\tau)$ has nonnegative curvature operator. In particular taking $\varepsilon=1$ and $A=2$, we obtain that for some $\delta(n,1,2)<1$

$$\left|\mathrm{Rm}_{g_{\tau_i}}(x,1)\right| \leq 2\delta(n,1,2)^{-1} \quad \text{for } x\in B_{g_{\tau_i}(1)}(q_{\tau_i},1).$$

Since $g(\theta)$ is κ-noncollapsed on all scales, we have $g_{\tau_i}(\theta)$ is κ-noncollapsed on $B_{g_{\tau_i}(1)}(q_{\tau_i}, \sqrt{\delta(n,1,2)/2})$ and hence

$$\mathrm{Vol}_{g_{\tau_i}(1)} B_{g_{\tau_i}(1)}\left(q_{\tau_i}, \sqrt{\delta(n,1,2)/2}\right) \geq \kappa\left(\sqrt{\delta(n,1,2)/2}\right)^n.$$

By a theorem of Cheeger, Gromov and Taylor we have the injectivity radius estimate

$$\mathrm{inj}_{g_{\tau_i}(1)}(q_{\tau_i}) \geq \delta_1(n,\kappa) \tag{5.6}$$

for some positive constant $\delta_1(n,\kappa)$ depending only on n and κ.

(5.5) and (5.6) enable us to apply Hamilton's Cheeger-Gromov-type compactness theorem to the sequence of solutions $g_{\tau_i}(\theta)$ of the backward Ricci flow to get a convergent subsequence

$$(M^n, g_{\tau_i}(\theta), (q_{\tau_i}, 1)) \longrightarrow (M^n_\infty, g_\infty(\theta), (q_\infty, 1)) \quad \text{for } \theta \in \left[A^{-1}, A\right]. \tag{5.7}$$

The limit $g_\infty(\theta)$ is a complete solution to the backward Ricci flow. Since each $g_{\tau_i}(\theta)$ satisfies the trace Harnack inequality, $g_\infty(\theta)$ satisfies the trace Harnack inequality (5.2) with $T = \infty$. Also $g_\infty(\theta)$ is κ-noncollapsed on all scales, has nonnegative curvature operator, and satisfies $\mathrm{inj}_{g_\infty(1)}(q_\infty) \geq \delta_1(n,\kappa)$. However because the curvature bound $A\delta(n,\varepsilon,A)^{-1}$ in (5.5) depends on ε and hence on the radius $\sqrt{\varepsilon^{-1}}$, we may not have curvature bound $\sup_{M_\infty} |\mathrm{Rm}_{g_\infty}(x,\theta)| < \infty$ for each $\theta \in [A^{-1}, A]$.

By choosing a sequence of $A_k \to \infty$ and using a diagonalization argument, we may assume that $(M^n_\infty, g_\infty(\theta))$ exists for $\theta \in (0,\infty)$ and that the convergence in (5.7) holds for $\theta \in (0,\infty)$.

5.2D. Finishing the proof of Theorem 5.6. To finish the proof of Theorem 5.6, we need to show that for each θ, $g_\infty(\theta)$ is a nonflat shrinking gradient Ricci soliton. Let $\ell_i(x,\theta)$ denote the reduced distance of the solution $g_{\tau_i}(\theta)$ with respect to the basepoint $(p,0)$. After scaling, (5.3) and (5.4) give the derivative estimates of $\ell_i(x,\theta)$, by Arzela-Ascoli theorem some subsequence $\ell_i(x,\theta)$ converges to a Lipschitz function $\ell_\infty(x,\theta)$ on M_∞ in the Cheeger-Gromov sense. Similar to the definition of reduced volume (4.1), we use $\ell_\infty(x,\theta)$ to define

$$\hat{V}_\infty(\theta) = \int_{M_\infty} (4\pi\theta)^{-n/2} \exp\left[-\ell_\infty(x,\theta)\right] d\mu_{g_\infty(\theta)}(x), \qquad \theta \in (0,\infty). \tag{5.8}$$

By certain estimates on the growth of reduced distance $\ell(x,\tau)$ (not covered in §3), one can prove the convergence of the reduced volume (defined above (5.1))

$$\lim_{i\to\infty} \tilde{V}(\tau_i\theta) = \hat{V}_\infty(\theta) \qquad \text{for each } \theta > 0.$$

The monotonicity of reduced volume then implies that $\hat{V}_\infty(\theta)$ is a constant function. Combining this and (3.28a) for $\ell_i(x,\theta)$, one can argue that

$$\frac{\partial \ell_\infty}{\partial\theta} - \Delta_{g_\infty}\ell_\infty + |\nabla_{g_\infty}\ell_\infty|^2 - R_{g_\infty} + \frac{n}{2\theta} = 0 \tag{5.9}$$

in weak sense (we suggest the reader to find the details of the argument in the literature). Regularity theory of parabolic partial differential equations implies that ℓ_∞ is a smooth function.

Define two functions on $M_\infty \times (0,\infty)$ by $u_\infty(x,\theta) = (4\pi\theta)^{-\frac{n}{2}} e^{-\ell_\infty(x,\theta)}$ and

$$v_\infty = \left(\theta(2\Delta_{g_\infty}\ell_\infty - |\nabla_{g_\infty}\ell_\infty|^2 + R_{g_\infty}) + \ell_\infty - n\right) u_\infty.$$

Let operator $\Box^* = \frac{\partial}{\partial\theta} - \Delta_{g_\infty} + R_{g_\infty}$. Equation (5.9) implies that $\Box^* u_\infty = 0$. By some calculation one can show

$$\Box^* v_\infty = -2\theta \left| \mathrm{Rc}\,(g_\infty) + \nabla_{g_\infty}\nabla_{g_\infty}\ell_\infty - \frac{1}{2\theta} g_\infty \right|^2 u_\infty. \tag{5.10}$$

Applying (3.28d) to $l_i(x,\theta)$ we have

$$2\frac{\partial \ell_i}{\partial\theta} + \left|\nabla_{g_{\tau_i}}\ell_i\right|^2 - R_{g_{\tau_i}} + \frac{\ell_i}{\theta} = 0.$$

It can be argued that when $i \to \infty$ the above equality implies

$$2\frac{\partial \ell_\infty}{\partial\theta} + |\nabla_{g_\infty}\ell_\infty|^2 - R_{g_\infty} + \frac{\ell_\infty}{\theta} = 0.$$

Combining this with (5.9) we get $v_\infty = 0$, and hence it follows from (5.10) that

$$\mathrm{Rc}\,(g_\infty) + \nabla_{g_\infty}\nabla_{g_\infty}\ell_\infty - \frac{1}{2\theta} g_\infty = 0. \tag{5.11}$$

We have proved that $g_\infty(\theta)$ is a shrinking gradient Ricci soliton.

The last part that $g_\infty(\theta)$ is nonflat, is argued by contradiction. If it is flat, then the soliton equation (5.11) gives enough information of g_∞ and ℓ_∞ (Euclidean shrinking solution) to conclude that $\hat{V}_\infty(\theta) = 1$. But the equation above (5.9) implies that $\hat{V}_\infty(\theta) = \lim_{\tau\to\infty}\tilde{V}(\tau) < 1$. We get a contradiction. Now we have finished the sketch of the proof of Theorem 5.6.

Acknowledgement: In preparing these notes the author benefits a lot from the book "The Ricci flow: Techniques and Applications, Part I", written by Bennett Chow, Sun-Chin Chu, David Glickenstein, Christine Guenther, Jim Isenberg, Tom Ivey, Dan Knopf, Peng Lu, Feng Luo, and Lei Ni.

The author thanks the organizers of 15th Gökova Geometry/Topology Conference (Turkey, May, 2008) for a wonderful conference.

References

[Pe02] G. Perelman, *The entropy formula for the Ricci flow and its geometric applications.* arXiv: math.DG/0211159

Department of Mathematics, University of Oregon, Eugene, OR 97403
E-mail address: penglu@uoregon.edu

Proceedings of 15$^{\text{th}}$ Gökova
Geometry-Topology Conference
pp. 31–63

Resolution of singularities of analytic spaces

Jarosław Włodarczyk

Abstract. Building upon work of Villamayor Bierstone-Milman and our recent paper we give a proof of the canonical Hironaka principalization and desingularization of analytic spaces. Though the inductive scheme of the proof is the same as in algebraic case there is a number of technical differences between analytic and algebraic situation.

1. Introduction

In the present paper we give a short proof of the Hironaka theorem on resolution of singularities of analytic spaces. The structure of the proof and its organization is very similar with the one given in the paper [38].

The strategy of the proof we formulate here is essentially the same as the one found by Hironaka and simplified by Bierstone-Milman and Villamayor ([8], [9], [10]), ([35], [36], [37]). In particular we apply here one of Villamayor's key simplifications, eliminating the use of the Hilbert-Samuel function and the notion of normal flatness (see [13]).

The main idea of the algorithm is to control the resolution procedure by two simple invariants: order of the weak transform of the ideal sheaf $\mathcal{I}$ and the dimension of the ambient manifold M. The process of dropping the order starts from the isolating the "worst singularity locus" -the set where the order is maximal $\operatorname{ord}_x(\mathcal{I}) = \mu$. This leads to considerations of ideal sheaves with assigned order $(\mathcal{I}, \mu)$.

Eliminating "worst singularity locus" $\operatorname{supp}(\mathcal{I}, \mu)$ builds upon reduction of the dimension of the ambient variety. It was observed by Abhyankhar and successfully implemented by Hironaka that $\operatorname{supp}(\mathcal{I}, \mu)$ is contained in a certain smooth hypersurface M' of M. The concept of hypersurface of maximal contact can be expressed nicely by using Giraud approach with derivations.

The blow-ups used for eliminating $\operatorname{supp}(\mathcal{I}, \mu)$ are performed only at centers which are contained in $\operatorname{supp}(\mathcal{I}, \mu)$. This has two major consequences:

1. The outside of the locus $\operatorname{supp}(\mathcal{I}, \mu)$ can be ignored in the process. Thus $(\mathcal{I}, \mu)$ can be considered as a "part of the ideal sheaf of $\mathcal{I}$ where the order is $\geq \mu$". Solving of $(\mathcal{I}, \mu)$ is merely eliminating $\operatorname{supp}(\mathcal{I}, \mu)$.

Key words and phrases. resolution of singularities, analytic spaces, sheaves of ideals.
The author was supported in part by NSF grant grant DMS-0500659 and Polish KBN grant GR-1784.

2. The total transform of ideal is divisible by μ-power of exceptional divisor. Thus the transformation of the ideal $\mathcal{I}$ can be described by explicit formula:

$$\sigma^c(\mathcal{I},\mu) = \mathcal{I}(E)^{-\mu}\sigma^*(\mathcal{I}).$$

This makes a basis for the reduction to the hypersuface of maximal contact. Although it is not possible to restrict $\mathcal{I}$ directly to $M' \subset M$ we can find an ideal sheaf $(\mathcal{I}',\mu')$, called "coefficient ideal ", which lives on M', and which is related to $(\mathcal{I},\mu)$ by the equality

$$\mathrm{supp}(\mathcal{I},\mu) = \mathrm{supp}(\mathcal{I}',\mu').$$

Now the problem of eliminating "bad locus" $\mathrm{supp}(\mathcal{I},\mu)$ is reduced to the lower dimension where we proceed by induction.

This approach has a major flaw. The procedure of restricting $\mathcal{I}$ to the hypersurface of maximal contact is not canonical and is defined locally. In fact for two different hypersurfaces of maximal contact we get two different objects which are loosely related. In order to resolve this issue Hironaka used the following approach: The local resolutions can be encoded by a certain invariant. Each single operation used in the above mentioned induction leaves its "trace" which is a single entry of the invariant. As a result the invariant is a sequence of the numbers occuring in local resolutions. The invariant is upper semicontinuous and defines a stratification of the ambient space. This invariant drops after the blow-up of the maximal stratum. It determines the centers of the resolution and allows one to patch up local desingularizations to a global one. What adds to the complexity is that the invariant is defined within some rich inductive scheme encoding the desingularization and assuring its canonicity (Bierstone-Milman's towers of local blow-ups with *admissible centers* and Villamayor's *general basic objects*) (see also Encinas-Hauser [17]).

Instead of considering the invariant as the key notion of the algorithm, in [38] we proposed a different approach. It is based upon two simple observations.

(1) The resolution process defined as a sequence of suitable blow-ups of ambient spaces can be applied simultaneously not only to the given singularities but rather to a class of equivalent singularities obtained by simple arithmetical modifications. This means that we can "tune" singularities before resolving them.
(2) In the equivalence class we can choose a convenient representative given by the *homogenized ideals* introduced in the paper. The restrictions of homogenized ideals to different hypersurfaces of maximal contact define locally analytically isomorphic singularities. Moreover the local isomorphism of hypersurfaces of maximal contact is defined by a local analytic automorphism of the ambient space preserving all the relevant resolutions.

"Homogenization" of the ideal makes the operation of restriction to hypersurface of maximal contact canonical- independent of any choices. In particular there is no necessity of describing and comparing local algorithms. The inductive structure of the process is reduced to the existence of a canonical functorial resolution in lower dimensions. This approach puts much less emphasis on the invariant. In fact as was observed by Kollár by

mere allowing reducible algebraic varieties (or analytic spaces) in the inductive scheme one eliminates the "long" invariant completely ([31]). What is left is a "bare" two- step induction.

In **Step 2** of the proof, given an ideal $(\mathcal{I}, \mu)$ we assign to it the worst singularity order μ'. Instead of dealing with $(\mathcal{I}, \mu)$ directly we form an auxiliary ideal (companion ideal) which is roughly $(\mathcal{I}, \mu')$. Its resolution determines the drop of the maximal order of the weak transform (nonmonomial part) of $\mathcal{I}$. By repeating this process sufficiently many times the weak transform of $(\mathcal{I}, \mu)$ disappear and $(\mathcal{I}, \mu)$ becomes principal monomial thus, easy to solve directly. The procedure in Step 2 uses the fact that companion ideals and, in general, all ideals $(\mathcal{I}, \mu')$, where $\mu' = \max\{\mathrm{ord}_x(\mathcal{I}) \mid x \in M\}$ are possible to solve by reduction to the hypersurface of maximal contact. This is done in **Step 1** of the proof. That's where the operation of tuning comes handy. The "tuning" of ideals has two aspects. First, homogenization gives us the canonicity of resolution and solves the glueing problem. Second, we can view a coefficient ideal as a part of the tuning too. In this approach coefficient ideal $\mathcal{C}(\mathcal{I}, \mu)$ lives on M and is equivalent to $\mathcal{I}$ but its "restricts well" not only to the hypersurface of maximal contact but to any smooth subvariety $Z \subset M$, that is,

$$\mathrm{supp}(\mathcal{C}(\mathcal{I}), \mu) \cap Z = \mathrm{supp}(\mathcal{C}(\mathcal{I})_{|Z}, \mu')$$

In the analytic situation, considered in the paper, in the algorithm of resolution of $(\mathcal{I}, \mu)$ the compactness condition is essential. In particular isolating "the worst singularity" locus is possible only under the assumption of compactness. Even if we start our considerations from ideal sheaves on compact manifolds the operation of local restriction to hypersurface of maximal contact leads to noncompact submanifolds. That is why in the analytic case it is natural to consider not manilfolds or compact manifolds but rather germs of manifolds at compact subsets. After establishing a few technical differences between analytic and algebraic case we can carry the inductive algorithm essentially in the same way as in the algebraic case. As a result we construct a resolution which is locally but not globally a sequence of blow-ups at smooth centers.

The presented proof is elementary, constructive and self-contained.

The paper is organized as follows. In section 1 we formulate three main theorems: the theorem of canonical principalization (Hironaka's "Desingularization II"), the theorem of canonical embedded resolution (a slightly weaker version of Hironaka's "Desingularization I") and the theorem of canonical resolution. In section 2 we introduce basic notions we are going to use throughout the paper. In section 3 we formulate the theorem of canonical resolution of marked ideals and show how it implies three main theorems (Hironaka's resolution principle). Section 4 gives important technical ingredients. In particular we introduce here the notion of homogenized ideals. In section 5 we formulate the resolution algorithm and prove the theorem of canonical resolution of marked ideals. In section 6 we make final conclusions from the proof.

1.1. Acknowledgements

The author would like to express his gratitude to E. Bierstone, A. Gabrielov, H. Hauser, Y. Kawamata, J. Kollár, J. Lipman, K. Matsuki, P. Milman, O.Villamayor and J. Wiśniewski for helpful comments.

2. Formulation of the main theorems

All analytic spaces in this paper are defined over a ground field $\mathbf{K} = \mathbb{C}$ or $\mathbb{R}$. We give a proof of the following Hironaka Theorems (see [26]):

Canonical resolution of singularities

Theorem 2.0.1. *Let Y be an analytic space. There exists a canonical desingularization of Y that is a manifold $\widetilde{Y}$ together with a proper bimeromorphic morphism $\mathrm{res}_Y : \widetilde{Y} \to Y$ such that*

(1) *$\mathrm{res}_Y : \widetilde{Y} \to Y$ is an isomorphism over the nonsingular part Y_{ns} of Y.*
(2) *The inverse image of the singular locus $\mathrm{res}_Y^{-1}(Y_{\mathrm{sing}})$ is a simple normal crossing divisor.*
(3) *res_Y is functorial with respect to local analytic isomorphisms. For any local analytic isomorphism $\phi : Y' \to Y$ there is a natural lifting $\widetilde{\phi} : \widetilde{Y'} \to \widetilde{Y}$ which is a local analytic isomorphism.*

Locally finite embedded desingularization

Theorem 2.0.2. *Let Y be an analytic subspace of an analytic manifold M. There exists a manifold $\widetilde{M}$, a simple normal crossing locally finite divisor E on $\widetilde{M}$, and a bimeromorphic proper morphism*

$$\mathrm{res}_{Y,M} : \widetilde{M} \to M$$

such that the strict transform $\widetilde{Y} \subset \widetilde{M}$ is smooth and have simple normal crossings with the divisor E. The support of the divisor E is the the exceptional locus of $\mathrm{res}_{Y,M}$. The morphism $\mathrm{res}_{Y,M}$ locally factors into a sequence of blow-ups at smooth centers. That is, for any compact set $Z \subset Y$ there is an open subset $U \subset M$ and $\widetilde{U} = \mathrm{res}_{Y,M}^{-1}(U) \subset \widetilde{M}$ and a sequence

$$U_0 = U \overset{\sigma_{U1}}{\longleftarrow} U_1 \overset{\sigma_{U2}}{\longleftarrow} U_2 \longleftarrow \ldots \longleftarrow U_i \longleftarrow \ldots \longleftarrow U_r = \widetilde{U} \quad (*)$$

of blow-ups $\sigma_{Ui} : U_{i-1} \longleftarrow U_i$ with smooth closed centers $C_{i-1} \subset U_{i-1}$ such that

(1) *The exceptional divisor E_{Ui} of the induced morphism $\sigma_U^i = \sigma_{U1} \circ \ldots \circ \sigma_{Ui} : U_i \to U$ has only simple normal crossings and C_i has simple normal crossings with E_i.*
(2) *Let $Y_{Ui} := Y \cap U_i$ be the strict transform of Y. All centers C_i are disjoint from the set $\mathrm{Reg}(Y) \subset Y_i$ of points where Y (not Y_i) is smooth (and are not necessarily contained in Y_i).*

(3) *The strict transform* $Y_{Ur} = \widetilde{Y} \cap U_r$ *of* $Y_U := Y \cap U$ *is smooth and has only simple normal crossings with the exceptional divisor* E_r.

(4) *The morphism* $\mathrm{res}_{Y,M} : (M, Y) \leftarrow (\widetilde{M}, \widetilde{Y})$ *defined by the embedded desingularization commutes with local analytic isomorphisms, embeddings of ambient varieties.*

(5) *For any compact sets* $Z_1 \subset Z_2$ *and corresponding open neighborhoods* $U_1 \subset U_2$ *the restriction of the factorization (*) of* $\mathrm{res}_{Y,M|\widetilde{U_2}} : \widetilde{U}_2 \to U_2$ *to* $\widetilde{U_1}$ *determines the factorization of* $\mathrm{res}_{Y,M|\widetilde{U_1}} : \widetilde{U}_1 \to U_1$.

(6) *(Strengthening of Bravo-Villamayor* [13]*)*

$$\sigma^*(\mathcal{I}_Y) = \mathcal{I}_{\widetilde{Y}} \mathcal{I}_{\widetilde{E}},$$

where $\mathcal{I}_{\widetilde{Y}}$ *is the sheaf of ideals of the subvariety* $\widetilde{Y} \subset \widetilde{M}$ *and* $\mathcal{I}_{\widetilde{E}}$ *is the sheaf of ideals of a simple normal crossing divisor* $\widetilde{E}$ *which is a locally finite combination of the irreducible components of the divisor* E_{Ur}.

Locally finite principalization of sheaves of ideals

Theorem 2.0.3. *Let* $\mathcal{I}$ *be a sheaf of ideals on a analytic manifold* M *(not necessarily compact). There exists a locally finite principalization of* $\mathcal{I}$*, that is, a manifold* $\widetilde{M}$*, a proper morphism* $\mathrm{prin}_{\mathcal{I}} : \widetilde{M} \to M$*, and a sheaf of ideals* $\widetilde{\mathcal{I}}$ *on* M *such that*

(1) *For any compact set* $Z \subset M$*, there is an open neighborhoods* $U \supset Z$ *and* $\widetilde{U} := \mathrm{prin}_{\mathcal{I}}^{-1}(U) \subset \widetilde{M}$ *for which the restriction* $\mathrm{prin}_{\mathcal{I}|\widetilde{U}} : \widetilde{U} \to U$ *splits into a finite sequence of blow-ups*

$$U = U_0 \overset{\sigma_{U1}}{\longleftarrow} U_1 \overset{\sigma_{U2}}{\longleftarrow} U_2 \longleftarrow \ldots \longleftarrow U_i \longleftarrow \ldots \longleftarrow U_r = \widetilde{U} \qquad (*)$$

of blow-ups $\sigma_{Ui} : U_{i-1} \leftarrow U_i$ *with smooth centers* $C_{i-1} \subset U_{i-1}$ *such that*

(2) *The exceptional divisor* E_{Ui} *of the induced morphism* $\sigma^i = \sigma_1 \circ \ldots \circ \sigma_i : U_i \to U$ *has only simple normal crossings and* C_i *has simple normal crossings with* E_i.

(3) *The total transform* $\mathrm{prin}^*_{\mathcal{I}|\widetilde{U}}(\mathcal{I}) = \sigma^{r*}(\mathcal{I})$ *is the ideal of a simple normal crossing divisor* $\widetilde{E}_U$ *which is a locally finite combination of the irreducible components of the divisor* E_{Ur}.

(4) *For any compact sets* $Z_1 \subset Z_2$ *and corresponding open neighborhoods* $U_1 \subset U_2$ *the restriction of the factorization (*) of* $\mathrm{prin}_{\mathcal{I}|\widetilde{U_2}} : \widetilde{U}_2 \to U_2$ *to* $\widetilde{U_1}$ *determines the factorization of* $\mathrm{prin}_{\mathcal{I}|\widetilde{U_1}} : \widetilde{U}_1 \to U_1$.

The morphism $\mathrm{prin} : (\widetilde{M}, \widetilde{\mathcal{I}}) \to (M, \mathcal{I})$ *commutes with local analytic isomorphisms, embeddings of ambient varieties.*

Remarks. (1) By the exceptional divisor of the blow-up $\sigma : M' \to M$ with a smooth center C we mean the inverse image $E := \sigma^{-1}(C)$ of the center C. By the exceptional divisor of the composite of blow-ups σ_i with smooth centers C_{i-1} we mean

the union of the strict transforms of the exceptional divisors of σ_i. This definition coincides with the standard definition of the exceptional set of points of the bimeromorphic morphism in the case when $\operatorname{codim}(C_i) \geq 2$ (as in Theorem 2.0.2). If $\operatorname{codim}(C_{i-1}) = 1$ the blow-up of C_{i-1} is an identical isomorphism and defines a formal operation of converting a subvariety $C_{i-1} \subset M_{i-1}$ into a component of the exceptional divisor E_i on M_i. This formalism is convenient for the proofs. In particular it indicates that C_{i-1} identified via σ_i with a component of E_i has simple normal crossings with other components of E_i.

(2) In the Theorem 2.0.2 we blow up centers of codimension ≥ 2 and both definitions coincide.

(3) Given a closed embedding of manifolds $i : M \hookrightarrow M'$, the coherent sheaf of ideals $\mathcal{I}$ on M defines a coherent subsheaf $i_*(\mathcal{I}) \subset i_*(\mathcal{O}_M)$ of $\mathcal{O}_{M'}$-module $i_*(\mathcal{O}_M)$. Let $i^\sharp : O_{M'} \to i_*(\mathcal{O}_M)$ be the natural surjection of $\mathcal{O}_{M'}$-modules. The inverse image $\mathcal{I}' = (i^\sharp)^{-1}(i_*(\mathcal{I}))$ defines a coherent sheaf of ideals on M'. By abuse of notation $\mathcal{I}'$ will be denoted as $i_*(\mathcal{I}) \cdot O_{M'}$.

3. Preliminaries

3.1. Germs of analytic spaces at compact subsets

Definition 3.1.1. Let M be an analytic space and $Z \subset M$ be a compact subset. By *a representative of germ* M_Z of M at Z we mean a pair (U, Z) where $U \subset M$ is any open subset of M containing Z. We say that for any two open subsets U, U' of M containing Z the representative of germs (U, Z), and (U', Z) define the same *germ* M_Z. We write $M_Z = (U, Z)$ and call U a *neighborhood* of a germ M_Z. By a *morphism* $f : M_Z \to M'_{Z'}$ we mean a morphism $f_U : U \to U'$ between some neighborhoods of M_Z and $M'_{Z'}$ such that $f(Z) \subset Z'$. The morphism f is proper, projective, (resp. is an open or closed inclusion) if f_U has this property for the corresponding neighborhoods U, U'.

We introduce the operation of union and intersection of germs : If $U, U' \subset M$ then

$$(U, Z) \cup (U', Z') := (U \cup U', Z \cup Z'), \quad (U, Z) \cap (U', Z') := (U \cap U', Z \cap Z')$$

Then $(U, Z) \to (U, Z) \cup (U', Z')$ and $(U, Z) \cap (U', Z') \to (U, Z)$ are open inclusions.

3.2. Resolution of marked ideals

We shall consider ideal sheaves and divisors on germs M_Z. If $U \subset M$ is a smooth open subset containing Z then we call the germ $M_Z = (U, Z)$ *smooth*. A sheaf of ideal on M_Z is a sheaf $\mathcal{I}$ on some neighborhood U of M_Z. For any sheaf of ideals $\mathcal{I}$ on a smooth germ $M_Z = (U, Z)$ and any point $x \in U$ we denote by

$$\operatorname{ord}_x(\mathcal{I}) := \max\{i \mid \mathcal{I}_x \subset m_x^i\}$$

the *order* of $\mathcal{I}$ at x. (Here m_x denotes the maximal ideal of x.)

Definition 3.2.1. (Hironaka [26], [28], Bierstone-Milman [8], Villamayor [35]) A *marked ideal* is a collection $(M_Z, \mathcal{I}, E, \mu)$, where M_Z is a smooth germ, $\mathcal{I}$ is a sheaf of ideals

on M_Z, μ is a nonnegative integer and E is a totally ordered collection of divisors on M_Z whose irreducible components are pairwise disjoint and all have multiplicity one. Moreover the irreducible components of divisors in E have simultaneously simple normal crossings.

Let $(M_Z, \mathcal{I}, E, \mu)$ be a marked ideal such that the ideal sheaf $\mathcal{I}$ is defined on an open neighborhood U of M_Z. One can show that the set

$$\mathrm{supp}_Z(M_Z, \mathcal{I}, E, \mu) := \{x \in Z \mid \mathrm{ord}_x(\mathcal{I}) \geq \mu\}$$

is compact. On the other hand the set

$$\mathrm{supp}_U(M_Z, \mathcal{I}, E, \mu) := \{x \in U \mid \mathrm{ord}_x(\mathcal{I}) \geq \mu\}$$

defines a closed analytic subspace of U. (see Lemma 5.2.2).

Definition 3.2.2. (Hironaka [26], [28], Bierstone-Milman [8], Villamayor [35]) By the *support* (originally *singular locus*) of $(M_Z, \mathcal{I}, E, \mu)$ we mean the germ of analytic space

$$\mathrm{supp}(M_Z, \mathcal{I}, E, \mu) := (\mathrm{supp}_U(M_Z, \mathcal{I}, E, \mu), \mathrm{supp}_Z(M_Z, \mathcal{I}, E, \mu)),$$

Remarks. (1) The ideals with assigned orders or functions with assigned multiplicities and their supports are key objects in the proofs of Hironaka, Villamayor and Bierstone-Milman. In particular Hironaka introduced the notion of *idealistic exponent*.

(2) To simplify notation we often write marked ideals $(M_Z, \mathcal{I}, E, \mu)$ as couples $(\mathcal{I}, \mu)$ or even ideals $\mathcal{I}$.

(3) For any sheaf of ideals $\mathcal{I}$ on $M_Z = (U, Z)$ we have

$$\mathrm{supp}(\mathcal{I}, 1) = V(\mathcal{I}) := \{x \in U \mid f(x) = 0, \quad \text{for any } f \in \mathcal{I}\}.$$

Definition 3.2.3. Let M_Z be a germ of an analytic manifold M. Let $C \subset U$ be a smooth closed subspace of a neighborhood $U \subset Z$. Let $\sigma_U : U' \to U$ denote the blow-up of a smooth center C. Set $Z' := \sigma_U^{-1}(Z)$, $M'_{Z'} := (U', Z')$. The germ of σ_U is a bimeromorphic morphism $\sigma : M'_{Z'} \to M_Z$ which is called a *blow-up* of M_Z at the center $C \subset M_Z$.

Definition 3.2.4. (Hironaka [26], [28], Bierstone-Milman [8], Villamayor [35]) By a *resolution* of $(M_Z, \mathcal{I}, E, \mu)$ we mean a sequence of blow-ups $\sigma_i : M_{i,Z_i} \to M_{i-1,Z_{i-1}}$ of disjoint unions of smooth centers $C_{i-1} \subset M_{i-1}$,

$$M_{0,Z_0} \xleftarrow{\sigma_1} M_{1,Z_1} \xleftarrow{\sigma_2} M_{2,Z_2} \xleftarrow{\sigma_3} \dots M_{i,Z_i} \longleftarrow \dots \xleftarrow{\sigma_r} M_{r,Z_r},$$

which defines a sequence of marked ideals $(M_{i,Z_i}, \mathcal{I}_i, E_i, \mu)$ where

(1) $C_i \subset \mathrm{supp}(M_{i,Z_i}, \mathcal{I}_i, E_i, \mu)$.

(2) C_i has simple normal crossings with E_i.

(3) $\mathcal{I}_i = \mathcal{I}(D_i)^{-\mu}\sigma_i^*(\mathcal{I}_{i-1})$, where $\mathcal{I}(D_i)$ is the ideal of the exceptional divisor D_i of σ_i.

(4) $E_i = \sigma_i^{\mathrm{c}}(E_{i-1}) \cup \{D_i\}$, where $\sigma_i^{\mathrm{c}}(E_{i-1})$ is the set of strict transforms of divisors in E_{i-1}.

(5) The order on $\sigma_i^c(E_{i-1})$ is defined by the order on E_{i-1} while D_i is the maximal element of E_i.
(6) $\mathrm{supp}(M_{r,Z_r}, \mathcal{I}_r, E_r, \mu) = \emptyset$.

Remark. Note that the resolution of $(M_Z, \mathcal{I}, E, \mu)$ coincides with the resolution of $(M_{Z'}, \mathcal{I}, E, \mu)$, where $Z' := Z \cap \mathrm{supp}(\mathcal{I}, \mu)$ so we can assume that

$$Z \subset \mathrm{supp}(\mathcal{I}, \mu).$$

Definition 3.2.5. The sequence of morphisms which are either isomorphisms or blow-ups satisfying conditions (1)-(5) is called a *multiple test blow-up*. The number of morphisms in a multiple test blow-up will be called its *length*.

Definition 3.2.6. An *extension* of a sequence of blow-ups $(M_{iZ_i})_{0\leq i\leq m}$ is a sequence $(M'_{jZ_j})_{0\leq j\leq m'}$ of blow-ups and isomorphisms $M'_{0Z_0} = M'_{j_0 Z_{j_0}} = \ldots = M'_{j_1-1,Z_{j_1-1}} \leftarrow M'_{j_1} = \ldots = M'_{j_2-1,Z_{j_2-1}} \leftarrow \ldots M'_{j_m,Z_{j_m}} = \ldots = M'_{m'}$, where $M'_{j_i Z_{j_i}} = M_{iZ_i}$.

In particular we shall consider *extensions of multiple test blow-ups.*

Remarks. (1) The definition of extension arises naturally when we pass to open subsets of the considered ambient manifold M.
(2) The notion of a *multiple test blow-up* is analogous to the notions of *test* or *admissible* blow-ups considered by Hironaka, Bierstone-Milman and Villamayor.

3.3. Transforms of marked ideals and controlled transforms of functions

In the setting of the above definition we shall call

$$(\mathcal{I}_i, \mu) := \sigma_i^c(\mathcal{I}_{i-1}, \mu)$$

a *transform of the marked ideal* or *controlled transform* of $(\mathcal{I}, \mu)$. It makes sense for a single blow-up in a multiple test blow-up as well as for a multiple test blow-up. Let $\sigma^i := \sigma_1 \circ \ldots \circ \sigma_i : M_i \to M$ be a composition of consecutive morphisms of a multiple test blow-up. Then in the above setting

$$(\mathcal{I}_i, \mu) = (\sigma^i)^c(\mathcal{I}, \mu).$$

We shall also denote the controlled transform $(\sigma^i)^c(\mathcal{I}, \mu)$ by $(\mathcal{I}, \mu)_i$ or $[\mathcal{I}, \mu]_i$.

The controlled transform can also be defined for local sections $f \in \mathcal{I}(U)$. Let $\sigma : M \leftarrow M'$ be a blow-up with a smooth center $C \subset \mathrm{supp}(\mathcal{I}, \mu)$ defining a transformation of marked ideals $\sigma^c(\mathcal{I}, \mu) = (\mathcal{I}', \mu)$. Let $f \in \mathcal{I}(U)$ be a section of a sheaf of ideals. Let $U' \subseteq \sigma^{-1}(U)$ be an open subset for which the sheaf of ideals of the exceptional divisor is generated by a function y. The function

$$g = y^{-\mu}(f \circ \sigma) \in \mathcal{I}(U')$$

is a *controlled transform* of f on U' (defined up to an invertible function). As before we extend it to any multiple test blow-up.

The following lemma shows that the notion of controlled transform is well defined.

Lemma 3.3.1. *Let $C \subset \mathrm{supp}(\mathcal{I}, \mu)$ be a smooth center of the blow-up $\sigma : M \leftarrow M'$ and let D denote the exceptional divisor. Let $\mathcal{I}_C$ denote the sheaf of ideals defined by C. Then*

(1) $\mathcal{I} \subset \mathcal{I}_C^{\mu}$.
(2) $\sigma^*(\mathcal{I}) \subset (\mathcal{I}_D)^{\mu}$.

Proof. (1) We can assume that the ambient manifold M is isomorphic to an open ball in $\mathring{A}^n$. Let $u_1, \ldots, u_k$ be coordinates generating $\mathcal{I}_C$. Suppose $f \in \mathcal{I} \setminus \mathcal{I}_C^{\mu}$. Then we can write $f = \sum_{\alpha} c_{\alpha} u^{\alpha}$, where either $|\alpha| \geq \mu$ or $|\alpha| < \mu$ and $c_{\alpha} \notin \mathcal{I}_C$. By assumption there is α with $|\alpha| < \mu$ such that $c_{\alpha} \notin \mathcal{I}_C$. Take α with the smallest $|\alpha|$. There is a point $x \in C$ for which $c_{\alpha}(x) \neq 0$ and in the Taylor expansion of f at x there is a term $c_{\alpha}(x)u^{\alpha}$. Thus $\mathrm{ord}_x(\mathcal{I}) < \mu$. This contradicts the assumption $C \subset \mathrm{supp}(\mathcal{I}, \mu)$.
(2) $\sigma^*(\mathcal{I}) \subset \sigma^*(\mathcal{I}_C)^{\mu} = (\mathcal{I}_D)^{\mu}$. □

3.4. Functorial properties of multiple test blow-ups

We can define the fiber products for the germs of analytic spaces

$$(X, Z_X) \times_{(Y, Z_Y)} (\overline{X}, \overline{Z_X}) := (X \times_Y \overline{X}, Z_X \times_{Z_Y} \overline{Z_X}).$$

Proposition 3.4.1. *Let M_{iZ_i} be a multiple test blow-up of a marked ideal $(M_Z, \mathcal{I}, E, \mu)$ defining a sequence of marked ideals $(M_{iZ_i}, \mathcal{I}_i, E_i, \mu)$. Given a local analytic isomophism $\phi : M'_{Z'} \to M_Z$, the induced sequence $M'_{iZ_i} := M' \times_{M_Z} M_{i,Z_i}$ is a multiple test blow-up of $(M'_{Z'}, \mathcal{I}', E', \mu)$ such that*

(1) *ϕ lifts to local analytic isomorphisms $\phi_{iZ_i} : M'_{iZ_i} \to M_{iZ_i}$.*
(2) *$(M'_{iZ'_i})$ defines a sequence of marked ideals $(M'_{Z'_i}, \mathcal{I}'_i, E'_i, \mu)$ where $\mathcal{I}'_i = \phi_i^*(\mathcal{I}_i)$, the divisors in E'_i are the inverse images of the divisors in E_i and the order on E'_i is defined by the order on E_i.*
(3) *If (M_{iZ_i}) is a resolution of $(M_Z, \mathcal{I}, E, \mu)$ then $(M'_{iZ'_i})$ is an extension of a resolution of $(M'_{Z'}, \mathcal{I}', E', \mu)$.*

Proof Follows from definition. □

Definition 3.4.2. We say that the above multiple test blow-up $(M'_{iZ'_i})$ is *induced* via ϕ_i by M_{iZ_i}. We shall denote $(M'_{iZ'_i})$ and the corresponding marked ideals $(M'_{iZ'_i}, \mathcal{I}', E', \mu)$ by

$$\phi^*(M_{iZ_i}) := M'_{iZ'_i}, \quad \phi^*(M_{iZ_i}, \mathcal{I}_i, E_i, \mu) := (M'_{iZ'_i}, \mathcal{I}'_i, E'_i, \mu).$$

The above proposition and definition generalize to any sequence of blow-ups with smooth centers.

Proposition 3.4.3. *Let M_{iZ_i} be a sequence blow-ups with smooth centers having simple normal crossings with exceptional divisors.*

(1) *Given a surjective local analytic isomophism $\phi : M'_{Z'} \to M_Z$, the induced sequence $M'_{i,Z'_i} := M'_{Z'} \times_{M_Z} M_{iZ_i}$ is a sequence of blow-ups with smooth centers having simple normal crossings with exceptional divisors.* □

(2) *Given a local analytic isomophism* $\phi : M'_{Z'} \to M_Z$*, the induced sequence* $M'_{i,Z'_i} := M'_{Z'} \times_{M_Z} M_{iZ_i}$ *is an extension of a sequence of blow-ups with smooth centers having simple normal crossings with exceptional divisors.*

3.5. Canonical resolution of marked ideals

Theorem 3.5.1. *With any marked ideal* $(M_Z, \mathcal{I}, E, \mu)$ *there is associated a resolution* (M_{iZ_i}) *called canonical such that*

(1) *For any surjective local analytic isomorphism* $\phi : M'_{Z'} \to M_Z$ *the induced resolution* $\phi^*(M_{iZ_i})$ *is the canonical resolution of* $\phi^*(M_Z, \mathcal{I}, E, \mu)$.
(2) *For any local analytic isomorphism* $\phi : M'_{Z'} \to M_Z$ *the induced resolution* $\phi^*(M_{iZ_i})$ *is an extension of the canonical resolution of* $\phi^*(M_Z, \mathcal{I}, E, \mu)$.
(3) *If* $E = \emptyset$ *then* (M_i) *commutes with closed embeddings of the ambient manifolds* $M_Z \hookrightarrow M'_{Z'}$*, that is, the canonical resolution* (M_{iZ_i}) *of* $(M_Z, \mathcal{I}, \emptyset, \mu)$ *with centers* C_i *defines the canonical resolution* $(M'_{iZ'_i})$ *of* $(M'_{Z'}, \mathcal{I}', \emptyset, \mu)$*, where* $\mathcal{I}' = i_*(\mathcal{I}) \cdot \mathcal{O}_{M'}$*, with the centers* $i(C_i)$.

3.6. Canonical principalization of germs of ideals

Theorem 3.6.1. *Let* $\mathcal{I}$ *be a sheaf of ideals on a germ* M_Z *of an analytic manifold* M*. There exists a principalization of* $\mathcal{I}$*, that is, a projective morphism* $\mathrm{prin}(\mathcal{I}) : \widetilde{M}_{\widetilde{Z}} \to M_Z$ *a finite sequence*

$$M_Z = M_{0,Z_0} \xleftarrow{\sigma_1} M_{1Z_1} \xleftarrow{\sigma_2} M_{2,Z_2} \longleftarrow \ldots \longleftarrow M_{i,Z_i} \longleftarrow \ldots \longleftarrow M_{r,Z_r} = \widetilde{M}_{\widetilde{Z}}$$

of blow-ups with smooth centers $C_{i-1} \subset M_{i-1,Z_{i-1}}$ *such that*

(1) *The exceptional divisor* E_i *of the induced morphism* $\sigma^i = \sigma_1 \circ \ldots \circ \sigma_i : U_i \to U$ *has only simple normal crossings and* C_i *has simple normal crossings with* E_i.
(2) *The total transform* $\mathrm{prin}^*_{|\widetilde{U}}(\mathcal{I}) = \sigma^{r*}(\mathcal{I})$ *is the ideal of a simple normal crossing divisor* $\widetilde{E}$ *which is a natural combination of the irreducible components of the divisor* E_r.

The morphism $\mathrm{prin} : (\widetilde{M}, \widetilde{\mathcal{I}}) \to (M, \mathcal{I})$ *commutes with local analytic isomorphisms, embeddings of ambient manifolds.*

3.7. Canonical embedded desingularization of germs of analytic spaces

Theorem 3.7.1. *Let* M_Z *be a germ of an analytic manifold and* Y_Z *be a germ of analytic subspace of a germ* M_Z*. There exists an embedded desingularization of* $Y_Z \subset M_Z$ *that is, a finite sequence*

$$M_Z = M_{0,Z_0} \xleftarrow{\sigma_1} M_{1Z_1} \xleftarrow{\sigma_2} M_{2,Z_2} \longleftarrow \ldots \longleftarrow M_{i,Z_i} \longleftarrow \ldots \longleftarrow M_{r,Z_r} = \widetilde{M}_{\widetilde{Z}}$$

of blow-ups with smooth centers $C_{i-1} \subset M_{i-1,Z_{i-1}}$ *such that*

(1) *The exceptional divisor* E_i *of the induced morphism* $\sigma^i = \sigma_1 \circ \ldots \circ \sigma_i : U_i \to U$ *has only simple normal crossings and* C_i *has simple normal crossings with* E_i.

(2) *The strict transform* $\widetilde{Y}_{\widetilde{Z}} := Y_{r,Z_r}$ *of* Y_Z *is smooth and has only simple normal crossings with the exceptional divisor* E_r.
(3) *The morphism* $(M_Z, Y_Z) \leftarrow (\widetilde{M}_{\widetilde{Z}}, \widetilde{Y}_{\widetilde{Z}})$ *defined by the embedded desingularization commutes with local analytic isomorphisms, embeddings of ambient manifolds.*

3.8. Canonical desingularization of germs of analytic spaces

Theorem 3.8.1. *Let* Y *be an analytic space and* $Z \subset Y$ *be a compact subset. There exists a canonical desingularization of* Y_Z *that is a germ of a manifold* $\widetilde{Y}_{\widetilde{Z}}$ *together with a proper bimeromorphic morphism* $\mathrm{res}_{Y_Z} : \widetilde{Y}_{\widetilde{Z}} \to Y_Z$ *such that*

(1) $\widetilde{Z} = \mathrm{res}^{-1}_{Y_Z}(Z)$.
(2) $\mathrm{res}_{Y_Z} : \widetilde{Y}_{\widetilde{Z}} \to Y_Z$ *is an isomorphism over the nonsingular part* Y_{ns} *of* Y.
(3) *The inverse image of the singular locus* $\mathrm{res}^{-1}_{Y_Z}(Y_{Z\mathrm{sing}})$ *is a simple normal crossing divisor.*
(4) res_{Y_Z} *is functorial with respect to local analytic isomorphisms. For any local analytic isomorphism* $\phi : Y'_{Z'} \to Y_Z$ *there is a natural lifting* $\widetilde{\phi} : \widetilde{Y'}_{\widetilde{Z}'} \to \widetilde{Y}_{\widetilde{Z}}$ *which is a local analytic isomorphism.*

4. Hironaka resolution principle

Our proof is based upon the following principle which can be traced back to Hironaka and was used by Villamayor in his simplification of Hironaka's algorithm:

Proposition 4.0.2. *The following implications hold true:*

$$\textbf{\textit{Canonical resolution of germs of marked ideals}}\ (M_Z, \mathcal{I}, E, \mu) \tag{1}$$
$$\Downarrow$$
$$\textbf{\textit{Canonical principalization of germs of sheaves}}\ \mathcal{I}\ \textbf{\textit{on manifolds}}\ M \tag{2}$$
$$\Downarrow$$
$$\textbf{\textit{Canonical embedded desingularization of germs}}\ Y_Z \subset M_Z \tag{3}$$
$$\Downarrow$$
$$\textbf{\textit{Canonical desingularization of germs of analytic spaces}} \tag{4}$$

Proof (1)⇒(2) **Canonical principalization**

Let $\sigma : M_Z \leftarrow \widetilde{M_Z}$ denote the morphism defined by the canonical resolution $M_Z = M_{0,Z_0} \leftarrow M_{1,Z_1} \leftarrow M_{2,Z_2} \leftarrow \ldots \leftarrow M_{k,Z_k} = \widetilde{M_Z}$ of $(M_Z, \mathcal{I}, \emptyset, 1)$. The controlled transform $(\widetilde{\mathcal{I}}, 1) = (\mathcal{I}_k, 1) = \sigma^{\mathrm{c}}(\mathcal{I}, 1)$ has empty support. Consequently, $V(\widetilde{\mathcal{I}}) = V(\mathcal{I}_k) = \emptyset$, which implies $\widetilde{\mathcal{I}}_{\widetilde{Z}} = \mathcal{I}_k = \mathcal{O}_{\widetilde{M}_{\widetilde{Z}}}$. By definition for $i = 1, \ldots, k$, we have $(\mathcal{I}_i, 1) = \sigma^c_i(\mathcal{I}_{i-1}) = \mathcal{I}(D_i)^{-1}\sigma^*(\mathcal{I}_{i-1})$, and thus

$$\sigma^*_i(\mathcal{I}_{i-1}) = \mathcal{I}_i \cdot \mathcal{I}(D_i).$$

Note that if $\mathcal{I}(D) = \mathcal{O}(-D)$ is the sheaf of ideals of a simple normal crossing divisor D on a smooth M_Z and $\sigma : M'_{Z'} \to M_Z$ is the blow-up with a smooth center C which has only simple normal crossings with D then $\sigma^*(\mathcal{I}(D)) = \mathcal{I}(\sigma^*(D))$ is the sheaf of ideals of the divisor with simple normal crossings. The components of the induced Cartier divisors $\sigma^*(D)$ are either the strict transforms of the components of D or the components of the exceptional divisors. (The local equation $y_1^{a_1} \cdot \ldots \cdot y_l^{a_k}$ of D is transformed by the blow-up $(y_1, \ldots, y_n) \to (y_1, y_1y_2, y_1y_3, \ldots, y_1y_l, y_{l+1}, \ldots, y_n)$ into the equation $y_1^{a_1+\ldots a_l} y_2^{a_2} \ldots y_n^{a_n}$.) This implies by induction on i that

$$\sigma_i^* \sigma_{i-1}^* \ldots \sigma_2^* \sigma_1^*(\mathcal{I}_0) = \mathcal{I}_i \cdot \mathcal{I}(E_i)$$

where E_i is an exceptional divisor with simple normal crossings constructed inductively as

$$\mathcal{I}(E_i) = \sigma^*(\mathcal{I}(E_{i-1}))\mathcal{I}(D_i).$$

Finally the full transform $\sigma_k^*(\mathcal{I}) = \mathcal{I}_k \cdot \mathcal{I}(E_k) = \mathcal{O}_{\widetilde{M}} \cdot \mathcal{I}(E_k) = \mathcal{I}(E_k)$ is principal and generated by the sheaf of ideals of a divisor whose components are the exceptional divisors. The canonicity conditions for principalization follow from the canonicity of resolution of marked ideals.

(2)⇒(3) **Canonical embedded desingularization of germs of analytic spaces**

Lemma 4.0.3. *The canonical principalization of $\mathcal{I}$ on M_Z defines an isomorphism over $M_Z \setminus V(\mathcal{I})$.*

Proof. Let $p = 0 \in \mathbf{A}^n$ denote the origin of the affine space $\mathbf{A}^n$. The canonical principalization of the germ $(\mathbf{A}^n_{\{p\}}, \mathcal{O}_{\mathbf{A}^n})$ is an isomorphism over generic points in a neighborhood of p and is equivariant with respect to Gl(n) action , thus it is an isomorphism. The restriction of the canonical principalization $(\widetilde{M_{\widetilde{Z}}}, \widetilde{\mathcal{I}})$ of $(M_Z, \mathcal{I})$ to an open subset $U_{Z_U} \subset M_Z$ determines the canonical principalization of $(U_{Z_U}, \mathcal{I}_{|U_{Z_U}})$. Let $\widetilde{M_{\widetilde{Z}}} \to M_Z$ be the canonical principalization of $(M_Z, \mathcal{O}_{M_Z})$ and $x \in Z \setminus V(\mathcal{I})$. Locally we find an open subset $U_{\{x\}} \subset M_Z \setminus V(\mathcal{I})$ isomorphic to $(\mathbf{A}^n_{\{p\}}, \mathcal{O}_{\mathbf{A}^n})$. The canonical principalization of $(U_{\{x\}}, \mathcal{I}_U) = ((U_{\{x\}}, \mathcal{O}_U) \simeq (\mathbf{A}^n_{\{p\}}, \mathcal{O}_{\mathbf{A}^n})$ is an isomorphism. □

Let $Y_Z \subset M_Z$ be a germ of a closed analytic subspace $Y \subset M$. Let $M_Z = M_{0,Z_0} \leftarrow M_{1,Z_1} \leftarrow M_{2,Z_2} \leftarrow \ldots \leftarrow M_{k,Z_k} = \widetilde{M_Z}$ be the canonical principalization of germs sheaves of ideals $\mathcal{I}_Y$. It defines a sequence of blow-ups $U_0 \leftarrow U_k$ which is a principalization of $\mathcal{I}_{U_0}$ for a suitable open neighborhood U_0 of Z.

Suppose all centers C_{i-1} of the blow-ups $\sigma_i : U_{i-1} \leftarrow U_i$ are disjoint from the generic points of strict transforms Y_{i-1} of $Y_0 = Y \cap U_0$. Then $\widetilde{\sigma}$ is an isomorphism over the generic points y of Y_0 and $\widetilde{\sigma}^*(\mathcal{I})_y = \sigma^*(\mathcal{I})_y$. Moreover no exceptional divisor pass through y. This contradicts the condition $\widetilde{\sigma}^*(\mathcal{I}) = \mathcal{I}_{\widetilde{E}}$. Thus there is a smallest i_{res} with the property that $C_{i_{\text{res}}}$ contains the strict transform $Y_{i_{\text{res}}}$ and all centers C_j for $j < i_{\text{res}}$ are disjoint from the generic points of strict transforms Y_j. Let $y \in Y_{i_{\text{res}}}$ be a generic point for which $U_{i_{\text{res}}} \to U_0$ is an isomorphism. Find an open set $U \subset U_0$ intersecting Y such that

$U_{i_{\rm res}} \to U_0$ is an isomorphism over U. Then $Y_{i_{\rm res}} \cap U = Y \cap U$ and $C_{i_{\rm res}} \cap U \supset Y_{i_{\rm res}} \cap U$ by the definition of $Y_{i_{\rm res}}$. On the other hand, by the previous lemma $C_{i_{\rm res}} \cap U \subseteq Y_{i_{\rm res}} \cap U$, which gives $C_{i_{\rm res}} \cap U = Y_{i_{\rm res}} \cap U$. Finally, $Y_{i_{\rm res}}$ is an irreducible component of a smooth (possibly reducible) center C_i. This implies that $Y_{i_{\rm res}}$ is smooth and has simple normal crossings with the exceptional divisors. We define the canonical embedded resolution of (M_Z, Y_Z) to be

$$(M_Z, Y_Z) = (U_{0Z}, Y_{0Z}) \leftarrow (U_{1Z_1}, Y_{1Z_1}) \leftarrow (U_{2Z_2}, Y_{2Z_2}) \leftarrow \ldots \leftarrow (U_{i_{\rm res},Z_{i_{\rm res}}}, Y_{i_{\rm res}Z_{i_{\rm res}}}).$$

It is independent of the choice of U. If $(M'_{Z'}, Y'_Z) \to (M_Z, Y_Z)$ is a local analytic isomophism then the induced sequence of blow-ups $(U'_{iZ_i})_{0\le i\le k} = (U'_{Z'}, \times_{M_Z} U_{iZ_i})_{0\le i\le k}$ is an extension of the canonical principalization $(U'_{j,Z'_j})_{0\le j\le k'}$ of $(U'_{0Z'_0}, \mathcal{I}_{Y'|U'_0})$. Moreover $U'_{j_{\rm res}} = U'_{i_{\rm res}}$ and $(U'_i)_{0\le i\le i_{\rm res}}$ is is an extension of the canonical resolution $(U'_j)_{0\le j\le j_{\rm res}}$ of $(M'_{Z'}, Y'_Z)$. Commutativity with closed embeddings for embedded desingularizations follows from the commutativity with closed embeddings for principalizations.

(3)⇒(4) **Canonical desingularization of germs**

Let Y be an analytic space. Every point of $y \in Y$ has a neighborhood V which is locally isomorphic to a closed analytic subset of an open ball $U \subset \mathbb{C}^n$. The coordinates $u_1, \ldots u_n$ on Y define a minimal embedding $Y \supset V \to U$ into an open subset U of $\mathbb{C}^n$. Let $Z \subset V = Y \cap U$ be a compact set. Then Y_Z can be identified with V_Z. Consider the canonical embedded desingularization $(\widetilde{U_Z}, \widetilde{Y_Z}) \to (U_Z, Y_Z)$. Then we define the canonical desingularization of Y_Z to be $\widetilde{Y_Z} \to Y_Z$. Two minimal embeddings $\phi_1 : Z \subset V_1 \to U_1 \supset Z_1 = \phi_1(Z)$ and $\phi_2 : Z \subset V_2 \to U_2 \supset Z_2 = \phi_2(Z)$ of two different open subsets V_1, V_2 containing Z are defined by two different sets of coordinates $u_1, \ldots u_n$ and $u'_1, \ldots u'_n$ differ by an isomorphism

$$\psi := \phi_2^{-1}\phi_1 : (U_{1Z_1}, (\phi_1(V_1)_{Z_1}) \to (U_{2Z_2}, (\phi_2(V_2)_{Z_2})$$

mapping coordinates $x_1, \ldots, x_n$ to $x'_1, \ldots, x'_n$. Note that both $\phi_1(V_1)_{Z_1}$ and $\phi_2(V_1)_{Z_2}$ can be identified with $\widetilde{Y}_{\widetilde{Z}}$. The isomorphism ψ, by canonicity, lifts to the isomorphisms between embedded desingularizations $\widetilde{\psi} : (\widetilde{U_{1\widetilde{Z}_1}}, \widetilde{Y_{1\widetilde{Z}}}) \to (\widetilde{U_{2Z}}, \widetilde{Y_{2Z}})$ and nonembedded desingularizations $\widetilde{Y_{1Z}} \to \widetilde{Y_{2Z}}$. The latter shows that $\widetilde{Y_Z} \to Y_Z$ is independent of the choice of ambient manifold U. Observe that if $Y_Z \subset Y'_{Z'}$ is an open embedding then it extends to an open embedding $U_Z \subset U'_{Z'}$ and it defines an open embeddings of desingularizations $\widetilde{Y_Z} \subset \widetilde{Y'_{Z'}}$.

Let Y_Z denote the analytic germ of Y at Z. Consider an open cover of Z with the open subsets $V_i \subset W_i \subset U_i$ of Y, such that $\overline{V_i} \subset W_i$ and $\overline{V_i} \subset U_i$ are compact and U_i is isomorphic to an open balls as above. Set $S_i := \overline{V_i}$, $Z_i := \overline{W_i} \cap Z$.

The desingularization of $Y_{S_i} = U_{iS_i}$ determines the desingularization $\widetilde{U'_i}$ of an open neighborhood U'_i of Y_{S_i} and thus the desingularization $\widetilde{V_i} \to V_i$ of $V_i \subset U'_i$.

For each i,j, the embedding $Y_{Z_i\cap Z_j}\to Y_{Z_i}$ lifts to embeddings of nonembedded desingularizations of germs $\widetilde{Y_{Z_i\cap Z_j}}\to\widetilde{Y_{Z_i}}$. Note that the open embedding $V_i\cap V_j\to V_i$ is the restriction of $Y_{Z_i\cap Z_j}\to Y_{Z_i}$. It defines an embedding of desingularizations $(V_i\cap V_j)\tilde{}\to\widetilde{V_i}$.

Let $\widetilde{V}$ be a manifold obtained by gluing V_i along $V_i\cap V_j$. The desingularization morphism $\mathrm{des}_V:\widetilde{V}\to V$ is bimeromorphic and proper. Let $\widetilde{Z}:=\mathrm{des}_V^{-1}(Z)$. Note that $Y_Z=\bigcup Y_{Z_i}=\bigcup(V_i)_{Z_i}$. We define the canonical desingularization of Y_Z to be

$$\widetilde{Y_Z}:=\widetilde{V}_{\widetilde{Z}}=\bigcup\widetilde{V_{iZ_i}}.$$

It follows from the definition that it commutes with local analytic isomorphisms. □

4.1. Canonical principalization of ideal sheaves on analytic spaces

Let $\mathcal{I}$ be an ideal sheaf on a manifold M. Consider an open cover $\{U_i\}_{i\in I}$ of M, such that $Z_i:=\overline{U_i}$ are compact. For every i let $\mathrm{prin}_i:(\widetilde{Y}_{Z_i},\widetilde{\mathcal{I}}_{Z_i})\to(Y_{Z_i},\mathcal{I}_{Z_i})$ be a canonical principalization of $\mathcal{I}$ on Y_{Z_i}. Let $\widetilde{U}_i:=\mathrm{prin}_i^{-1}(U_i)\to(U_i,\mathcal{I}_{|U_i})$ be its restriction. By canonicity, $\mathrm{prin}_i:\mathrm{prin}_i^{-1}(Y_{Z_i}\cap Y_{Z_j})$ is isomorphic over $Y_{Z_i}\cap Y_{Z_j}$ to $\widetilde{Y}_{Z_i\cap Z_j}$. Thus the meromorphic map

$$\widetilde{U}_{ij}:=\mathrm{prin}_i^{-1}(U_i\cap U_j)\simeq\widetilde{U}_{ji}:=\mathrm{prin}_j^{-1}(U_i\cap U_j)$$

is an isomorphism. We define $\widetilde{M}$ to be a manifold obtained by gluing $\widetilde{U}_i$ along $\widetilde{U}_{ij}$. Then $\mathrm{prin}:\widetilde{M}\to M$ is a proper bimeromorphic morphism. Moreover for any compact $Z\subset M$, $(\widetilde{M}_{\widetilde{Z}},\widetilde{\mathcal{I}}_{\widetilde{Z}})\to(M_Z,\mathcal{I}_Z)$ is a canonical principalization of $\mathcal{I}$ on the germ M_Z.

4.2. Canonical embedded desingularization of analytic spaces

Let $Y\subset M$ be an analytic subspace of a manifold. Consider an open cover $\{U_i\}_{i\in I}$ of M, such that $Z_i:=\overline{U_i}$ are compact. For every i let $\mathrm{des}_i:(\widetilde{M}_{Z_i},\widetilde{Y}_{Z_i})\to(M_{Z_i},Y_{Z_i})$ be the canonical desingularization of Y_{Z_i}. Let $(\widetilde{U}_i,\widetilde{U}_i^Y):=\mathrm{des}_i^{-1}(U_i,Y\cap U_i)\to(U_i,Y\cap U_i)$ be its restriction. As before we define $\widetilde{M}$ to be a manifold obtained by gluing $\widetilde{U}_i$ along $\widetilde{U}_{ij}$. A subspace $\widetilde{Y}\subset\widetilde{M}$ is a manifold obtained by gluing $\widetilde{U^Y}_i$ along $\widetilde{U^Y}_{ij}$. Then $\mathrm{des}:(\widetilde{M},\widetilde{Y})\to(M,Y)$ is a proper bimeromorphic morphism. Moreover for any compact $Z\subset M$, $(\widetilde{M}_{\widetilde{Z}}\widetilde{Y}_{\widetilde{Z}})\to(M_Z,Y_Z)$ is a canonical embedded desingularization of the germ $Y_Z\subset M_Z$.

4.3. Canonical desingularization of analytic spaces

Let Y be an analytic space. Consider an open cover $\{U_i\}_{i\in I}$ of Y, such that $Z_i:=\overline{U_i}$ are compact. For every i let $\mathrm{des}_i:\widetilde{Y}_{Z_i}\to Y_{Z_i}$ be the canonical desingularization of the germ Y_{Z_i}. Let $\widetilde{U}_i:=\mathrm{des}_i^{-1}(U_i)\to U_i$ be its restriction. As before we define $\widetilde{Y}$ to be a manifold obtained by gluing $\widetilde{U}_i$ along $\widetilde{U}_{ij}$. Then $\mathrm{des}:\widetilde{Y}\to Y$ is a proper bimeromorphic morphism. Moreover for any compact $Z\subset Y$, $\widetilde{Y}_{\widetilde{Z}}\to Y_Z$ is a canonical desingularization of germ Y_Z.

5. Marked ideals

5.1. Equivalence relation for marked ideals

Let us introduce the following equivalence relation for marked ideals:

Definition 5.1.1. Let $(M_Z, \mathcal{I}, E_\mathcal{I}, \mu_\mathcal{I})$ and $(M_Z, \mathcal{J}, E_\mathcal{J}, \mu_\mathcal{J})$ be two marked ideals on the manifold M_Z. Then $(M_Z, \mathcal{I}, E_\mathcal{I}, \mu_\mathcal{I}) \simeq (M_Z, \mathcal{J}, E_\mathcal{J}, \mu_\mathcal{J})$ if

(1) $E_\mathcal{I} = E_\mathcal{J}$ and the orders on $E_\mathcal{I}$ and on $E_\mathcal{J}$ coincide.
(2) $\mathrm{supp}(M_Z, \mathcal{I}, E_\mathcal{I}, \mu_\mathcal{I}) = \mathrm{supp}(M_Z, \mathcal{J}, E_\mathcal{J}, \mu_\mathcal{J})$.
(3) All the multiple test blow-ups $M_{Z0} = M_Z \overset{\sigma_1}{\longleftarrow} M_{1Z_1} \overset{\sigma_2}{\longleftarrow} \ldots \longleftarrow M_{iZ_i} \overset{\sigma_r}{\longleftarrow} \ldots \longleftarrow M_{rZ_r}$ of $(M_Z, \mathcal{I}, E_\mathcal{I}, \mu_\mathcal{I})$ are exactly the multiple test blow-ups of $(M_Z, \mathcal{J}, E_\mathcal{J}, \mu_\mathcal{J})$ and moreover we have
$$\mathrm{supp}(M_{iZ_i}, \mathcal{I}_i, E_i, \mu_\mathcal{I}) = \mathrm{supp}(M_{iZ_i}, \mathcal{J}_i, E_i, \mu_\mathcal{J}).$$

It is easy to show the lemma:

Lemma 5.1.2. *For any $k \in \mathbf{N}$, $(\mathcal{I}, \mu) \simeq (\mathcal{I}^k, k\mu)$.*

Remark. The marked ideals considered in this paper satisfy a stronger equivalence condition: For any local analytic isomorphisms $\phi : M'_Z \to M_Z$, $\phi^*(\mathcal{I}, \mu) \simeq \phi^*(\mathcal{J}, \mu)$. This condition will follow and is not added in the definition.

5.2. Ideals of derivatives

Ideals of derivatives were first introduced and studied in the resolution context by Giraud. Villamayor developed and applied this language to his *basic objects*.

Definition 5.2.1. (Giraud, Villamayor) Let $\mathcal{I}$ be a coherent sheaf of ideals on a germ of manifold M_Z. By the *first derivative* (originally *extension*) $\mathcal{D}_{M_Z}(\mathcal{I})$ of $\mathcal{I}$ (or simply $\mathcal{D}(\mathcal{I})$) we mean the coherent sheaf of ideals generated by all functions $f \in \mathcal{I}$ with their first derivatives. Then the *i-th derivative* $\mathcal{D}^i(\mathcal{I})$ is defined to be $\mathcal{D}(\mathcal{D}^{i-1}(\mathcal{I}))$. If $(\mathcal{I}, \mu)$ is a marked ideal and $i \leq \mu$ then we define
$$\mathcal{D}^i(\mathcal{I}, \mu) := (\mathcal{D}^i(\mathcal{I}), \mu - i).$$

Recall that on a manifold M there is a locally free sheaf of differentials $\Omega_{M/K}$ generated locally by $du_1, \ldots, du_n$ for a set of local coordinates $u_1, \ldots, u_n$. The dual sheaf of derivations $\mathrm{Der}_K(\mathcal{O}_M)$ is locally generated by the derivations $\frac{\partial}{\partial u_i}$. Immediately from the definition we observe that $\mathcal{D}(\mathcal{I})$ is a coherent sheaf defined locally by generators f_j of $\mathcal{I}$ and all their partial derivatives $\frac{\partial f_j}{\partial u_i}$. We see by induction that $\mathcal{D}^i(\mathcal{I})$ is a coherent sheaf defined locally by the generators f_j of $\mathcal{I}$ and their derivatives $\frac{\partial^{|\alpha|} f_j}{\partial u^\alpha}$ for all multiindices $\alpha = (\alpha_1, \ldots, \alpha_n)$, where $|\alpha| := \alpha_1 + \ldots + \alpha_n \leq i$.

Lemma 5.2.2. *(Giraud, Villamayor) For any $i \leq \mu - 1$,*
$$\mathrm{supp}(\mathcal{I}, \mu) = \mathrm{supp}(\mathcal{D}^i(\mathcal{I}), \mu - i)).$$

In particular $\mathrm{supp}(\mathcal{I},\mu)=\mathrm{supp}(\mathcal{D}^{\mu-1}(\mathcal{I}),1)=V(\mathcal{D}^{\mu-1}(\mathcal{I}))$ *is a closed set* ($i=\mu-1$).

Proof. It suffices to prove the lemma for $i=1$. If $x\in\mathrm{supp}(\mathcal{I},\mu)$ then for any $f\in\mathcal{I}$ we have $\mathrm{ord}_x(f)\geq\mu$. This implies $\mathrm{ord}_x(Df)\geq\mu-1$ for any derivative D and consequently $x\in\mathrm{supp}(\mathcal{D}(\mathcal{I}),\mu-1))$. Now, let $x\in\mathrm{supp}(\mathcal{D}(\mathcal{I}),\mu-1))$. Then for any $f\in\mathcal{I}$ we have $\mathrm{ord}_x(f)\geq\mu-1$. Suppose $\mathrm{ord}_x(f)=\mu-1$ for some $f\in\mathcal{I}$. Then $f=\sum_{|\alpha|\geq\mu-1}c_\alpha x^\alpha$ and there is α such that $\alpha=\mu-1$ and $c_\alpha\neq 0$. We find $\frac{\partial}{\partial x_i}$ for which $\mathrm{ord}_x(\frac{\partial x^\alpha}{\partial x_i})=\mu-2$ and thus $\mathrm{ord}_x(\frac{\partial f}{\partial x_i})=\mu-2$ and $x\notin\mathrm{supp}(\mathcal{D}(\mathcal{I}),\mu-1))$. □

We write $(\mathcal{I},\mu)\subset(\mathcal{J},\mu)$ if $\mathcal{I}\subset\mathcal{J}$.

Lemma 5.2.3. *(Giraud, Villamayor) Let* $(\mathcal{I},\mu)$ *be a marked ideal and* $C\subset\mathrm{supp}(\mathcal{I},\mu)$ *be a smooth center and* $r\leq\mu$. *Let* $\sigma:M_Z\leftarrow M'_Z$ *be a blow-up at* C. *Then*

$$\sigma^{\mathrm{c}}(\mathcal{D}^r_{M_Z}(\mathcal{I},\mu))\subseteq\mathcal{D}^r_{M_Z{}'}(\sigma^{\mathrm{c}}(\mathcal{I},\mu)).$$

Proof. First assume that $r=1$. Let $u_1,\ldots,u_n$ denote the local coordinates at $x\in C$ such that C is a coordinate subspace. Then the local coordinates at $x'\in\sigma^{-1}(x)$ are of the form $u'_i=\frac{u_i}{u_m}$ for $i<m$ and $u'_i=u_i$ for $i\geq m$, where $u_m=u'_m=y$ denotes the local equation of the exceptional divisor.

The derivations $\frac{\partial}{\partial u_i}$ of $\mathcal{O}_{x,M}$ extend to derivations of the rational field $K(\mathcal{O}_{x,M})$. Note also that

$$\frac{\partial u'_j}{\partial u_i}=\frac{\delta_{ij}}{u_m},\quad i<m,1\leq j\leq n;\qquad \frac{\partial u'_j}{\partial u_m}=-\frac{1}{u_m^2}u_j,\quad j<m;\qquad \frac{\partial u'_m}{\partial u_m}=1;$$
$$\frac{\partial u'_j}{\partial u_m}=0,j>m;\qquad \frac{\partial u'_i}{\partial u_j}=\delta_{ij},\quad i\geq m.$$

This gives

$$\frac{\partial}{\partial u_i}=\frac{1}{u_m}\frac{\partial}{\partial u'_i}=\frac{1}{y}\frac{\partial}{\partial u'_i},\quad 1\leq i<m;\qquad \frac{\partial}{\partial u'_i}=\frac{\partial}{\partial u_i},\quad m<i\leq n,$$
$$\frac{\partial}{\partial u_m}=-\frac{1}{y}(u'_1\frac{\partial}{\partial u'_1}+\ldots+u'_{m-1}\frac{\partial}{\partial u'_{m-1}}-u'_m\frac{\partial}{\partial u'_m}).$$

We see that any derivation D of $\mathcal{O}_{x,M}$ induces a derivation $y\sigma^*(D)$ of $\mathcal{O}_{x',M'}$. Let E be the exceptional divisor $\mathcal{I}(E)$ be its ideal sheaf (locally generated by y). Thus the sheaf of derivations $\mathcal{I}(E)\sigma^*(\mathrm{Der}_K(\mathcal{O}_M))$ is a subsheaf of $\mathrm{Der}_K(\mathcal{O}_{M'})$ locally generated by

$$\frac{\partial}{\partial u'_i},i<m;\quad y\frac{\partial}{\partial y},\quad\text{and}\quad y\frac{\partial}{\partial u'_i},i>m.$$

In particular $\mathcal{I}(E)\sigma^*(\mathcal{D}_M(\mathcal{I}))\subset\mathcal{D}_{M'}(\sigma^*(\mathcal{I}))$. For any sheaf of ideals $\mathcal{J}$ on M' denote by $\mathcal{I}(E)\sigma^*(\mathcal{D}_M)(\mathcal{J})\subset\mathcal{D}_{M'}(\mathcal{J})$ the ideal generated by $\mathcal{J}$ and the derivatives $D'(f)$, where $f\in\mathcal{J}$ and $D'\in\mathcal{I}(E)\sigma^*(\mathrm{Der}_K(\mathcal{O}_M))$. Note that for a neighborhood $U'\ni x'$ and any $f\in\mathcal{J}(U')$ and $D'\in y\sigma^*(\mathrm{Der}_K(\mathcal{O}_M))$, y divides $D'(y)$ and

$$D'(yf)=yD'(f)+D'(y)f\quad\in\quad y\sigma^*(\mathcal{D}_M)(\mathcal{J})+y\mathcal{J}=y\sigma^*(\mathcal{D}_M)(\mathcal{J}).$$

Consequently, $y\sigma^*(\mathcal{D}_M)(y\mathcal{J}) \subseteq yy\sigma^*(\mathcal{D}_M)(\mathcal{J})$ and more generally $y\sigma^*(\mathcal{D}_M)(y^\mu\mathcal{J}) \subseteq y^\mu y\sigma^*(\mathcal{D}_{M'})(\mathcal{J})$. Then

$$\begin{aligned} y\sigma^*(\mathcal{D}_M(\mathcal{I})) \subseteq y\sigma^*(\mathcal{D}_M)(\sigma^*(\mathcal{I})) &= y\sigma^*(\mathcal{D}_M)(y^\mu\sigma^{\rm c}(\mathcal{I})) \\ &\subseteq y^\mu y\sigma^*(\mathcal{D}_M)(\sigma^{\rm c}(\mathcal{I})) \subseteq y^\mu\mathcal{D}_{M'}(\sigma^{\rm c}(\mathcal{I})). \end{aligned}$$

Then

$$\sigma^{\rm c}(\mathcal{D}_M(\mathcal{I})) = y^{-\mu+1}\sigma^*(\mathcal{D}_M(\mathcal{I})) \subseteq \mathcal{D}_{M'}(\sigma^{\rm c}(\mathcal{I})).$$

Assume now that r is arbitrary. Then $C \subset \mathrm{supp}(\mathcal{I},\mu) = \mathrm{supp}(\mathcal{D}^i_M(\mathcal{I},\mu))$ for $i \leq r$ and by induction on r,

$$\sigma^{\rm c}(\mathcal{D}^r_M\mathcal{I}) = \sigma^{\rm c}(\mathcal{D}_M(\mathcal{D}^{r-1}_M(\mathcal{I}))) \subseteq \mathcal{D}_{M'}(\sigma^{\rm c}\mathcal{D}^{r-1}_M(\mathcal{I})) \subseteq \mathcal{D}^r_{M'}(\sigma^{\rm c}(\mathcal{I})).$$

□

As a corollary from Lemma 5.2.3 we prove the following

Lemma 5.2.4. *A multiple test blow-up $(M_i)_{0\leq i\leq k}$ of $(\mathcal{I},\mu)$ is a multiple test blow-up of $\mathcal{D}^j(\mathcal{I},\mu)$ for $0 \leq j \leq \mu$ and*

$$[\mathcal{D}^j(\mathcal{I},\mu)]_k \subset \mathcal{D}^j(\mathcal{I}_k,\mu).$$

Proof. Induction on k. For $k = 0$ evident. Let $\sigma_{k+1} : M_k \leftarrow M_{k+1}$ denote the blow-up with a center $C_k \subseteq \mathrm{supp}(\mathcal{I}_k,\mu) = \mathrm{supp}(\mathcal{D}^j(\mathcal{I}_k,\mu)) \subseteq \mathrm{supp}([\mathcal{D}^j(\mathcal{I},\mu)]_k)$. Then by induction $[\mathcal{D}^j(\mathcal{I},\mu)]_{k+1} = \sigma^{\rm c}_{k+1}[\mathcal{D}^j(\mathcal{I},\mu)]_k \subseteq \sigma^{\rm c}_{k+1}(\mathcal{D}^j(\mathcal{I}_k,\mu))$. Lemma 5.2.3 gives $\sigma^{\rm c}_{k+1}(\mathcal{D}^j(\mathcal{I}_k,\mu)) \subseteq \mathcal{D}^j\sigma^{\rm c}_{k+1}(\mathcal{I}_k,\mu) = \mathcal{D}^j(\mathcal{I}_{k+1},\mu)$. □

5.3. Hypersurfaces of maximal contact

The concept of the *hypersurfaces of maximal contact* is one of the key points of this proof. It was originated by Hironaka, Abhyankhar and Giraud and developed in the papers of Bierstone-Milman and Villamayor.

In our terminology we are looking for a smooth hypersurface containing the supports of marked ideals and whose strict transforms under multiple test blow-ups contain the supports of the induced marked ideals. Existence of such hypersurfaces allows a reduction of the resolution problem to codimension 1.

First we introduce marked ideals which locally admit hypersurfaces of maximal contact.

Definition 5.3.1. (Villamayor [35]) We say that $(M_Z,\mathcal{I},E,\mu)$ be a marked ideal of *maximal order* (originally *simple basic object*) if there exists an open neighborhood U of $M_Z = (U,Z)$ such that $\mathcal{I}$ is defined on $U \supset Z$ and $\max\{\mathrm{ord}_x(\mathcal{I}) \mid x \in U\} \leq \mu$ or equivalently $\mathcal{D}^\mu(\mathcal{I}) = \mathcal{O}_{M_Z}$.

Lemma 5.3.2. *(Villamayor [35]) Let $(\mathcal{I},\mu)$ be a marked ideal of maximal order and $C \subset \mathrm{supp}(\mathcal{I},\mu)$ be a smooth center. Let $\sigma : M_Z \leftarrow M'_{Z'}$ be a blow-up at $C \subset \mathrm{supp}(\mathcal{I},\mu)$. Then $\sigma^{\rm c}(\mathcal{I},\mu)$ is of maximal order.*

Proof. If $(\mathcal{I},\mu)$ is a marked ideal of maximal order then $\mathcal{D}^\mu(\mathcal{I}) = \mathcal{O}_{M_Z}$. Then by Lemma 5.2.3, $\mathcal{D}^\mu(\sigma^{\rm c}(\mathcal{I},\mu)) \supset \sigma^{\rm c}(\mathcal{D}^\mu(\mathcal{I}),0) = \mathcal{O}_{M_Z}$. □

Lemma 5.3.3. *(Villamayor [35]) If $(\mathcal{I},\mu)$ is a marked ideal of maximal order and $0 \leq i \leq \mu$ then $\mathcal{D}^i(\mathcal{I},\mu)$ is of maximal order.*

Proof. $\mathcal{D}^{\mu-i}(\mathcal{D}^i(\mathcal{I},\mu)) = \mathcal{D}^{\mu}(\mathcal{I},\mu) = \mathcal{O}_{M_Z}$. □

In particular $(\mathcal{D}^{\mu-1}(\mathcal{I}), 1)$ is a marked ideal of maximal order.

Lemma 5.3.4. *(Giraud) Let $(\mathcal{I},\mu)$ be a marked ideal of maximal order and let $\sigma : M_Z \leftarrow M'_{Z'}$ be a blow-up at a smooth center $C \subset \operatorname{supp}(\mathcal{I},\mu)$. Let $u \in \mathcal{D}^{\mu-1}(\mathcal{I},\mu)(U)$ be a function of multiplicity one on U, that is, for any $x \in V(u)$, $\operatorname{ord}_x(u) = 1$. In particular $\operatorname{supp}(\mathcal{I},\mu) \cap U \subset V(u)$. Let $U' \subset \sigma^{-1}(U) \subset M'_{Z'}$ be an open set where the exceptional divisor is described by y. Let $u' := \sigma^{\mathrm{c}}(u) = y^{-1}\sigma^*(u)$ be the controlled transform of u. Then*

(1) *$u' \in \mathcal{D}^{\mu-1}(\sigma^{\mathrm{c}}(\mathcal{I}_{|U'},\mu))$.*
(2) *u' is a function of multiplicity one on U'.*
(3) *$V(u')$ is the restriction of the strict transform of $V(u)$ to U'.*

Proof. (1) $u' = \sigma^{\mathrm{c}}(u) = u/y \in \sigma^{\mathrm{c}}(\mathcal{D}^{\mu-1}(\mathcal{I})) \subset \mathcal{D}^{\mu-1}(\sigma^{\mathrm{c}}(\mathcal{I}))$.

(2) Since u was one of the local coordinates describing the center of blow-ups, $u' = u/y$ is a parameter, that is, a function of order one.

(3) follows from (2). □

Definition 5.3.5. We shall call a function

$$u \in T(\mathcal{I})(U) := \mathcal{D}^{\mu-1}(\mathcal{I}(U))$$

of multiplicity one a *tangent direction* of $(\mathcal{I},\mu)$ on U.

As a corollary from the above we obtain the following lemma:

Lemma 5.3.6. *(Giraud) Let $u \in T(\mathcal{I})(U)$ be a tangent direction of $(\mathcal{I},\mu)$ on U. Then for any multiple test blow-up (U_i) of $(\mathcal{I}_{|U},\mu)$ all the supports of the induced marked ideals $\operatorname{supp}(\mathcal{I}_i,\mu)$ are contained in the strict transforms $V(u)_i$ of $V(u)$.* □

Remarks. (1) Tangent directions are functions defining locally hypersurfaces of maximal contact.

(2) The main problem leading to complexity of the proofs is that of noncanonical choice of the tangent directions. We overcome this difficulty by introducing *homogenized ideals*.

Lemma 5.3.7. *(Villamayor) Let $(\mathcal{I},\mu)$ be a marked ideal of maximal order whose support is of codimension 1. Then all codimension one components of $\operatorname{supp}(\mathcal{I},\mu)$ are smooth and isolated. After the blow-up $\sigma : M_Z \leftarrow M'_{Z'}$ at such a component $C \subset \operatorname{supp}(\mathcal{I},\mu)$ the induced support $\operatorname{supp}(\mathcal{I}',\mu)$ does not intersect the exceptional divisor of σ.*

Proof. By the previous lemma there is a tangent direction $u \in \mathcal{D}^{\mu-1}(\mathcal{I})$ whose zero set is smooth and contains $\operatorname{supp}(\mathcal{I},\mu)$. Then $\mathcal{D}^{\mu-1}(\mathcal{I}) = (u)$ and $\mathcal{I}$ is locally described as $\mathcal{I} = (u^\mu)$. Suppose there is $g \in \mathcal{I}$ written as $g = c_\mu(x,u)u^\mu + c_{\mu-1}(x)u^{\mu-1} + \ldots + c_0(x)$, where at least one function $c_i(x) \neq 0$ for $0 \leq i \leq \mu-1$. Then there is a multiindex α such $|\alpha| = \mu - i - 1$ and $\frac{\partial^{|\alpha|}c_i}{\partial x^\alpha}$ is not the zero function. Then the derivative $\frac{\partial^{\mu-1}g}{\partial u^i \partial x^\alpha} \in \mathcal{D}^{\mu-1}(\mathcal{I})$ does not belong to the ideal (u).

The blow-up at the component C locally defined by u transforms $(\mathcal{I},\mu)=((u^\mu),\mu)$ to $(\mathcal{I}',\mu)$, where $\sigma^*(\mathcal{I})=y^\mu\mathcal{O}_M$, and $\mathcal{I}'=\sigma^c(\mathcal{I})=y^{-\mu}\sigma^*(\mathcal{I})=\mathcal{O}_M$, where $y=u$ describes the exceptional divisor. □

Remark. Note that the blow-up of codimension one components is an isomorphism. However it defines a nontrivial transformation of marked ideals. In the actual desingularization process this kind of blow-up may occur for some marked ideals induced on subvarieties of ambient varieties. Though they define isomorphisms of those subvarieties they determine blow-ups of ambient varieties which are not isomorphisms.

5.4. Arithmetical operations on marked ideals

In this section all marked ideals are defined for the germ of the manifold M and the same set of exceptional divisors E. Define the following operations of addition and multiplication of marked ideals:

(1) $(\mathcal{I},\mu_\mathcal{I})+(\mathcal{J},\mu_\mathcal{J}):=(\mathcal{I}^{\mathrm{lcm}(\mu_\mathcal{I},\mu_\mathcal{J})/\mu_\mathcal{I}}+\mathcal{J}^{\mathrm{lcm}(\mu_\mathcal{I},\mu_\mathcal{J})/\mu_\mathcal{J}},\mathrm{lcm}(\mu_\mathcal{I},\mu_\mathcal{J}))$
or more generally (the operation of addition is not associative)
$(\mathcal{I}_1,\mu_1)+\ldots+(\mathcal{I}_m,\mu_m):=(\mathcal{I}_1^{\mathrm{lcm}(\mu_1,\ldots,\mu_m)/\mu_1}+\mathcal{I}_2^{\mathrm{lcm}(\mu_1,\ldots,\mu_m)/\mu_2}$
$+\ldots+\mathcal{I}_m^{\mathrm{lcm}(\mu_1,\ldots,\mu_m)/\mu_m},\mathrm{lcm}(\mu_1,\ldots,\mu_m))$.

(2) $(\mathcal{I},\mu_\mathcal{I})\cdot(\mathcal{J},\mu_\mathcal{J}):=(\mathcal{I}\cdot J,\mu_\mathcal{I}+\mu_\mathcal{J})$.

Lemma 5.4.1. (1) $\mathrm{supp}((\mathcal{I}_1,\mu_1)+\ldots+(\mathcal{I}_m,\mu_m))=\mathrm{supp}(\mathcal{I}_1,\mu_1)\cap\ldots\cap\mathrm{supp}(\mathcal{I}_m,\mu_m)$. *Moreover multiple test blow-ups (M_k) of $(\mathcal{I}_1,\mu_1)+\ldots+(\mathcal{I}_m,\mu_m)$ are exactly those which are simultaneous multiple test blow-ups for all $(\mathcal{I}_j,\mu_j)$ and for any k we have the equality for the controlled transforms $(\mathcal{I}_j,\mu_\mathcal{I})_k$*

$$(\mathcal{I}_1,\mu_1)_k+\ldots+(\mathcal{I}_m,\mu_m)_k=[(\mathcal{I}_1,\mu_1)+\ldots+(\mathcal{I}_m,\mu_m)]_k$$

(2)

$$\mathrm{supp}(\mathcal{I},\mu_\mathcal{I})\cap\mathrm{supp}(\mathcal{J},\mu_\mathcal{J})\subseteq\mathrm{supp}((\mathcal{I},\mu_\mathcal{I})\cdot(\mathcal{J},\mu_\mathcal{J})).$$

Moreover any simultaneous multiple test blow-up M_i of both ideals $(\mathcal{I},\mu_\mathcal{I})$ and $(\mathcal{J},\mu_\mathcal{J})$ is a multiple test blow-up for $(\mathcal{I},\mu_\mathcal{I})\cdot(\mathcal{J},\mu_\mathcal{J})$, and for the controlled transforms $(\mathcal{I}_k,\mu_\mathcal{I})$ and $(\mathcal{J}_k,\mu_\mathcal{J})$ we have the equality

$$(\mathcal{I}_k,\mu_\mathcal{I})\cdot(\mathcal{J}_k,\mu_\mathcal{J})=[(\mathcal{I},\mu_\mathcal{I})\cdot(\mathcal{J},\mu_\mathcal{J})]_k.$$

Proof.

(1) Follows from two simple observations:

(i) $(\mathcal{I},\mu)\simeq(\mathcal{I}^k,k\mu)$

(ii) $\mathrm{supp}(\mathcal{I},\mu)\cap\mathrm{supp}(\mathcal{I}',\mu)=\mathrm{supp}(\mathcal{I}+\mathcal{I}',\mu)$ and the property is persistent for controlled transforms.

(2) Follows from the following fact:

If $\mathrm{ord}_x(\mathcal{I})\geq\mu_\mathcal{I}$ and $\mathrm{ord}_x(\mathcal{J})\geq\mu_\mathcal{J}$ then $\mathrm{ord}_x(\mathcal{I}\cdot\mathcal{J})\geq\mu_\mathcal{I}+\mu_\mathcal{J}$. This implies that $\mathrm{supp}(\mathcal{I},\mu_\mathcal{I})\cap\mathrm{supp}(\mathcal{J},\mu_\mathcal{J})\subseteq\mathrm{supp}((\mathcal{I},\mu_\mathcal{I})\cdot(\mathcal{J},\mu_\mathcal{J}))$. Then by induction we have the equality:

$$(\mathcal{I}_k,\mu_\mathcal{I})\cdot(\mathcal{J}_k,\mu_\mathcal{J})=[(\mathcal{I},\mu_\mathcal{I})\cdot(\mathcal{J},\mu_\mathcal{J})]_k.$$

□

5.5. Homogenized ideals and tangent directions

Let $(\mathcal{I},\mu)$ be a marked ideal of maximal order. Set $T(\mathcal{I}) := \mathcal{D}^{\mu-1}\mathcal{I}$. By the *homogenized ideal* we mean

$$\mathcal{H}(\mathcal{I},\mu) := (\mathcal{H}(\mathcal{I}),\mu) = (\mathcal{I} + \mathcal{D}\mathcal{I}\cdot T(\mathcal{I}) + \ldots + \mathcal{D}^i\mathcal{I}\cdot T(\mathcal{I})^i + \ldots + \mathcal{D}^{\mu-1}\mathcal{I}\cdot T(\mathcal{I})^{\mu-1},\mu).$$

Lemma 5.5.1. *Let $(\mathcal{I},\mu)$ be a marked ideal of maximal order.*

(1) *If $\mu = 1$, then $(\mathcal{H}(\mathcal{I}),1) = (\mathcal{I},1)$.*
(2) $\mathcal{H}(\mathcal{I}) = \mathcal{I}+\mathcal{D}\mathcal{I}\cdot T(\mathcal{I})+\ldots+\mathcal{D}^i\mathcal{I}\cdot T(\mathcal{I})^i+\ldots+\mathcal{D}^{\mu-1}\mathcal{I}\cdot T(\mathcal{I})^{\mu-1}+\mathcal{D}^\mu\mathcal{I}\cdot T(\mathcal{I})^\mu+\ldots$.
(3) $(\mathcal{H}(\mathcal{I}),\mu) = (\mathcal{I},\mu) + \mathcal{D}(\mathcal{I},\mu)\cdot(T(\mathcal{I}),1) + \ldots + \mathcal{D}^i(\mathcal{I},\mu)\cdot(T(\mathcal{I}),1)^i + \ldots + \mathcal{D}^{\mu-1}(\mathcal{I},\mu)\cdot(T(\mathcal{I}),1)^{\mu-1}$.
(4) *If $\mu > 1$ then $\mathcal{D}(\mathcal{H}(\mathcal{I},\mu)) \subseteq \mathcal{H}(\mathcal{D}(\mathcal{I},\mu))$.*
(5) $T(\mathcal{H}(\mathcal{I},\mu)) = T(\mathcal{I},\mu)$.

Proof. (1) $T(\mathcal{I}) = \mathcal{I}$ and $\mathcal{D}^i(\mathcal{I})T(\mathcal{I})^i \subseteq \mathcal{I}$. (2) $\mathcal{D}^{\mu-1}(\mathcal{I})T(\mathcal{I}) = T(\mathcal{I})^\mu$ and $\mathcal{D}^i(\mathcal{I})T(\mathcal{I})^i \subset T(\mathcal{I})^\mu$ for $i \geq \mu$. (3) By definition. (4) Note that $T(\mathcal{D}(\mathcal{I})) = T(\mathcal{I})$ and $\mathcal{D}(\mathcal{D}^i(\mathcal{I})T^i(\mathcal{I})^i) \subseteq \mathcal{D}^i(\mathcal{D}(\mathcal{I}))T(\mathcal{D}(\mathcal{I})) + \mathcal{D}^{i-1}(\mathcal{D}\mathcal{I})T(\mathcal{D}(\mathcal{I}))^{i-1} \subseteq \mathcal{H}(\mathcal{D}(\mathcal{I},\mu))$. (5) $T(\mathcal{I}) = \mathcal{D}^{\mu-1}(\mathcal{I}) \subseteq \mathcal{D}^{\mu-1}(\mathcal{H}(\mathcal{I})) \subseteq \mathcal{H}(\mathcal{D}^{\mu-1}(\mathcal{I})) = \mathcal{H}(T(\mathcal{I})) = T(\mathcal{I})$. □

Remark. A homogenized ideal features two important properties:

(1) It is equivalent to the given ideal.
(2) It "looks the same" from all possible tangent directions.

By the first property we can use the homogenized ideal to construct resolution via the Giraud Lemma 5.3.6. By the second property such a construction does not depend on the choice of tangent directions.

Lemma 5.5.2. *Let $(\mathcal{I},\mu)$ be a marked ideal of maximal order. Then*

(1) $(\mathcal{I},\mu) \simeq (\mathcal{H}(\mathcal{I}),\mu)$.
(2) *For any multiple test blow-up (M_k) of $(\mathcal{I},\mu)$,*
$(\mathcal{H}(\mathcal{I}),\mu)_k = (\mathcal{I},\mu)_k + [\mathcal{D}(\mathcal{I},\mu)]_k\cdot[(T(\mathcal{I}),1)]_k + \ldots + [\mathcal{D}^{\mu-1}(\mathcal{I},\mu)]_k\cdot[(T(\mathcal{I}),1)]_k^{\mu-1}$.

Proof. Since $\mathcal{H}(\mathcal{I}) \supset \mathcal{I}$, every multiple test blow-up of $\mathcal{H}(\mathcal{I},\mu)$ is a multiple test blow-up of $(\mathcal{I},\mu)$. By Lemma 5.2.4, every multiple test blow-up of $(\mathcal{I},\mu)$ is a multiple test blow-up for all $\mathcal{D}^i(\mathcal{I},\mu)$ and consequently, by Lemma 5.4.1 it is a simultaneous multiple test blow-up of all $(\mathcal{D}^i(\mathcal{I})\cdot T(\mathcal{I})^i,\mu) = (\mathcal{D}^i(\mathcal{I}),\mu-i)\cdot(T(\mathcal{I})^i,i)$ and

$$\begin{aligned}\operatorname{supp}(\mathcal{H}(\mathcal{I},\mu)_k) &= \textstyle\bigcap_{i=0}^{\mu-1}\operatorname{supp}(\mathcal{D}^i(\mathcal{I})\cdot T(\mathcal{I})^i,\mu)_k \\ &= \textstyle\bigcap_{i=0}^{\mu-1}\operatorname{supp}(\mathcal{D}^i(\mathcal{I}),\mu-i)_k\cdot(T(\mathcal{I})^i,i)_k \\ &\supseteq \textstyle\bigcap_{i=0}^{\mu-1}\operatorname{supp}(\mathcal{D}^i(\mathcal{I},\mu))_k = \operatorname{supp}(\mathcal{I}_k,\mu).\end{aligned}$$

Therefore every multiple test blow-up of $(\mathcal{I},\mu)$ is a multiple test blow-up of $\mathcal{H}(\mathcal{I},\mu)$ and by Lemmas 5.5.1(3) and 5.4.1 we get (2). □

Although the following Lemma 5.5.3 are used in this paper only in the case $E = \emptyset$ we formulate them in slightly more general versions.

Lemma 5.5.3. *(**Glueing Lemma**) Let $(M_Z, \mathcal{I}, E, \mu)$ be a marked ideal of maximal order. Assume there exist tangent directions $u, v \in T(\mathcal{I},\mu)_x = \mathcal{D}^{\mu-1}(\mathcal{I},\mu)$ at $x \in \text{supp}(\mathcal{I},\mu)$ which are transversal to E. Then there exists an open neighborhood V of x such that $\overline{V}$ is compact and an automorphism ϕ_{uv} of M_S where $S := Z \cap \overline{V}$ such that*

(1) $\phi^*_{uv}(\mathcal{H}\mathcal{I})_{|M_S} = \mathcal{H}\mathcal{I}_{|M_S}$.
(2) $\phi^*_{uv}(E) = E$.
(3) $\phi^*_{uv}(u) = v$.
(4) $\text{supp}(\mathcal{I},\mu) := V(T(\mathcal{I},\mu))$ *is contained in the fixed point set of* ϕ.
(5) *Any test resolution* M_{iS_i} *of* $(M_S, \mathcal{I}, E, \mu)$ *is equivariant with respect to* ϕ_{uv} *and moreover the properties (1)-(4) are satisfied for the lifting* $\phi_{uvi} : M_{iS_i} \to M_{iS_i}$ *of* $\phi_{uv} : M_S \to M_S$ *and the induced marked ideal* $\mathcal{H}\mathcal{I})_i$.

Proof. (0) **Construction of the automorphism ϕ_{uv}.**

Find coordinates $u_2, \ldots, u_n$ transversal to u and v such that $u = u_1, u_2, \ldots, u_n$ and $v, u_2, \ldots, u_n$ form two sets of coordinates at x and divisors in E are described by some coordinates u_i where $i \geq 2$. Set

$$\phi_{uv}(u_1) = v, \quad \phi_{uv}(u_i) = u_i \quad \text{for} \quad i > 1.$$

The morphism $\phi_{uv} : U \to U'$ defines an open embedding from some neighborhood U of x to another neighborhood U' of x.

(1) Let $h := v - u \in T(\mathcal{I})$. For any $f \in \mathcal{I}$,

$$\phi^*_{uv}(f) = f(u_1+h, u_2, \ldots, u_n) = f(u_1, \ldots, u_n) + \frac{\partial f}{\partial u_1} \cdot h + \frac{1}{2!}\frac{\partial^2 f}{\partial u_1^2} \cdot h^2 + \ldots + \frac{1}{i!}\frac{\partial^i f}{\partial u_1^i} \cdot h^i + \ldots$$

The latter element belongs to

$$\mathcal{I} + \mathcal{D}\mathcal{I} \cdot T(\mathcal{I}) + \ldots + \mathcal{D}^i\mathcal{I} \cdot T(\mathcal{I})^i + \ldots + \mathcal{D}^{\mu-1}\mathcal{I} \cdot T(\mathcal{I})^{\mu-1} = \mathcal{H}\mathcal{I}.$$

Hence $\phi^*_{uv}(\mathcal{I}) \subset \mathcal{H}\mathcal{I}$. Analogously $\phi^*_{uv}(\mathcal{D}^i\mathcal{I}) \subset \mathcal{D}^i\mathcal{I} + \mathcal{D}^{i+1}\mathcal{I} \cdot T(\mathcal{I}) + \ldots + \mathcal{D}^{\mu-1}\mathcal{I} \cdot T(\mathcal{I})^{\mu-i-1} = \mathcal{H}\mathcal{D}^i I$. In particular by Lemma 5.5.1, $\phi^*_{uv}(T(\mathcal{I}), 1) \subset \mathcal{H}(T(\mathcal{I}), 1) = (T(\mathcal{I}), 1)$. This gives

$$\phi^*_{uv}(\mathcal{D}^i\mathcal{I} \cdot T(\mathcal{I})^i) \subset \mathcal{D}^i\mathcal{I} \cdot T(\mathcal{I})^i + \ldots + \mathcal{D}^{\mu-1}\mathcal{I} \cdot T(\mathcal{I})^{\mu-1} \subset \mathcal{H}\mathcal{I}.$$

By the above $\phi^*_{uv}(\mathcal{H}\mathcal{I})_x \subset (\mathcal{H}\mathcal{I})_x$ and since the scheme is noetherian, $\phi^*_{uv}(\mathcal{H}\mathcal{I})_x = (\mathcal{H}\mathcal{I})_x$. Consequently $\phi^*_{uv}(\mathcal{H}\mathcal{I})_y = (\mathcal{H}\mathcal{I})_y$ for all points y in some neighborhood $V \subset U$ of x. We can assume that $\overline{V} \subset U$ is compact.

(2)(3) Follow from the construction.

(4) The fixed point set of ϕ_{uv} is defined by $u_i = \phi^*_{uv}(u_i)$, $i = 1, \ldots, n$, that is, $h = 0$. But $h \in \mathcal{D}^{\mu-1}(\mathcal{I})$ is 0 on $\text{supp}(\mathcal{I},\mu)$. In particular ϕ_{uv} defines an automorphism of M_S identical on $S = \overline{V} \cap M$.

(5) Let $C_0 \subset \text{supp}(\mathcal{I},\mu)$ be the center of σ_0. Then we can find coordinates $u'_1, u'_2, \ldots, u'_n$ transversal to $u = u'_1$ and $v = u + h$ such that C_0 is described by coordinates $u'_1 = u'_2 =$

$\ldots = u'_m = 0$ for some $m \geq 0$ or equivalently $v = u'_2 = \ldots = u'_m = 0$. By (4), the automorphism ϕ_{uv} is described by

$$\phi_{uv}(u'_i) = u'_i + h'_i, \quad \text{where} \quad h'_i \in (h) \in T(\mathcal{I}) \subset \mathcal{D}^{\mu-1}\mathcal{I}.$$

By (3), C is invariant with respect to ϕ_{uv} and it lifts to an automorphism ϕ_{uv1} of M_1. Note also that at any point $p \in \sigma_0^{-1}(x) \cap \operatorname{supp}(\mathcal{I}_1, \mu)$ there is a set of coordinates $u''_1, u''_2, \ldots, u''_n$ where $u''_i = \frac{u'_i}{u'_m}$, $u''_i = u'_i$ for $i > m$. Then the form of ϕ_{uv1} is the same as ϕ_{uv}.

$$\phi_{uv1}(u''_i) = u''_1 + h''_i, \quad \text{where} \quad h'' \in T(\mathcal{I})_1 \subset \mathcal{D}^{\mu-1}\mathcal{I}_1$$

The fixed point set of ϕ_{uv} is defined by $h'' = 0$ in a neighborhood U_p of p and it contains $\operatorname{supp}(\mathcal{I}_1, \mu) \cap U_p$. In particular all points $p \in \operatorname{supp}(\mathcal{I}_1, \mu) \cap (\sigma_1)^{-1}(x)$ are fixed under ϕ_{uv1}. Thus ϕ_{uv1} defines an automorphism of $M_{1,S_1} = \sigma_1^{-1}(M_S)$. We continue the reasoning by induction. □

5.6. Coefficient ideals and Giraud Lemma

The idea of coefficient ideals was originated by Hironaka and then developed in papers of Villamayor and Bierstone-Milman. The following definition modifies and generalizes the definition of Villamayor.

Definition 5.6.1. Let $(\mathcal{I}, \mu)$ be a marked ideal of maximal order. By the *coefficient ideal* we mean

$$\mathcal{C}(\mathcal{I}, \mu) = (\mathcal{I}, \mu) + (\mathcal{D}\mathcal{I}, \mu - 1) + \ldots + (\mathcal{D}^{\mu-1}\mathcal{I}, 1).$$

Remark. The coefficient ideals $\mathcal{C}(\mathcal{I})$ feature two important properties.

(1) $\mathcal{C}(\mathcal{I})$ is equivalent to $\mathcal{I}$.
(2) The intersection of the support of $(\mathcal{I}, \mu)$ with any submanifold S is the support of the restriction of $\mathcal{C}(\mathcal{I})$ to S:

$$\operatorname{supp}(\mathcal{I}) \cap S = \operatorname{supp}(\mathcal{C}(\mathcal{I})_{|S}).$$

Moreover this condition is persistent under relevant multiple test blow-ups.

These properties allow one to control and modify the part of support of $(\mathcal{I}, \mu)$ contained in S by applying multiple test blow-ups of $\mathcal{C}(\mathcal{I})_{|S}$.

Lemma 5.6.2. $\mathcal{C}(\mathcal{I}, \mu) \simeq (\mathcal{I}, \mu)$.

Proof. By Lemma 5.4.1 multiple test blow-ups of $\mathcal{C}(\mathcal{I}, \mu)$ are simultaneous multiple test blow-ups of $\mathcal{D}^i(\mathcal{I}, \mu)$ for $0 \leq i \leq \mu - 1$. By Lemma 5.2.4 multiple test blow-ups of $(\mathcal{I}, \mu)$ define a multiple test blow-up of all $\mathcal{D}^i(\mathcal{I}, \mu)$. Thus multiple test blow-ups of $(\mathcal{I}, \mu)$ and $\mathcal{C}(\mathcal{I}, \mu)$ are the same and $\operatorname{supp}(\mathcal{C}(\mathcal{I}, \mu))_k = \bigcap \operatorname{supp}(\mathcal{D}^i\mathcal{I}, \mu - i)_k = \operatorname{supp}(\mathcal{I}_k, \mu)$. □

Lemma 5.6.3. *Let $(M_Z, \mathcal{I}, E, \mu)$ be a marked ideal of maximal order whose support $\operatorname{supp}(\mathcal{I}, \mu)$ does not contain a submanifold S of M_Z. Assume that S has only simple normal crossings with E. Then*

$$\operatorname{supp}(\mathcal{I}, \mu) \cap S \subseteq \operatorname{supp}((\mathcal{I}, \mu)_{|S}).$$

Proof. The order of an ideal does not drop but may rise after restriction to a submanifold. □

Proposition 5.6.4. *Let $(M_Z, \mathcal{I}, E, \mu)$ be a marked ideal of maximal order whose support $\mathrm{supp}(\mathcal{I}, \mu)$ does not contain the germ of a submanifold S_T of M_Z. Assume that S has only simple normal crossings with E and $T := Z \cap S$. Let $E' \subset E$ be the set of divisors transversal to S. Set $E'_{|S} := \{D \cap S \mid D \in E'\}$, $\mu_c := \mathrm{lcm}(1, 2, \ldots, \mu)$, and consider the marked ideal $\mathcal{C}(\mathcal{I}, \mu)_{|S} = (S, \mathcal{C}(\mathcal{I}, \mu)_{|S}, E'_{|S}, \mu_c)$. Then*

$$\mathrm{supp}(\mathcal{I}, \mu) \cap S = \mathrm{supp}(\mathcal{C}(\mathcal{I}, \mu)_{|S}).$$

Moreover let (M_{iZ_i}) be a multiple test blow-up with centers C_i contained in the strict transforms $S_i \subset M_i$ of S. Then

1. *The restrictions $\sigma_{i|S_i} : S_{iT_i} \to S_{i-1T_{i-1}}$ of the morphisms $\sigma_i : M_{iZ_i} \to M_{i-1Z_{i-1}}$ define a multiple test blow-up (S_{iT_i}) of $\mathcal{C}(\mathcal{I}, \mu)_{|S_T}$ (where $T_i := Z_i \cap S_i$.)*
2. *$\mathrm{supp}(\mathcal{I}_i, \mu) \cap S_i = \mathrm{supp}[\mathcal{C}(\mathcal{I}, \mu)_{|S}]_i$.*
3. *Every multiple test blow-up (S_{iT_i}) of $\mathcal{C}(\mathcal{I}, \mu)_{|S}$ defines a multiple test blow-up (M_{iZ_i}) of $(\mathcal{I}, \mu)$ with centers C_i contained in the strict transforms $S_{iT_i} \subset M_{iZ_i}$ of $S_T \subset M_T$.*

Proof. By Lemmas 5.6.2 and 5.6.3, $\mathrm{supp}(\mathcal{I}, \mu) \cap S = \mathrm{supp}(\mathcal{C}(\mathcal{I}, \mu)) \cap S \subseteq \mathrm{supp}(\mathcal{C}(\mathcal{I}, \mu)_{|S})$.

Let $x_1, \ldots, x_k, y_1, \ldots, y_{n-k}$ be local coordinates at p such that $\{x_1 = 0, \ldots, x_k = 0\}$ describes S. Then write a function $f \in \mathcal{I}$ can be written as

$$f = \sum c_{\alpha f}(y) x^\alpha.$$

Now $x \in \mathrm{supp}(\mathcal{I}, \mu) \cap S$ iff $\mathrm{ord}_x(c_{\alpha,f}) \geq \mu - |\alpha|$ for all $f \in \mathcal{I}$ and $0 \leq |\alpha| < \mu$. Note that

$$c_{\alpha f|S} = \left(\frac{1}{\alpha!} \frac{\partial^{|\alpha|}(f)}{\partial x^\alpha} \right)_{|S} \in \mathcal{D}^{|\alpha|}(\mathcal{I})_{|S}$$

and hence $\mathrm{supp}(\mathcal{I}, \mu) \cap S = \bigcap_{f \in \mathcal{I}, |\alpha| \leq \mu} \mathrm{supp}(c_{\alpha f|S}, \mu - |\alpha|) \supseteq \bigcap_{0 \leq i < \mu} \mathrm{supp}((\mathcal{D}^i \mathcal{I})_{|S}) = \mathrm{supp}(\mathcal{C}(\mathcal{I}, \mu)_{|S})$.

Assume that all multiple test blow-ups of $(\mathcal{I}, \mu)$ of length k with centers $C_i \subset S_i$ are defined by multiple test blow-ups of $\mathcal{C}(\mathcal{I}, \mu)_{|S}$ and moreover for $i \leq k$,

$$\mathrm{supp}(\mathcal{I}_i, \mu) \cap S_i = \mathrm{supp}[\mathcal{C}(\mathcal{I}, \mu)_{|S}]_i.$$

For any $f \in \mathcal{I}$ define $f = f_0 \in \mathcal{I}$ and $f_{i+1} = \sigma_i^{\mathrm{c}}(f_i) = y_i^{-\mu} \sigma^*(f_i) \in \mathcal{I}_{i+1}$. Assume that

$$f_k = \sum c_{\alpha f k}(y) x^\alpha,$$

where $c_{\alpha f k|S_k} \in (\sigma^k_{|S_k})^{\mathrm{c}}(\mathcal{D}^{\mu - |\alpha|}(\mathcal{I})_{|S})$. Consider the effect of the blow-up of C_k at a point p_{k+1} in the strict transform $S_{k+1} \subset M_{k+1}$. By Lemmas 5.6.2 and 5.6.3,

$$\begin{aligned} \mathrm{supp}(\mathcal{I}_{k+1}, \mu) \cap S_{k+1} &= \mathrm{supp}[\mathcal{C}(\mathcal{I}, \mu)]_{k+1} \cap S_{k+1} \\ &\subseteq \mathrm{supp}[\mathcal{C}(\mathcal{I}, \mu)]_{k+1|S_{k+1}} = \mathrm{supp}[\mathcal{C}(\mathcal{I}, \mu)_{|S}]_{k+1} \end{aligned}$$

Let $x_1, \ldots, x_k$ describe the submanifold S_k of M_k. We can find coordinates $x_1, \ldots, x_k$, $y_1, \ldots, y_{n-k}$ at the point p_k, by taking if necessary linear combinations of $y_1, \ldots, y_{n-k}$,

such that the center of the blow-up is described by $x_1, \ldots, x_k, y_1, \ldots, y_m$ and the coordinates at p_{k+1} are given by

$$x'_1 = x_1/y_m, \ldots, x'_k = x_k/y_m, y'_1 = y_1/y_m, \ldots, y'_m = y_m, y'_{m+1} = y_{m+1}, \ldots, y'_n = y_n.$$

Note that replacing $y_1, \ldots, y_{n-k}$ with their linear combinations does not modify the form $f_k = \sum c_{\alpha f k}(y)x^\alpha$. Then the function $f_{k+1} = \sigma^{\mathrm{c}}(f_k)$ can be written as

$$f_{k+1} = \sum c_{\alpha f,k+1}(y){x'}^\alpha,$$

where $c_{\alpha f k+1} = y_m^{-\mu+|\alpha|}\sigma^*_{k+1}(c_{\alpha f k})$. Thus

$$c_{\alpha f k+1|S_{k+1}} = (\sigma_{k+1|S_{k+1}})^{\mathrm{c}}(c_{\alpha f k|S_k}) \in (\sigma^{k+1}_{|S_{k+1}})^{\mathrm{c}}(\mathcal{D}^{\mu-|\alpha|}(\mathcal{I})_{|S}) = (\sigma^{k+1})^{\mathrm{c}}(\mathcal{D}^{\mu-|\alpha|}(\mathcal{I}))_{|S_{k+1}}$$

and consequently

$$\begin{array}{rcl} \mathrm{supp}(\mathcal{I}_{k+1}, \mu) \cap S_{k+1} & = & \bigcap_{f\in\mathcal{I}, |\alpha|\leq\mu} \mathrm{supp}(c_{\alpha f k+1|S_{k+1}}, \mu - |\alpha|) \\ & \supseteq & \mathrm{supp}[\mathcal{C}(\mathcal{I}, \mu)_{|S}]_{k+1} = \mathrm{supp}(\mathcal{C}(\mathcal{I}, \mu)_{k+1})_{|S_{k+1}}. \end{array}$$

□

As a simple consequence of Lemma 5.6.4 we formulate the following refinement of the Giraud Lemma.

Lemma 5.6.5. *Let $(M_Z, \mathcal{I}, \emptyset, \mu)$ be a marked ideal of maximal order whose support $\mathrm{supp}(\mathcal{I}, \mu)$ has codimension at least 2 at some point x. Let $U \ni x$ be an open subset for which there is a tangent direction $u \in T(\mathcal{I})$ and such that $\mathrm{supp}(\mathcal{I}, \mu) \cap U$ is of codimension at least 2. Let $V(u)$ be the regular subscheme of U defined by u. Then for any multiple test blow-up (M_{iZ_i}) of M_Z,*

1. *$\mathrm{supp}(\mathcal{I}_i, \mu)$ is contained in the strict transform $V(u)_{iT_i}$ of $V(u)_T$ as a proper subset (where $T = Z \cap V(u)$ and $T_i = Z_i \cap V(u)_i$).*
2. *The sequence $(V(u)_{iT_i})$ is a multiple test blow-up of $\mathcal{C}(\mathcal{I}, \mu)_{|V(u)_T}$.*
3. *$\mathrm{supp}(\mathcal{I}_i, \mu) \cap V(u)_{iT_i} = \mathrm{supp}[\mathcal{C}(\mathcal{I}, \mu)_{|V(u)_T}]_i$.*
4. *Every multiple test blow-up $(V(u)_{iT_i})$ of $\mathcal{C}(\mathcal{I}, \mu)_{|V(u)_T}$ defines a multiple test blow-up (M_{iZ_i}) of $(\mathcal{I}, \mu)$.*

□

6. Algorithm for canonical resolution of marked ideals

The presentation of the following resolution algorithm builds upon Villamayor's and Bierstone-Milman's proofs.

Theorem 6.0.6. *For any marked ideal $(M_Z, \mathcal{I}, E, \mu)$ such that $\mathcal{I} \neq 0$ there is an associated resolution $(M_{iZ_i}))_{0\leq i\leq m_M}$, called canonical, satisfying the following conditions:*

1. *For any surjective local analytic isomophism $\phi : M'_{Z'} \to M_Z$ the induced sequence $(M'_{iZ'_i}) = \phi^*(M_{iZ_i})$ is the canonical resolution of M'.*
2. *For any local analytic isomophism $\phi : M' \to M$ the induced sequence $(M'_{iZ_i}) = \phi^*(M_{iZ_i})$ is an extension of the canonical resolution of $M'_{Z'}$.*

Remarks. (1) In Step 2 we resolve general marked ideals by reducing the algorithm to resolving some marked ideals of maximal order (companion ideals).

(2) In Step 1 we resolve marked ideals of maximal order. It is the heart part of the algorthm.
(3) The main idea of the algorithm of resolving marked ideals of maximal order in Step 1 is to reduce the procedure to the hypersurface of maximal contact (Step 1b).
(4) By Lemma 5.3.4 hypersurfaces of maximal contact can be constructed locally. They are in general not transversal to E and can not be used for the reduction procedure. We think of E and its strict transforms as an obstacle to existence of a hypersurface of maximal contact (transversal to E). These divisors are often referred to as "old" ones.
(5) In Step 1a we move "old" divisors apart from the support of the marked ideal. In this process we create "new" divisors but these divisors are "born" from centers lying in the hypersurface of maximal contact. The "new" divisors are transversal to hypersurfaces of maximal contact. After eliminating "old" divisors from the support in Step 1a all divisors are "new" and we may reduce the resolving procedure to hypersurfaces of maximal contact (Step 1b).

Proof. Induction on the dimension of M_Z. If M is 0-dimensional, $\mathcal{I} \neq 0$ and $\mu > 0$ then $\operatorname{supp}(M, \mathcal{I}, \mu) = \emptyset$ and all resolutions are trivial.

Step 1. **Resolving a marked ideal** $(M_Z, \mathcal{J}, E, \mu)$ **of maximal order.**

Before we start our resolution algorithm for the marked ideal $(\mathcal{J}, \mu)$ of maximal order we shall replace it with the equivalent homogenized ideal $\mathcal{C}(\mathcal{H}(\mathcal{J}, \mu))$. Resolving the ideal $\mathcal{C}(\mathcal{H}(\mathcal{J}, \mu))$ defines a resolution of $(\mathcal{J}, \mu)$ at this step. To simplify notation we shall denote $\mathcal{C}(\mathcal{H}(\mathcal{J}, \mu))$ by $(\overline{\mathcal{J}}, \overline{\mu})$.

Step 1a. Reduction to the nonboundary case. For any multiple test blow-up (M_{iZ_i}) of $(M_Z, \overline{\mathcal{J}}, E, \overline{\mu})$ we shall identify (for simplicity) strict transforms of E on M_{iZ_i} with E. For any $x \in Z_i$, let $s(x)$ denote the number of divisors in E through x and set

$$s_i = \max\{s(x) \mid x \in \operatorname{supp}(\overline{\mathcal{J}}_i) \cap Z_i\}.$$

Let $s = s_0$. By assumption the intersections of any $s > s_0$ components of the exceptional divisors are disjoint from $\operatorname{supp}(\overline{\mathcal{J}}, \overline{\mu})$. Each intersection of divisors in E on M_Z is locally defined by intersection of some irreducible components of these divisors. Find all intersections $H^s_\alpha \subset M_Z, \alpha \in A$, of s irreducible components of divisors E such that $\operatorname{supp}(\overline{\mathcal{J}}, \overline{\mu}) \cap H^s_\alpha \cap Z \neq \emptyset$. By the maximality of s, the supports $\operatorname{supp}(\overline{\mathcal{J}}_{|H^s_\alpha}) \subset H^s_\alpha$ are disjoint from $H^s_{\alpha'}$ (in a neighborhood of Z), where $\alpha' \neq \alpha$.

Step 1aa. Eliminating the components H^s_α **contained in** $\operatorname{supp}(\overline{\mathcal{J}}, \overline{\mu})$**.**

Let $H^s_\alpha \subset \operatorname{supp}(\overline{\mathcal{J}}, \overline{\mu})$ (in a neighborhood of Z). If $s \geq 2$ then by blowing up $C = H^s_a$ we separate divisors contributing to H^s_a, thus creating new points all with $s(x) < s$. If $s = 1$ then by Lemma 5.3.7, $H^s_\alpha \subset \operatorname{supp}(\overline{\mathcal{J}}, \overline{\mu})$ is a codimension one component and by blowing up H^s_α we create all new points off $\operatorname{supp}(\overline{\mathcal{J}}, \overline{\mu})$.

Note that all $H^s_\alpha \subset \operatorname{supp}(\overline{\mathcal{J}}, \overline{\mu})$ will be blown up first and we reduce the situation to the case where no H^s_α is contained in $\operatorname{supp}(\overline{\mathcal{J}}, \overline{\mu})$.

Step 1ab. Moving $\operatorname{supp}(\overline{\mathcal{J}}, \overline{\mu})$ **and** H^s_α **apart.**

After the blow-up in Step 1aa we arrive at M_{pZ_p} for which no H^s_α is contained in $\mathrm{supp}(\overline{\mathcal{J}}_p, \overline{\mu})$ (in a neighborhood of Z), where $p = 0$ if there were no such components and $p = 1$ if there were some. Let $U^s_\alpha := M_p \setminus \bigcup_{\beta\neq\alpha} H^s_\beta \quad Z^s_\alpha := Z \cap H^s_\alpha \cap \mathrm{supp}(\overline{\mathcal{J}}_p, \overline{\mu})$. Note that by the maximality condition for s all $H^s_\alpha \cap \mathrm{supp}(\overline{\mathcal{J}}_p, \overline{\mu})$ are disjoint for two different $\alpha \in A_s$. By definition $Z^s_\alpha \subset \mathrm{supp}(\overline{\mathcal{J}}_p, \overline{\mu}) \cap H^s_\alpha \subset U^s_\alpha$ is compact. Set

$$\widetilde{Z}^s = \coprod Z^s_\alpha \quad Z^s = \bigcup Z^s_\alpha = Z \cap \mathrm{supp}(\overline{\mathcal{J}}_p, \overline{\mu}) \quad \widetilde{M}_p := \coprod U^s_\alpha \qquad \widetilde{H^s} := \coprod H^s_\alpha \cap U^s_\alpha$$

Consider the surjective local analytic isomorphism $\phi : \widetilde{M}_p := \coprod U^s_\alpha \to M_p$. Note that Z^s_α is disjoint from $U^s_{\alpha'}$, where $\alpha' \neq \alpha$. The morphism ϕ defines a morphisms of germs $\phi_Z : \widetilde{M}_{\widetilde{Z^s}} \to M_{Z^s}$ which is locally an isomorphism

$$\widetilde{M}_{\widetilde{Z^s}} \supseteq \phi^{-1}(U^s_{\alpha Z^s_\alpha}) \simeq U^s_{\alpha Z^s_\alpha} \subseteq M_{Z^s}.$$

Denote by $\widetilde{J}$ the pull back of the ideal sheaf $\mathcal{J}$ via ϕ_Z. The closed embeddings $H^s_\alpha \cap U^s_\alpha \subset U^s_\alpha$ define the closed embedding $\widetilde{H^s} \subset \widetilde{M}$. Let $Z_H := Z \cap H$.

Construct by the inductive assumption the canonical resolution $(\widetilde{H}^{s_i}_{iZ_{Hi}})$ of $\widetilde{\mathcal{J}}_{p|\widetilde{H^s}}$. By Proposition 5.6.4 such a resolution defines a multiple test blow-up $(\widetilde{M}_{iZ_i})$ of $(\widetilde{\mathcal{J}}_p, \overline{\mu})$ (and of $(\overline{\mathcal{J}}, \overline{\mu})$). By Proposition 5.6.4,

$$\mathrm{supp}((\widetilde{\mathcal{J}}_i, \overline{\mu})_{|\widetilde{H}^s}) = \mathrm{supp}(\widetilde{\mathcal{J}}_i, \overline{\mu}) \cap \widetilde{H}^s.$$

Descending the multiple test blow-up to M_{Z^s}, defines a multiple test blow-up of $(\overline{\mathcal{J}}_i, \overline{\mu})$ such that

$$\mathrm{supp}((\overline{\mathcal{J}}_i, \overline{\mu})_{|H^s_\alpha}) = \mathrm{supp}(\overline{\mathcal{J}}_i, \overline{\mu}) \cap H^s_\alpha.$$

This creates a marked ideal $(\overline{\mathcal{J}}_{j_1}, \overline{\mu})$ with support disjoint from all H^s_α.

Conclusion of the algorithm in Step 1a. After performing the blow-ups in Steps 1aa and 1ab for the marked ideal $(\overline{\mathcal{J}}, \overline{\mu})$ we arrive at a marked ideal $(\overline{\mathcal{J}}_{j_1}, \overline{\mu})$ with $s_{j_1} < s_0$. Now we put $s = s_{j_1}$ and repeat the procedure of Steps 1aa and 1ab for $(\overline{\mathcal{J}}_{j_1}, \overline{\mu})$. Note that any $H^s_{\alpha j_1}$ on M_{j_1} is the strict transform of some intersection $H^{s_{j_1}}_\alpha$ of $s = s_{j_1}$ divisors in E on M. Moreover by the maximality condition for all s_i, where $i \leq j_1$ and $\alpha \neq \alpha'$, the set $\mathrm{supp}(\overline{\mathcal{J}}_i, \overline{\mu}) \cap H^{s_i}_{\alpha' i}$ is either disjoint from $H^{s_{j_1}}_{\alpha i}$ or contained in it. Thus for $0 \leq i \leq j_1$, all centers C_i have components either contained in $H^{s_{j_1}}_{\alpha i} = H^s_{\alpha i}$ or disjoint from them and by Proposition 5.6.4,

$$\mathrm{supp}((\overline{\mathcal{J}}_i, \overline{\mu})_{|H^s_{\alpha i}}) = \mathrm{supp}(\overline{\mathcal{J}}_i, \overline{\mu}) \cap H^s_{\alpha i}.$$

Moreover if we repeat the procedure in Steps 1aa and 1ab the above property will still be satisfied until either $(\overline{\mathcal{J}}_i, \overline{\mu})_{|H^s_\alpha}$ are resolved as in Step 1ab or H^s_α disappear as in Step 1aa.

We continue the above process till $s_{j_k} = s_r = 0$. Then $(M_j)_{0\leq j\leq r}$ is a multiple test blow-up of $(M, \overline{\mathcal{J}}, E, \overline{\mu})$ such that $\mathrm{supp}(\overline{\mathcal{J}}_r, \overline{\mu})$ does not intersect any divisor in E. Therefore $(M_j)_{0\leq j\leq r}$ and further longer multiple test blow-ups $(M_j)_{0\leq j\leq r_0}$ for any $r \leq r_0$ can be considered as multiple test blow-ups of $(M, \overline{\mathcal{J}}, \emptyset, \overline{\mu})$ since starting from M_r the

strict transforms of E play no further role in the resolution process since they do not intersect $\mathrm{supp}(\overline{\mathcal{J}}_j, \overline{\mu})$ for $j \geq r$.

Step 1b. Nonboundary case

Let $(M_{jZ_j})_{0\leq j\leq r}$ be the multiple test blow-up of $(M, \overline{\mathcal{J}}, \emptyset, \overline{\mu})$ defined in Step 1a.

Step 1ba. Eliminating the codimension one components of $\mathrm{supp}(\overline{\mathcal{J}}_r, \overline{\mu})$**.**

If $\mathrm{supp}(\overline{\mathcal{J}}_r, \overline{\mu})$ is of codimension 1 then by Lemma 5.3.7 all its codimension 1 components are smooth and disjoint from the other components of $\mathrm{supp}(\overline{\mathcal{J}}_r, \overline{\mu})$. These components are strict transforms of the codimension 1 components of $\mathrm{supp}(\overline{\mathcal{J}}, \overline{\mu})$. Moreover the irreducible components of the centers of blow-ups were either contained in the strict transforms or disjoint from them. Therefore E_r will be transversal to all the codimension 1 components. Let $\mathrm{codim}(1)(\mathrm{supp}(\overline{\mathcal{J}}_i, \overline{\mu}))$ be the union of all components of $\mathrm{supp}(\overline{\mathcal{J}}_i, \overline{\mu}))$ of codimension 1. By Lemma 5.3.7 blowing up the components reduces the situation to the case when $\mathrm{supp}(\overline{\mathcal{J}}, \overline{\mu})$ is of codimension ≥ 2.

Step 1bb. Eliminating the codimension ≥ 2 **components of** $\mathrm{supp}(\overline{\mathcal{J}}_r, \overline{\mu})$**.**

For any $x \in Z \cap \mathrm{supp}(\overline{\mathcal{J}}, \overline{\mu}) \setminus \mathrm{codim}(1)(\mathrm{supp}(\overline{\mathcal{J}}, \overline{\mu})) \subset M_Z$ find a tangent direction $u_\alpha \in \mathcal{D}^{\overline{\mu}-1}(\overline{\mathcal{J}})$ on some neighborhood U_α of x. Then $H_\alpha := V(u_\alpha) \subset U_\alpha$ is a hypersurface of maximal contact. Take a finite open covers (U_α) and (V_α) of Z such that the ideal sheaf is defined on each U_α, $\overline{V_\alpha} \subset U_\alpha$ is compact , and U_α satisfies the property of Glueing Lemma. Let $Z_\alpha := Z \cap \overline{V_\alpha}$ and $Z_{V,\alpha} \subset Z_\alpha$ be any compact set contained in $Z \cap V_\alpha$. Set

$$\widetilde{V} := \coprod \overline{V}_\alpha \qquad \widetilde{Z}_V := \coprod Z_{V\alpha} \quad \widetilde{M} := \coprod U_\alpha \qquad \widetilde{Z} := \coprod Z_\alpha \quad \widetilde{H} := \coprod H_\alpha \subseteq \widetilde{M}$$

The closed embeddings $H_\alpha \subseteq U_\alpha$ define the closed embedding $\widetilde{H} \subset \widetilde{M}$ of a hypersurface of maximal contact $\widetilde{H}$.

Consider the surjective local analytic isomorphism

$$\phi_U : \widetilde{M} := \coprod U^s_\alpha \to M.$$

It defines a morphism of germs $\phi_{Z_V} : \widetilde{M}_{\widetilde{Z}} \to M_{\overline{Z}}$. Denote by $\widetilde{J}$ the pull back of the ideal sheaf $\overline{\mathcal{J}}$ via ϕ_U. The multiple test blow-up $(M_{iZ_i})_{0\leq i\leq p}$ of $\overline{J}$ defines a multiple test blow-up $(\widetilde{M}_{\widetilde{Z}_i i})_{0\leq i\leq p}$ of $\widetilde{J}$ and a multiple test blow-up $(\widetilde{\widetilde{H}}_i)_{0\leq i\leq p}$ of $\widetilde{J}_H$.

Let $U_{\alpha,i} \subset M_i$ be the inverse image of U_α and let $H_{\alpha i} \subset U_{\alpha i}$ denote the strict transform of H_α. By Lemma 5.6.5, $(H_{\alpha i})_{0\leq i\leq p}$ is a multiple test blow-up of $(H_\alpha, \overline{\mathcal{J}}_{|H_\alpha}, \emptyset, \overline{\mu})$. In particular the induced marked ideal for $i = p$ is equal to

$$\overline{\mathcal{J}}_{p|H_{\alpha p}} = (H_{\alpha p}, \overline{\mathcal{J}}_{p|H_{\alpha p}}, (E_p \setminus E)_{|H_{\alpha p}}, \overline{\mu}).$$

Construct the canonical resolution of $(\widetilde{H}_{iZ_i})_{p\leq i\leq m_u}$ of the marked ideal $\widetilde{\mathcal{J}}_{p|\widetilde{H}_p}$ on $\widetilde{H}_{\widetilde{Z}}$. It defines, by Lemma 5.6.5, a resolution $(\widetilde{M}_{i\widetilde{Z}_i})_{p\leq i\leq m}$ of $\widetilde{\mathcal{J}}_p$ and thus also a resolution $(\widetilde{M}_{i\widetilde{Z}_i})_{0\leq i\leq m}$ of $(\widetilde{M_Z}, \widetilde{\mathcal{J}}, \emptyset, \overline{\mu})$. Moreover both resolutions are related by the property

$$\mathrm{supp}(\widetilde{\mathcal{J}}_i) = \mathrm{supp}(\widetilde{\mathcal{J}}_{i|\widetilde{H}_i}).$$

The resolution $(\widetilde{M}_{i\widetilde{Z}_i})_{0\leq i\leq m}$ defines the canonical resolution $(\widetilde{V}_{i\widetilde{Z}_{Vi}})_{0\leq i\leq m}$ for any compact $Z_V \subset \widetilde{V}$.

Consider the surjective local analytic isomorphism

$$\phi_V : \widetilde{V} := \coprod V_\alpha \to M.$$

We have to show that the resolution $(\widetilde{V}_{i\widetilde{Z_{Vi}}})_{0\leq i\leq m}$ descends to the resolution $(M_{iZ_i})_{0\leq i\leq m}$ which is independent of the choice of local hypersurfaces of maximal contact and $\widetilde{M}$. We show by induction that there exists a resolution $(M_{iZ_{V\,i}}))_{0\leq i\leq m}$ such that its restriction $((H_\alpha)_{iZ_{V\,i}^{\,\alpha}}))_{k\leq i\leq m}$ is an extension of the part of the canonical resolution.

Consider the inverse image

$$\phi_j^{-1}(V_{\beta,i}) = \coprod V_{\beta,j} \cap V_{\alpha,j}.$$

Let $\widetilde{C}_j$ be the center of the blow-up $\widetilde{\sigma}_j : \widetilde{V}_{j+1} \to \widetilde{V}_j$. If $\widetilde{C}_j \cap V_{\beta,j} \cap V_{\alpha,j} \neq \emptyset$ then $\widetilde{C}_j \cap V_{\beta,j}$ defines the center of an extension of the part of the canonical resolution $((H_{\beta j})_{Z_{V\beta j}})_{p\leq j\leq m}$. By the canonicity the intersection $\widetilde{C}_j \cap V_{\beta,j} \cap V_{\alpha,j}$ defines the center of an extension of the part of the canonical resolution $((H_{\beta j} \cap V_{\alpha j})_{Z_{V\beta j}\cap Z_{V\beta j}})_{p\leq j\leq m}$.

By Glueing Lemma 5.5.3 for the tangent directions u_α and u_β we find an automorphism $\phi_{i\alpha\beta}$ of $(U_{\beta i} \cap U_{\alpha i})_{Z_{\beta j}\cap Z_{\beta j}}$ and its restriction to $(V_\alpha \cap V_\beta)_{Z_{V\beta j}\cap Z_{V\beta j}})_{p\leq j\leq m}$.

such that

(1) $(\phi_{i\alpha\beta})(H_{\alpha i}) = H_{\beta i}$.
(2) $\phi_{\alpha\beta i}$ is the identity for $\mathrm{supp}(\overline{\mathcal{J}}_i)$
(3) $\phi_{\alpha\beta i}$ preserves the marked ideal $\overline{J}_i$
(4) $\phi_{i\alpha\beta}(\overline{J}_{i|H_{\alpha i}}) = \overline{J}_{i|H_{\beta i}}$

Its restriction to $(V_\alpha \cap V_\beta)_{Z_{V\alpha}\cap Z_{V\beta}}$ defines an automoprhism for any compact $Z_{V\alpha} \subset Z \cap V_\alpha$ and $Z_{V\beta} \subset Z \cap V_\beta$. By the above $\widetilde{C}_j \cap (V_{\beta,j} \cap V_{\alpha,i})_{Z_{V\beta j}\cap Z_{V\beta j}}$ is the center of the canonical resolution of $\overline{J}_{i|H_{\beta i}}$ and of $\overline{J}_{i|H_{\alpha i}}$. Thus the restriction of the natural embedding $\widetilde{C}_j \cap (V_{\beta,j} \cap V_{\alpha,i})_{Z_{V\alpha j}\cap Z_{V\beta j}} \subset (\widetilde{C}_j \cap V_{\alpha,i})_{Z_{V\alpha i}}$ is an open embedding and $\widetilde{C}_j$ descends to a smooth center $C_j := \bigcup \widetilde{C}_j \cap V_{\alpha,j Z_{V\alpha j}} \subset \bigcup V_{\alpha j Z_{V\alpha j}} = M_{jZ_j}$.

Step 2. **Resolving marked ideals** $(M_Z, \mathcal{I}, E, \mu)$**.**

For any marked ideal $(M_Z, \mathcal{I}, E, \mu)$ write

$$I = \mathcal{M}(\mathcal{I})\mathcal{N}(\mathcal{I}),$$

where $\mathcal{M}(\mathcal{I})$ is the *monomial part* of $\mathcal{I}$, that is, the product of the principal ideals defining the irreducible components of the divisors in E, and $\mathcal{N}(\mathcal{I})$ is a *nonmonomial part* which is not divisible by any ideal of a divisor in E. Let

$$\mathrm{ord}_{\mathcal{N}(\mathcal{I})} := \max\{\mathrm{ord}_x(\mathcal{N}(\mathcal{I})) \mid x \in Z \cap \mathrm{supp}(\mathcal{I}, \mu)\}.$$

Definition 6.0.7. (Hironaka, Bierstone-Milman, Villamayor, Encinas-Hauser) By the *companion ideal* of $(\mathcal{I},\mu)$ where $I=\mathcal{N}(\mathcal{I})\mathcal{M}(\mathcal{I})$ we mean the marked ideal of maximal order

$$O(\mathcal{I},\mu)=\begin{cases}(\mathcal{N}(\mathcal{I}),\mathrm{ord}_{\mathcal{N}(\mathcal{I})})+(\mathcal{M}(\mathcal{I}),\mu-\mathrm{ord}_{\mathcal{N}(\mathcal{I})}) & \text{if } \mathrm{ord}_{\mathcal{N}(\mathcal{I})}<\mu,\\ (\mathcal{N}(\mathcal{I}),\mathrm{ord}_{\mathcal{N}(\mathcal{I})}) & \text{if } \mathrm{ord}_{\mathcal{N}(\mathcal{I})}\geq\mu.\end{cases}$$

Step 2a. **Reduction to the monomial case by using companion ideals.**

By Step 1 we can resolve the marked ideal of maximal order $(\mathcal{J},\mu_{\mathcal{J}}):=O(\mathcal{I},\mu)$. By Lemma 5.4.1, for any multiple test blow-up of $O(\mathcal{I},\mu)$,

$$\begin{aligned}\mathrm{supp}(O(\mathcal{I},\mu))_i&=\mathrm{supp}[\mathcal{N}(\mathcal{I}),\mathrm{ord}_{\mathcal{N}(\mathcal{I})}]_i\ \cap\ \mathrm{supp}[M(\mathcal{I}),\mu-\mathrm{ord}_{\mathcal{N}(H\mathcal{I})}]_i\\&=\mathrm{supp}[\mathcal{N}(\mathcal{I}),\mathrm{ord}_{\mathcal{N}(\mathcal{I})}]_i\ \cap\ \mathrm{supp}(\mathcal{I}_i,\mu).\end{aligned}$$

Consequently, such a resolution leads to the ideal $(\mathcal{I}_{r_1},\mu)$ such that $\mathrm{ord}_{\mathcal{N}(\mathcal{I}_{r_1})}<\mathrm{ord}_{\mathcal{N}(\mathcal{I})}$.Then we repeat the procedure for $(\mathcal{I}_{r_1},\mu)$. We find marked ideals $(\mathcal{I}_{r_0},\mu)=(\mathcal{I},\mu),(\mathcal{I}_{r_1},\mu),\ldots,(\mathcal{I}_{r_m},\mu)$ such that $\mathrm{ord}_{\mathcal{N}(\mathcal{I}_0)}>\mathrm{ord}_{\mathcal{N}(\mathcal{I}_{r_1})}>\ldots>\mathrm{ord}_{\mathcal{N}(\mathcal{I}_{r_m})}$. The procedure terminates after a finite number of steps when we arrive at the ideal $(\mathcal{I}_{r_m},\mu)$ with $\mathrm{ord}_{\mathcal{N}(\mathcal{I}_{r_m})}=0$ or with $\mathrm{supp}(\mathcal{I}_{r_m},\mu)=\emptyset$. In the second case we get the resolution. In the first case $\mathcal{I}_{r_m}=\mathcal{M}(\mathcal{I}_{r_m})$ is monomial.

Step 2b. **Monomial case** $\mathcal{I}=\mathcal{M}(\mathcal{I})$.

Let $\mathrm{Sub}(E_i)$ denote the set of all subsets of E_i. For any subset in $\mathrm{Sub}(E_i)$ write a sequence $(D_1,D_2,\ldots,0,\ldots)$ consisting of all elements of the subset in increasing order followed by an infinite sequence of zeros. We shall assume that $0\leq D$ for any $D\in E_i$. Consider the lexicographic order $\leq$ on the set of such sequences. Then for any two subsets $A_1=\{D_i^1\}_{i\in I}$ and $A_2=\{D_j^2\}_{j\in J}$ we write

$$A_1\leq A_2$$

if for the corresponding sequences $(D_1^1,D_2^1,\ldots,0,\ldots)\leq(D_1^2,D_2^2,\ldots,0,\ldots)$.

Let $x_1,\ldots,x_k$ define equations of the components $D_1^x,\ldots,D_k^x\in E$ through $x\in\mathrm{supp}(M_Z,\mathcal{I},E,\mu)$ and $\mathcal{I}$ be generated by the monomial $x^{a_1,\ldots,a_k}$ at x. Note that $\mathrm{ord}_x(\mathcal{I})=a_1+\ldots+a_k$.

Let $\rho(x)=\{D_{i_1},\ldots,D_{i_l}\}\in\mathrm{Sub}(E)$ be the maximal subset satisfying the properties

(1) $a_{i_1}+\ldots+a_{i_l}\geq\mu$.
(2) For any $j=1,\ldots,l$, $a_{i_1}+\ldots+\check{a}_{i_j}+\ldots+a_{i_l}<\mu$.

Let $R(x)$ denote the subsets in $\mathrm{Sub}(E)$ satisfying the properties (1) and (2). The maximal components of $\mathrm{supp}(\mathcal{I},\mu)$ through x are described by the intersections $\bigcap_{D\in A}D$ where $A\in R(x)$. The maximal locus of ρ determines at most one maximal component of $\mathrm{supp}(\mathcal{I},\mu)$ through each x.

After the blow-up at the maximal locus $C=\{x_{i_1}=\ldots=x_{i_l}=0\}$ of ρ, the ideal $\mathcal{I}=(x^{a_1,\ldots,a_k})$ is equal to $\mathcal{I}'=(x'^{a_1,\ldots,a_{i_j}-1,a,a_{i_j+1},\ldots,a_k})$ in the neighborhood corresponding to x_{i_j}, where $a=a_{i_1}+\ldots+a_{i_l}-\mu<a_{i_j}$. In particular the invariant ν drops for all points of some maximal components of $\mathrm{supp}(\mathcal{I},\mu)$. Thus the maximal value of ν on the maximal components of $\mathrm{supp}(\mathcal{I},\mu)$ which were blown up is bigger than the maximal value

of $\mathrm{ord}_x(\mathcal{I})$ on the new maximal components of $\mathrm{supp}(\mathcal{I},\mu)$. It follows that the algorithm terminates after a finite number of steps. □

Remarks. (1) (*) The ideal $\mathcal{J}$ is replaced with $\mathcal{H}(\mathcal{J})$ to ensure that the algorithm in Step 1b is independent of the choice of the tangent direction u. We replace $\mathcal{H}(\mathcal{J})$ with $\mathcal{C}(\mathcal{H}(\mathcal{J}))$ to ensure the equalities $\mathrm{supp}(\mathcal{J}_{|S}) = \mathrm{supp}(\mathcal{J}) \cap S$, where $S = H^s_\alpha$ in Step 1a and $S = V(u)$ in Step 1b.

(2) If $\mu = 1$ the companion ideal is equal to $O(\mathcal{I},1) = (\mathcal{N}(\mathcal{I}), \mu_{\mathcal{N}(\mathcal{I})})$ so the general strategy of the resolution of $\mathcal{I},\mu$ is to decrease the order of the nonmonomial part and then to resolve the monomial part.

(3) In particular if we desingularize Y we put $\mu = 1$ and $\mathcal{I} = \mathcal{I}_Y$ to be equal to the sheaf of the submanifold Y and we resolve the marked ideal $(M_Z, \mathcal{I}, \emptyset, \mu)$. The nonmonomial part $\mathcal{N}(\mathcal{I}_i)$ is nothing but the weak transform $(\sigma^i)^{\mathrm{w}}(\mathcal{I})$ of $\mathcal{I}$.

7. Conclusion of the resolution algorithm

7.1. Commutativity of resolving marked ideals $(M_Z, \mathcal{I}, \emptyset, 1)$ with embeddings of ambient manifolds

Let $(M_Z, \mathcal{I}, \emptyset, 1)$ be a marked ideal and $\phi : M_Z \hookrightarrow M'_Z$ be a closed embedding of germs of manifolds. Then ϕ defines the marked ideal $(M'_Z, \mathcal{I}', \emptyset, 1)$, where $\mathcal{I}' = \phi_*(\mathcal{I}) \cdot \mathcal{O}_{M'_Z}$ (see remark after Theorem 2.0.1). We may assume that M_Z is a germ of the submanifold M of M' which is locally generated by coordinates $u_1, \ldots, u_k$. Then $u_1, \ldots, u_k$ in $\mathcal{I}'(U') = T(\mathcal{I})(U')$ define tangent directions on some open $U' \subset M'_Z$. We run Steps 2a and 1bb of our algorithm. That is, we pass to the hypersurface $V(u_1)$ and replace $\mathcal{I}$ with its restriction. By Step 1bb resolving $(M'_Z, \mathcal{I}', \emptyset, \mu)$ is locally equivalent to resolving $(V(u_1)_Z, \mathcal{I}'_{|V(u_1)}, \emptyset, \mu)$.

By repeating the procedure k times and restricting to the tangent directions $u_1, \ldots, u_k$ of the marked ideal $\mathcal{I}$ on M_Z we obtain:

Resolving $(M'_Z, \mathcal{I}', \emptyset, \mu)$ is equivalent to resolving $(M_Z, \mathcal{I}, \emptyset, \mu)$.

7.2. Principalization

Resolving the marked ideal $(M_Z, \mathcal{I}, \emptyset, 1)$ determines a principalization commuting with local analytic isomorphisms and embeddings of the ambient manifolds.

The principalization is often reached at an earlier stage upon transformation to the monomial case (Step 2b) (However the latter procedure does not commute with embeddings of ambient manifolds)

7.3. Weak embedded desingularization

Let Y be a closed subspace of the germ M_Z. Consider the marked ideal $(M_Z, \mathcal{I}_Y, \emptyset, 1)$. Its support $\mathrm{supp}(\mathcal{I}_Y, 1)$ is equal to Y. In the resolution process of $(M_Z, \mathcal{I}_Y, \emptyset, 1)$, the strict transform of Y is blown up. Otherwise the generic points of Y would be transformed isomorphically, which contradicts the resolution of $(M_Z, \mathcal{I}_Y, \emptyset, 1)$.

7.4. Bravo-Villamayor strengthening of the weak embedded desingularization

Theorem 7.4.1. *(Bravo-Villamayor* [13], [11]*) Let Y be a closed subspace of a manifold M and $Y = \bigcup Y_i$ be its decomposition into the union of irreducible components. There is a canonical locally finite resolution of a subspace $Y \subset M$, subject to the conditions from Theorem 2.0.2 such that the strict transforms $\widetilde{Y}_i$ of Y_i are smooth and disjoint. Moreover the full transform of Y is of the form*

$$(\widetilde{\sigma})^*(\mathcal{I}_Y) = \mathcal{M}((\widetilde{\sigma})^*(\mathcal{I}_Y)) \cdot \mathcal{I}_{\widetilde{Y}},$$

where $\widetilde{Y} := \bigcup \widetilde{Y}_i \subset \widetilde{M_Z}$ is a disjoint union of the strict transforms $\widetilde{Y}_i$ of Y_i, $\mathcal{I}_{\widetilde{Y}}$ is the sheaf of ideals of $\widetilde{Y}$ and $\mathcal{M}((\widetilde{\sigma})^*(\mathcal{I}_Y))$ is the monomial part of $(\widetilde{\sigma})^*(\mathcal{I}_Y)$.

Proof. Let $\mathcal{I} := \mathcal{I}_Y$ be the ideal sheaf of Y. Fix any compact set $Z \subset M$.

We use the following:

Modified algorithm in Step 2. We run the algorithm of resolving $(M_Z, \mathcal{I}, \emptyset, 1)$ as before until we drop the $\max\{\mathrm{ord}_x(\mathcal{N}(\mathcal{I})) : x \in \mathrm{supp}(\mathcal{I})\}$ to 1. The control transform of $(\mathcal{I}, 1)$ becomes equal to $(\mathcal{M}(\mathcal{I})\mathcal{N}(\mathcal{I}), 1)$. At this point algorithm is altered. We resolve the monomial ideal $(\mathcal{M}(\mathcal{I}), 1)$. The blow-ups are performed at exceptional divisors for which $\rho(x)$ is maximal. We arrive at the purely nonmonomial case $\mathcal{I}' = \mathcal{N}(\mathcal{I}')$, where $\max\{\mathrm{ord}_x(\mathcal{I}) : x \in \mathrm{supp}(\mathcal{I}') \cap Z\} = 1$. This concludes the altered procedure in Step 2. At this point we perform the altered Step 1 described below.

Modified algorithm in Step 1. In Step 1a we move the "old divisors" E as before. In Step 1b we consider two possibilities. If $\mathcal{I}' = (u)$ is the ideal of smooth hypersurface of (maximal contact) as in Step 1ba) the algorithm is stopped. Otherwise we restrict $(\mathcal{I}', 1) = \mathcal{C}(\mathcal{H}(\mathcal{I}))$ to a hypersurface of maximal contact $V(u_1)$.

The modified algorithm in Step 2 and Step 1 is then repeated for the restriction $\mathcal{I}'_{|}V(u_1)$.

We continue this procedure until it terminates. Then the resulting controlled transform of $(\mathcal{I}, 1)$ is locally equal to $\mathcal{I}'' = (u_1, \ldots, u_k)$, where u_i are coordinates transversal to exceptional divisors. The sheaf $\mathcal{I}''$ desribes the germ of submanifold which is a union of disjoint irreducible components. Some of them are the strict transforms of Y_i. Other components are possible strict transforms of embedding components occuring the process. At the end we blow-up all the irreducible components which are not strict transforms of Y_i. The procedure is canonical. It is defined for germs of analytic subspace at compact sets and it glues to the algorthm for whole subspace of manifolds.

References

[1] S. S. Abhyankar, *Desingularization of plane curves,* In Algebraic Geometry, Arcata 1981, Proc. Symp. Pure Appl. Math. 40, Amer. Math. Soc., 1983.

[2] S. S. Abhyankar, *Good points of a hypersurface,* Adv. in Math. 68 (1988), 87-256.

[3] D. Abramovich and A. J. de Jong, *Smoothness, semistability, and toroidal geometry,* J. Alg. Geom. 6 (1997), 789–801.

[4] D. Abramovich and J. Wang, *Equivariant resolution of singularities in characteristic 0,* Math. Res. Letters 4 (1997), 427–433.

[5] J. M. Aroca, H. Hironaka, and J. L. Vicente, *Theory of maximal contact,* Memo Math. del Inst. Jorge Juan, 29, 1975.

[6] J. M. Aroca, H. Hironaka, and J. L. Vicente, *Desingularization theorems,* Memo Math. del Inst. Jorge Juan, 30, 1977.

[7] E. Bierstone and P. Milman, *Semianalytic and subanalytic sets,* Publ. Math. IHES 67 (1988), 5–42.

[8] E. Bierstone and P. Milman, *Uniformization of analytic spaces,* J. Amer. Math. Soc. 2 (1989), 801–836.

[9] E. Bierstone and P. Milman, *A simple constructive proof of canonical resolution of singularities,* In T. Mora and C. Traverso, eds., Effective methods in algebraic geometry, pages 11–30, Birkhäuser, 1991.

[10] E. Bierstone and D. Milman, *Canonical desingularization in characteristic zero by blowing up the maximum strata of a local invariant,* Invent. Math. 128 (1997), 207–302.

[11] E. Bierstone and D. Milman, *Desingularization algorithms, I. Role of exceptional divisors,* IHES/M/03/30.

[12] E. Bierstone and D. Milman, *Desingularization of toric and binomial varieties,* preprint math.AG/0411340

[13] A. Bravo and O. Villamayor, *A strengthening of resolution of singularities in characteristic zero,* Proc. London Math. Soc. 2003 86(2), 327–357

[14] F. Bogomolov and T. Pantev, *Weak Hironaka theorem,* Math. Res. Letters 3 (1996), 299–309.

[15] V. Cossart, *Desingularization of embedded excellent surfaces,* Tohoku Math. J. 33 (1981), 25–33.

[16] S.D. Cutkosky, *Resolution of singularities,* AMS Graduate Studies in Mathematics, volume 63

[17] S. Encinas and H. Hauser, *Strong resolution of singularities in characteristic zero,* Comment. Math. Helv. 77 (2002), 821–845.

[18] S. Encinas and O. Villamayor, *Good points and constructive resolution of singularities,* Acta Math. 181 (1998), 109–158.

[19] S. Encinas and O. Villamayor, *A course on constructive desingularization and equivariance,* In H. Hauser et al., eds., Resolution of Singularities, A research textbook in tribute to Oscar Zariski, volume 181 of Progress in Mathematics, Birkhäuser, 2000.

[20] S. Encinas and O. Villamayor, *A new theorem of desingularization over fields of characteristic zero,* Preprint, 2001.

[21] J. Giraud, *Sur la théorie du contact maximal,* Math. Zeit. 137 (1974), 285–310.

[22] R. Hartshorne, *Algebraic Geometry,* Graduate Texts in Mathematics 52, Springer-Verlag, 1977.

[23] H. Hauser, *Resolution of singularities 1860-1999,* Resolution of singularities (Obergurgl, 1997), Progr. Math., 181, Birkhäuser, pp. 5–36

[24] H. Hauser, *The Hironaka theorem on resolution of singularities (or: A proof we always wanted to understand),* Bull Amer. Math. Soc. 40 (2003), 323–403.

[25] H. Hironaka, *An example of a non-Kählerian complex-analytic deformation of Kählerian complex structure,* Annals of Math. (2), 75 (1962), 190–208.

[26] H. Hironaka, *Resolution of singularities of an algebraic variety over a field of characteristic zero,* Annals of Math. 79 (1964), 109–326.

[27] H. Hironaka, *Introduction to the theory of infinitely near singular points,* Memo Math. del Inst. Jorge Juan, 28, 1974.

[28] H. Hironaka, *Idealistic exponents of singularity,* In Algebraic Geometry, The Johns Hopkins centennial lectures, pages 52–125, Johns Hopkins University Press, Baltimore, 1977.
[29] R. Goldin and B. Teissier, *Resolving singularities of plane analytic branches with one toric morphism,* Preprint ENS Paris, 1995.
[30] A. J. de Jong. *Smoothness, semistability, and alterations,* Publ. Math. I.H.E.S. 83 (1996), 51–93.
[31] J. Kollár, *Lectures on Resolution of singularities*, Princeton University Press, 2007.
[32] J. Lipman, *Introduction to the resolution of singularities,* In Arcata 1974, volume 29 of Proc. Symp. Pure Math, pages 187–229, 1975.
[33] K. Matsuki, *Notes on the inductive algorithm of resolution of singularities,* Preprint.
[34] T. Oda, *Infinitely very near singular points,* Adv. Studies Pure Math. 8 (1986), 363–404.
[35] O. Villamayor, *Constructiveness of Hironaka's resolution,* Ann. Scient. Ecole Norm. Sup. 22 (1989), 1–32.
[36] O. Villamayor, *Patching local uniformizations,* Ann. Scient. Ecole Norm. Sup. 25 (1992), 629–677.
[37] O. Villamayor, *Introduction to the algorithm of resolution,* In Algebraic geometry and singularities, La Rabida 1991, pages 123–154, Birkhäuser, 1996.
[38] J. Włodarczyk, *Simple Hironaka resolution in characteristic zero*, J. Amer. Math. Soc. 18 (2005), 779-822

Department of Mathematics, Purdue University, West Lafayette, IN 47907, USA
E-mail address: wlodar@math.purdue.edu, jwlodar@mimuw.edu.pl

Proceedings of 15th Gökova
Geometry-Topology Conference
pp. 64–90

Floor decompositions of tropical curves: the planar case

Erwan Brugallé and Grigory Mikhalkin

Abstract. In [BM07] we announced a formula to compute Gromov-Witten and Welschinger invariants of some toric varieties, in terms of combinatorial objects called floor diagrams. We give here detailed proofs in the tropical geometry framework, in the case when the ambient variety is a complex surface, and give some examples of computations using floor diagrams. The focusing on dimension 2 is motivated by the special combinatoric of floor diagrams compared to arbitrary dimension.

We treat a general toric surface case in this dimension: the curve is given by an arbitrary lattice polygon and include computation of Welschinger invariants with pairs of conjugate points. See also [FM] for combinatorial treatment of floor diagrams in the projective case.

1. Introduction

Let Δ be a lattice polygon in $\mathbb{R}^2$, g a non-negative integer, and ω a generic configuration of $Card(\partial\Delta \cap \mathbb{Z}^2) - 1 + g$ points in $(\mathbb{C}^*)^2$. Then, there exists a finite number $N(\Delta, g)$ of complex algebraic curves in $(\mathbb{C}^*)^2$ of genus g and Newton polygon Δ passing through all points in ω. Moreover, $N(\Delta, g)$ doesn't depend on ω as long as it is generic. If the toric surface $Tor(\Delta)$ corresponding to Δ is Fano, then the numbers $N(\Delta, g)$ are known as *Gromov-Witten invariants* of the surface $Tor(\Delta)$. Kontsevich first computed in [KM94] the series $N(\Delta, 0)$ for convex surfaces $Tor(\Delta)$, and Caporaso and Harris computed in [CH98] all $N(\Delta, g)$'s where $Tor(\Delta)$ is Fano or a Hirzebruch surface.

Suppose now that the surface $Tor(\Delta)$ is equipped with a real structure $conj$, i.e. $conj$ is an antiholomorphic involution on $Tor(\Delta)$. For example, one can take the tautological real structure given in $(\mathbb{C}^*)^2$ by the standard complex conjugation. Suppose moreover that ω is a real configuration, i.e. $conj(\omega) = \omega$. Then it is natural to study the set $\mathbb{R}\mathcal{C}(\omega)$ of real algebraic curves in $(\mathbb{C}^*)^2$ of genus g and Newton polygon Δ passing through all points in ω. It is not hard to see that, unlike in the enumeration of complex curves, the cardinal of this set depends heavily on ω. However, Welschinger proved in [Wel05] that when $g = 0$ and $Tor(\Delta)$ is Fano, one can define an invariant. A real nodal curve C in $Tor(\Delta)$ has two types of real nodes, isolated ones (locally given by the equation $x^2 + y^2 = 0$) and

2000 *Mathematics Subject Classification.* Primary 14N10, 14P05; Secondary 14N35, 14N05.
Key words and phrases. tropical geometry, enumerative geometry, Welschinger invariants, Gromov-Witten invariants.

non-isolated ones (locally given by the equation $x^2 - y^2 = 0$). Welschinger defined the *mass* $m(C)$ of the curve C as the number of isolated nodes of C, and proved that if $g = 0$ and $Tor(\Delta)$ is Fano, then the number

$$W(\Delta, r) = \sum_{C \in \mathbb{RC}(\omega)} (-1)^{m(C)}$$

depends only on Δ and the number r of pairs of complex conjugated points in ω.

Tropical geometry is an algebraic geometry over the tropical semi-field $\mathbb{T} = \mathbb{R} \cup \{-\infty\}$ where the tropical addition is taking the maximum, and the tropical multiplication is the classical addition. As in the classical setting, given Δ a lattice polygon in $\mathbb{R}^2$, g a non-negative integer, and ω a generic configuration of $Card(\partial\Delta \cap \mathbb{Z}^2) - 1 + g$ points in $(\mathbb{R}^*)^2$, we can enumerate tropical curves in $\mathbb{R}^2$ of genus g and Newton polygon Δ passing through all points in ω. It was proved in [Mik05] that provided that we count tropical curves with an appropriate multiplicity, then the number of tropical curves does not depend on ω and is equal to $N(\Delta, g)$. Moreover, tropical geometry allows also one to compute quite easily Welschinger invariants $W(\Delta, r)$ of Fano toric surfaces equipped with the tautological real structure (see [Mik05], [Shu06]). This has been the first systematic method to compute Welschinger invariants of these surfaces.

In [BM07], we announced a formula to compute the numbers $N(\Delta, g)$ and $W(\Delta, r)$ easily in terms of combinatorial objects called *floor diagrams*. This diagrams encode degeneracies of tropical curves passing through some special configuration of points. This paper is devoted to explain how floor diagrams can be used to compute the numbers $N(\Delta, g)$ and $W(\Delta, r)$ (Theorems 3.6 and 3.9), and to give some examples of concrete computations (section 6). In [BM07], we announced a more general formula computing Gromov-Witten and Welschinger invariants of some toric varieties of any dimension. However, floor diagrams corresponding to plane curves have a special combinatoric compared with the general case, and deserve some special attention. Details of the proof of the general formula given in [BM07] will appear soon.

In section 2 we remind some convention we use throughout this paper about graphs and lattice polygons. Then, we state in section 3 our main formulas computing the numbers $N(\Delta, g)$ and $W(\Delta, r)$ when Δ is a *h-transverse* polygon. We present tropical enumerative geometry in section 4, and prove our main formulas in section 5. We give some examples of computations using floor diagrams in section 6, and we end this paper with some remarks in section 7.

2. convention

2.1. Graphs

In this paper, graphs are considered as (non necessarily compact) abstract 1 dimensional topological objects. Recall that a *leaf* of a graph is an edge which is non-compact or adjacent to a 1-valent vertex. Given a graph Γ, we denote by

- Vert(Γ) the set of its vertices,
- End(Γ) the set of its 1-valent vertices,
- Edge(Γ) the set of its edges,
- Edge$^\infty(\Gamma)$ the set of its non-compact leaves.

If in addition Γ is oriented so that there are no oriented cycles, then there exists a natural partial ordering on Γ: an element a of Γ is greater than another element b if there exists an oriented path from b to a. In this case, we denote by Edge$^{+\infty}(\Gamma)$ (resp. Edge$^{-\infty}(\Gamma)$) the set of edges e in Edge$^\infty(\Gamma)$ such that no vertex of Γ is greater (resp. smaller) than a point of e.

We say that Γ is a *weighted graph* if each edge of Γ is prescribed a natural weight, i.e. we are given a function $\omega : \text{Edge}(\Gamma) \to \mathbb{N}^*$. Weight and orientation allow one to define the *divergence* at the vertices. Namely, for a vertex $v \in \text{Vert}(\Gamma)$ we define the divergence $\text{div}(v)$ to be the sum of the weights of all incoming edges minus the sum of the weights of all outgoing edges.

2.2. Lattice polygons

We remind that a *primitive integer vector*, or shortly a primitive vector, is a vector (α, β) in $\mathbb{Z}^2$ whose coordinates are relatively prime. A *lattice polygon* Δ is a convex polygon in $\mathbb{R}^2$ whose vertices are in $\mathbb{Z}^2$. For such a polygon, we define

$$\partial_l\Delta = \{p \in \partial\Delta \mid \forall t > 0,\ p + (-t, 0) \notin \Delta\},$$
$$\partial_r\Delta = \{p \in \partial\Delta \mid \forall t > 0,\ p + (t, 0) \notin \Delta\}.$$

A lattice polygon Δ is said to be *h-transverse* if any primitive vector parallel to an edge of $\partial_l\Delta$ or $\partial_r\Delta$ is of the form $(\alpha, \pm 1)$ with α in $\mathbb{Z}$.

If e is a lattice segment in $\mathbb{R}^2$, we define the *integer length* of e by $l(e) = Card(e \cap \mathbb{Z}^2) - 1$. If Δ is a h-transverse polygon, we define its *left directions* (resp. *right directions*), denoted by $d_l(\Delta)$ (resp. $d_r(\Delta)$), as the unordered list that consists of the elements α repeated $l(e)$ times for all edge vectors $e = \pm l(e)(\alpha, -1)$ of $\partial_l\Delta$ (resp. $\partial_r\Delta$). If Δ has a bottom (resp. top) horizontal edge e then we set $d_-(\Delta) = l(e)$ (resp. $d_+(\Delta) = l(e)$) and $d_-(\Delta) = 0$ (resp. $d_+(\Delta) = 0$) otherwise.

There is a natural one-to-one correspondence between quadruples (d_l, d_r, d_-, d_+) and h-transverse polygons Δ considered up to translation as the polygon can be easily reconstructed from such a quadruple.

We have

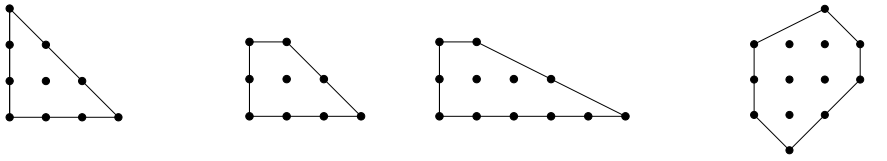
$$Card(d_l(\Delta)) = Card(d_r(\Delta)) = Card(\partial_l\Delta \cap \mathbb{Z}^2) - 1 = Card(\partial_r\Delta \cap \mathbb{Z}^2) - 1 \qquad (1)$$

and

$$2Card(d_l(\Delta)) + d_-(\Delta) + d_+(\Delta) = Card(\partial\Delta \cap \mathbb{Z}^2). \qquad (2)$$

We call the cardinality $Card(d_l(\Delta))$ the *height* of h-transversal polygon Δ.

Example 2.1. Some h-transverse polygons are depicted in Figure 1. By abuse of notation, we write unordered lists within brackets $\{\}$.



a)	$d_l = \{0,0,0\}$ $d_r = \{1,1,1\}$ $d_- = 3$ $d_+ = 0$	*b)*	$d_l = \{0,0\}$ $d_r = \{1,1\}$ $d_- = 3$ $d_+ = 1$	*c)*	$d_l = \{0,0\}$ $d_r = \{2,2\}$ $d_- = 5$ $d_+ = 1$	*d)*	$d_l = \{1,0,0,-2\}$ $d_r = \{-1,-1,0,1\}$ $d_- = 0$ $d_+ = 0$

FIGURE 1. Examples of h-transverse polygons

Remark 2.2. If Δ is a lattice polygon and if v is a primitive integer vector such that for any edge e of Δ we have $|det(v,e)| \leq l(e)$, then Δ is a h-transverse polygon after a suitable change of coordinates in $SL_2(\mathbb{Z})$.

In this paper, we denote by Δ_d the lattice polygon with vertices $(0,0)$, $(d,0)$, and $(0,d)$.

3. Floor diagrams

Here we define the combinatorial objects that can be used to replace the algebraic curves in real and complex enumerative problems. In this section, we fix an h-transverse lattice polygon Δ.

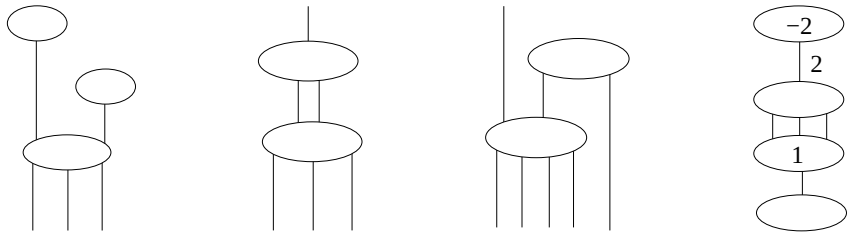
3.1. Enumeration of complex curves

Definition 3.1. *A (plane) floor diagram $\mathcal{D}$ of genus g and Newton polygon Δ is the data of a connected weighted oriented graph Γ and a map $\theta : Vert(\Gamma) \to \mathbb{Z}$ which satisfy the following conditions*

- *the oriented graph Γ is acyclic,*
- *the first Betti number $b_1(\Gamma)$ equals g,*
- *there are exactly $d_\pm(\Delta)$ edges in $Edge^{\pm\infty}(\Gamma)$, and all of them are of weight 1,*
- *the (unordered) collection of numbers $\theta(v)$, where v goes through vertices of Γ, coincides with $d_l(\Delta)$,*
- *the (unordered) collection of numbers $\theta(v) + div(v)$, where v goes through vertices of Γ, coincides with $d_r(\Delta)$.*

In order to avoid too many notation, we will denote by the same letter $\mathcal{D}$ a floor diagram and its underlying graph Γ. Here are the convention we use to depict floor diagrams: vertices of $\mathcal{D}$ are represented by ellipses. We write $\theta(v)$ inside the ellipse v only if $\theta(v) \neq 0$. Edges of $\mathcal{D}$ are represented by vertical lines, and the orientation is implicitly from down to up. We write the weight of an edge close to it only if this weight is at least 2. In the following, we define $s = Card(\partial\Delta \cap \mathbb{Z}^2) + g - 1$.

Example 3.2. Figure 2 depicts an example of floor diagram for any h-transverse polygon depicted in Figure 1.



a) $g = 0$ b) $g = 1$ c) $g = 0$ d) $g = 2$

FIGURE 2. Examples of floor diagrams whose Newton polygon are depicted in Figure 1

Note that Equations (1) and (2) combined with Euler's formula imply that for any floor diagram $\mathcal{D}$ of genus g and Newton polygon Δ we have

$$Card(\mathrm{Vert}(\mathcal{D})) + Card(\mathrm{Edge}(\mathcal{D})) = s.$$

A map m between two partially ordered sets is said *increasing* if

$$m(i) > m(j) \Longrightarrow i > j$$

Definition 3.3. *A marking of a floor diagram $\mathcal{D}$ of genus g and Newton polygon Δ is an increasing map $m : \{1, \dots, s\} \to \mathcal{D}$ such that for any edge or vertex x of $\mathcal{D}$, the set $m^{-1}(x)$ consists of exactly one element.*

A floor diagram enhanced with a marking is called a *marked floor diagram* and is said to be marked by m.

Definition 3.4. *Two marked floor diagrams $(\mathcal{D}, m)$ and $(\mathcal{D}', m')$ are called equivalent if there exists a homeomorphism of oriented graphs $\phi : \mathcal{D} \to \mathcal{D}'$ such that $w = w' \circ \phi$, $\theta = \theta' \circ \phi$, and $m = m' \circ \phi$.*

Hence, if $m(i)$ is an edge e of $\mathcal{D}$, only the knowledge of e is important to determine the equivalence class of $(\mathcal{D}, m)$, not the position of $m(i)$ on e. From now on, we consider marked floor diagrams up to equivalence. To any (equivalence class of) marked floor diagram, we assign a sequence of non-negative integers called *multiplicities* : a *complex* multiplicity, and some *r-real* multiplicities.

Definition 3.5. *The complex multiplicity of a marked floor diagram $\mathcal{D}$, denoted by $\mu^{\mathbb{C}}(\mathcal{D})$, is defined as*

$$\mu^{\mathbb{C}}(\mathcal{D}) = \prod_{e \in Edge(\mathcal{D})} w(e)^2$$

Note that the complex multiplicity of a marked floor diagram depends only on the underlying floor diagram. Next theorem is the first of our two main formulas.

Theorem 3.6. *For any h-transverse polygon Δ and any genus g, one has*

$$N(\Delta, g) = \sum \mu^{\mathbb{C}}(\mathcal{D})$$

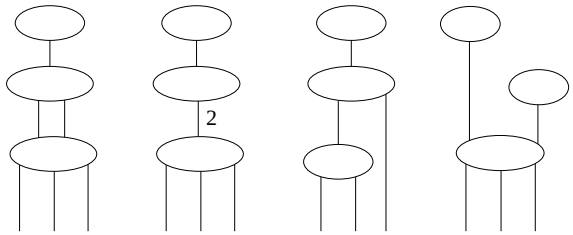
where the sum is taken over all marked floor diagrams of genus g and Newton polygon Δ.

Theorem 3.6 is a corollary of Proposition 5.9 proved in section 5.

Example 3.7. Using marked floor diagrams depicted in Figures 3 and 4 we verify that

$N(\Delta_3, 1) = 1$ (see Figure 3a), $N(\Delta_3, 0) = 12$ (see Figure 3b,c,d).

$N(\Delta, 0) = 84$ (see Figure 4), where Δ is the polygon depicted in Figure 1c.



a) $\mu^{\mathbb{C}} = 1$, 1 marking b) $\mu^{\mathbb{C}} = 4$, 1 marking c) $\mu^{\mathbb{C}} = 1$, 5 markings d) $\mu^{\mathbb{C}} = 1$, 3 markings

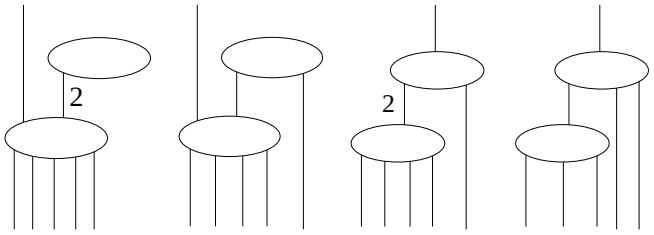
FIGURE 3. Floor diagrams of genus 1 and 0, and Newton polygon Δ_3

3.2. Enumeration of real curves

First of all, we have to define the notion of real marked floor diagrams. Like before, we define $s = Card(\partial\Delta \cap \mathbb{Z}^2) + g - 1$. Choose an integer $r \geq 0$ such that $s - 2r \geq 0$, and $\mathcal{D}$ a floor diagram of genus 0 and Newton polygon Δ marked by a map m.

The set $\{i, i+1\}$ is a called *r-pair* if $i = s - 2k + 1$ with $1 \leq k \leq r$. Denote by $\Im(m, r)$ the union of all the r-pairs $\{i, i+1\}$ where $m(i)$ is not adjacent to $m(i+1)$. Let $\rho_{m,r} : \{1, \ldots, s\} \to \{1, \ldots, s\}$ be the bijection defined by $\rho_{m,r}(i) = i$ if $i \notin \Im(m, r)$, and by $\rho_{m,r}(i) = j$ if $\{i, j\}$ is a r-pair contained in $\Im(m, r)$. Note that $\rho_{m,r}$ is an involution.

We define o_r to be the half of the number of vertices v of $\mathcal{D}$ in $m(\Im(m, r))$ with odd divergence $\mathrm{div}(v)$, and we set $A = \mathrm{Edge}(\mathcal{D}) \setminus m(\{1, \ldots, s - 2r\})$.



a) $\mu^{\mathbb{C}} = 4$, 3 markings | b) $\mu^{\mathbb{C}} = 1$, 23 markings | c) $\mu^{\mathbb{C}} = 4$, 7 markings | d) $\mu^{\mathbb{C}} = 1$, 21 markings

FIGURE 4. Floor diagrams of genus 0 and Newton polygon depicted in Figure 1c

Definition 3.8. *A marked floor diagram $(\mathcal{D}, m)$ is called r-real if the two marked floor diagrams $(\mathcal{D}, m)$ and $(\mathcal{D}, m \circ \rho_{m,r})$ are equivalent.*

The r-real multiplicity of a r-real marked floor diagram, denoted by $\mu_r^{\mathbb{R}}(\mathcal{D}, m)$, is defined as

$$\mu_r^{\mathbb{R}}(\mathcal{D}, m) = (-1)^{o_r} \prod_{e \in A} w(e)$$

if all edges of $\mathcal{D}$ of even weight contains a point of $m(\Im(m, r))$, and as

$$\mu_r^{\mathbb{R}}(\mathcal{D}, m) = 0$$

otherwise.

For convenience we set $\mu_r^{\mathbb{R}}(\mathcal{D}, m) = 0$ also in the case when $(\mathcal{D}, m)$ is not r-real.

Note that $\mu_0^{\mathbb{R}}(\mathcal{D}, m) = 1$ or 0 and is equal to $\mu^{\mathbb{C}}(\mathcal{D})$ modulo 2, hence doesn't depend on m. However, $\mu_r^{\mathbb{R}}(\mathcal{D}, m)$ depends on m as soon as $r \geq 1$. Next theorem is the second main formula of this paper.

Theorem 3.9. *Let Δ be a h-transverse polygon such that Welschinger invariants are defined for the corresponding toric surface $Tor(\Delta)$ equipped with its tautological real structure. Then for any integer r such that $s - 2r \geq 0$, one has*

$$W(\Delta, r) = \sum \mu_r^{\mathbb{R}}(\mathcal{D}, m)$$

where the sum is taken over all marked floor diagrams of genus 0 and Newton polygon Δ.

Theorem 3.9 is a corollary of Proposition 5.9 proved in section 5.

Example 3.10. All marked floor diagrams of genus 0 and Newton polygon Δ_3 are depicted in Table 1 together with their real multiplicities. The first floor diagram has an edge of weight 2, but we didn't mention it in the picture to avoid confusion. According to Theorem 3.9 we find $W(\Delta_3, r) = 8 - 2r$.

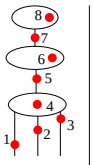
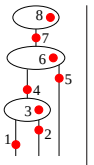
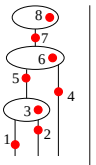
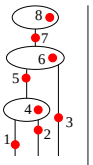
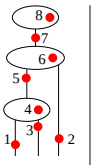
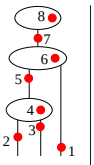
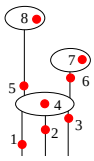
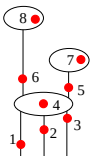
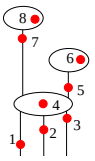
									
$\mu^{\mathbb{C}}$	4	1	1	1	1	1	1	1	1
$\mu_0^{\mathbb{R}}$	0	1	1	1	1	1	1	1	1
$\mu_1^{\mathbb{R}}$	0	1	1	1	1	1	0	0	1
$\mu_2^{\mathbb{R}}$	0	1	1	1	1	1	-1	-1	1
$\mu_3^{\mathbb{R}}$	0	1	0	0	1	1	-1	-1	1
$\mu_4^{\mathbb{R}}$	0	1	0	0	0	0	-1	-1	1

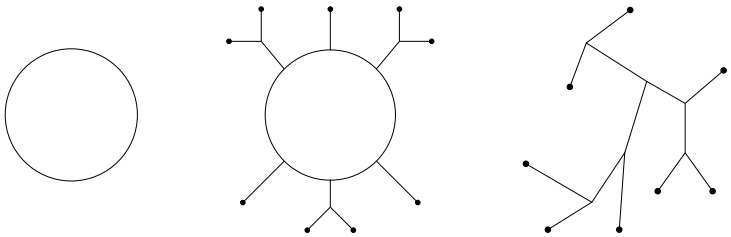
TABLE 1. Computation of $W(\Delta_3, r)$

4. Enumerative tropical geometry

4.1. Tropical curves

Definition 4.1. *An irreducible tropical curve C is a connected compact metric graph whose leaves are exactly the edges of infinite length. This means that $C \setminus End(C)$ is a complete metric space with inner metric. In other words the 1-valent vertices are at the infinite distance from all the other points of C. The genus of C is defined as its first Betti number $b_1(C)$.*

Example 4.2. Examples of tropical curves are depicted in Figure 5. 1-valent vertices are represented with bullets.



a) $g = 1$ b) $g = 1$ c) $g = 0$

FIGURE 5. Examples of tropical curves

Given e an edge of a tropical curve C, we choose a point p in the interior of e and a unit vector u_e of the tangent line to C at p. Of course, the vector u_e depends on the choice of p

and is not well defined, but this will not matter in the following. We will sometimes need u_e to have a prescribed direction, and we will then precise this direction. The standard inclusion of $\mathbb{Z}^2$ in $\mathbb{R}^2$ induces a standard inclusion of $\mathbb{Z}^2$ in the tangent space of $\mathbb{R}^2$ at any point of $\mathbb{R}^2$.

Definition 4.3. *A map $f : C \setminus End(C) \to \mathbb{R}^2$ is called a tropical morphism if the following conditions are satisfied*

- *for any edge e of C, the restriction $f_{|e}$ is a smooth map with $df(u_e) = w_{f,e}u_{f,e}$ where $w_{f,e}$ is a non-negative integer and $u_{f,e} \in \mathbb{Z}^2$ is a primitive vector,*
- *for any vertex v of C whose adjacent edges are $e_1, \dots, e_k$, one has the balancing condition*
$$\sum_{i=1}^{k} w_{f,e_i} u_{f,e_i} = 0$$
where u_{e_i} is chosen so that it points away from v.

Let $f : C \setminus \mathrm{End}(C) \to \mathbb{R}^2$ be a tropical morphism, and define $L(C, f)$ as the unordered list composed by elements $u_{f,e}$ repeated $w_{f,e}$ times where e goes through leaves of C and u_e is chosen so that it points to the 1-valent vertex. Then, there exists a unique, up to translation by a vector in $\mathbb{Z}^2$, lattice polygon $\Delta(C, f)$ such that the unordered list composed by the primitive vector normal to e and outward to $\Delta(C, f)$ repeated $l(e)$ times where e goes through edges of $\Delta(C, f)$ equals the list $L(C, f)$.

Definition 4.4. *The polygon $\Delta(C, f)$ is called the Newton polygon of the pair (C, f).*

Not any tropical curve admits a non-constant tropical morphism to $\mathbb{R}^2$. The tropical curve depicted in Figure 5a does not admit any tropical morphism since a circle cannot be mapped to a segment in $\mathbb{R}^2$ by a dilatation. However, up to *modification*, every tropical curve can be tropically immersed to $\mathbb{R}^2$ (see [Mik]).

The pair (C, f) where $f : C \setminus \mathrm{End}(C) \to \mathbb{R}^2$ is a tropical morphism with Newton polygon Δ is called a *parameterized tropical curve with Newton polygon* Δ. The integer $w_{f,e}$ is called the *weight* of the edge e. The genus of (C, f) is naturally defined as the genus of C.

Example 4.5. If C is the tropical curve depicted in Figure 5b (resp. c) then an example of the image $f(C)$ for some parameterization with Newton polygon Δ_3 is depicted in Figure 6a (resp. b). The second tropical morphism has an edge of weight 2.

Definition 4.6. *A tropical curve with n marked points is a $(n+1)$-tuple $(C, x_1, \dots, x_n)$ where C is a tropical curve and the x_i's are n points on C.*

A parameterized tropical curve with n marked points is a $(n+2)$-tuple $(C, x_1, \dots, x_n, f)$ where $(C, x_1, \dots, x_n)$ is a tropical curve with n marked points, and (C, f) is a parameterized tropical curve.

Note that in this paper we do not require the marked points on a marked tropical curve to be distinct. In the following, we consider tropical curves (with n marked points)

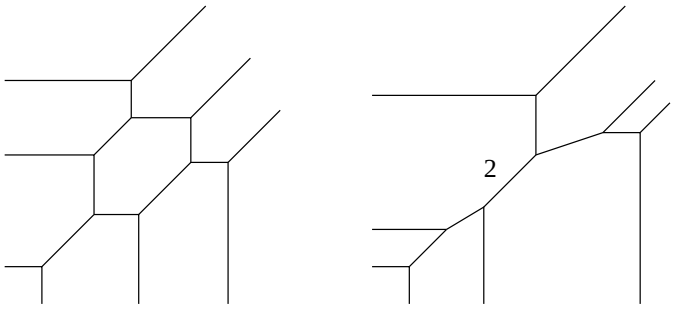


FIGURE 6. Images of tropical morphisms with Newton polygon Δ_3

up to homeomorphism of metric graphs (which send the i^{th} point to the i^{th} point). The notions of vertices, edges, Newton polygon, ... also make sense for a parameterized marked tropical curve as the corresponding notions for the underlying (parameterized) tropical curve.

4.2. Complex multiplicity of a tropical curve

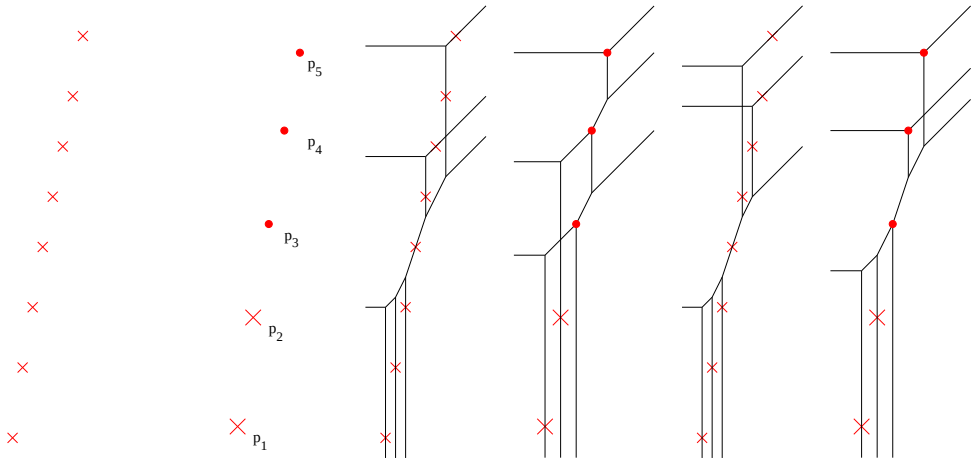
Let us now turn to tropical enumerative geometry, and let's relate it first to complex enumerative geometry. More details about this section can be found in [Mik05] or [GM07].

Fix a lattice polygon Δ, a non-negative integer number g, and define the number $s = Card(\partial\Delta \cap \mathbb{Z}^2) - 1 + g$. Choose a collection $\omega = \{p_1, \dots, p_s\}$ of s points in $\mathbb{R}^2$, and denote by $\mathcal{C}(\omega)$ the set of parameterized tropical curves with s-marked points $(C, x_1, \dots, x_s, f)$ satisfying the following conditions

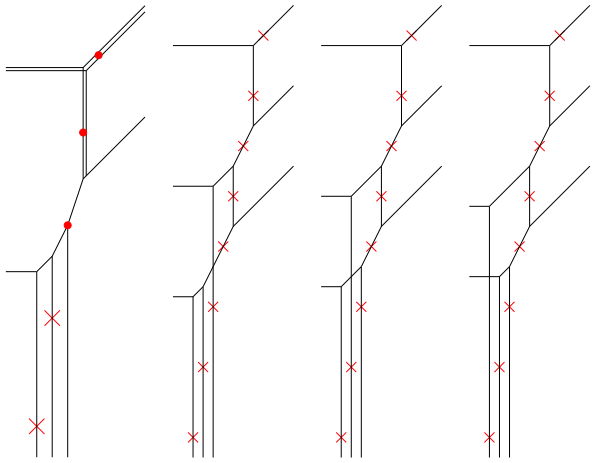
- the tropical curve C is irreducible and of genus g,
- $\Delta(C, f) = \Delta$,
- for any $1 \leq i \leq s$, $f(x_i) = p_i$.

Proposition 4.7 (Mikhalkin, [Mik05])**.** *For a generic configuration of points ω, the set $\mathcal{C}(\omega)$ is finite. Moreover, for any parameterized tropical curve $(C, x_1, \dots, x_s, f)$ in $\mathcal{C}(\omega)$, the curve C has only 1 or 3-valent vertices, the set $\{x_1, \dots, x_s\}$ is disjoint from $Vert(C)$, any leaf of C is of weight 1, and f is a topological immersion. In particular, any neighborhood of any 3-valent vertex of C is never mapped to a segment by f.*

Given a generic configuration ω, we associate a complex multiplicity $\mu^{\mathbb{C}}(\widetilde{C})$ to any element $\widetilde{C} = (C, x_1, \dots, x_s, f)$ in $\mathcal{C}(\omega)$. Let v be a vertex of $C \setminus \mathrm{End(C)}$ and e_1 and e_2 two of its adjacent edges. As v is trivalent, the balancing condition implies that the number $\mu^{\mathbb{C}}(v, f) = w_{f,e_1} w_{f,e_2} |\det(u_{f,e_1}, u_{f,e_2})|$ does not depend on the choice of e_1 and e_2.



a) b) $\mu^{\mathbb{C}}=4$ c) $\mu^{\mathbb{C}}=1$ d) $\mu^{\mathbb{C}}=1$ e) $\mu^{\mathbb{C}}=1$ f) $\mu^{\mathbb{C}}=1$



g) $\mu^{\mathbb{C}}=1$ h) $\mu^{\mathbb{C}}=1$ i) $\mu^{\mathbb{C}}=1$ j) $\mu^{\mathbb{C}}=1$

FIGURE 7. $N(\Delta_3,0)=12$

Definition 4.8. *The complex multiplicity of an element $\widetilde{C}$ of $\mathcal{C}(\omega)$, denoted by $\mu^{\mathbb{C}}(\widetilde{C})$, is defined as*

$$\mu^{\mathbb{C}}(\widetilde{C})=\prod_{v\in Vert(\widetilde{C})}\mu^{\mathbb{C}}(v,f)$$

Theorem 4.9 (Mikhalkin, [Mik05])**.** *For any lattice polygon Δ, any genus g, and any generic configuration ω of $Card(\partial\Delta \cap \mathbb{Z}^2) - 1 + g$ points in $\mathbb{R}^2$, one has*

$$N(\Delta, g) = \sum_{\widetilde{C} \in \mathcal{C}(\omega)} \mu^{\mathbb{C}}(\widetilde{C})$$

Example 4.10. Images $f(C)$ of all irreducible tropical curves of genus 0 and Newton polygon Δ_3 in $\mathcal{C}(\omega)$ for the configuration ω of 8 points depicted in Figure 7a are depicted in Figure 7b, ..., j. We verify that $N(\Delta_3, 0) = 12$ (compare with Table 1).

4.3. Real multiplicities of a tropical curve

We explain now how to adapt Theorem 4.9 to real enumerative geometry. Naturally, we need to consider tropical curves endowed with a real structure.

Definition 4.11. *A real parameterized tropical curve with n marked points is an $(n+3)$-uplet $(C, x_1, \dots, x_n, f, \phi)$ where $(C, x_1, \dots, x_n, f)$ is a parameterized marked tropical curve and $\phi : C \to C$ is an isometric involution such that*

- *there exists a permutation σ such that for any $1 \leq i \leq n$, $\phi(x_i) = x_{\sigma(i)}$,*
- *$f = f \circ \phi$.*

The *real and imaginary parts* of a real parameterized tropical curve with n marked points $\widetilde{C} = (C, x_1, \dots, x_n, f, \phi)$ are naturally defined as

$$\Re(\widetilde{C}) = \mathrm{Fix}(\phi) \quad \text{and} \quad \Im(\widetilde{C}) = C \setminus \Re(\widetilde{C})$$

Example 4.12. Two examples of real parameterized tropical curves with 4 marked points are depicted in Figure 8, the abstract curve is depicted on the left and its image by f in $\mathbb{R}^2$ is depicted on the right. Very close edges in the image represent edges which are mapped to the same edge by f. The parameterized tropical curve in Figure 8a has 2 equal marked points, and ϕ is the symmetry with respect to the non-leaf edge. In Figure 8b, ϕ exchanges the edges containing x_1 and x_2.

As usual, we fix a lattice polygon Δ and define $s = Card(\partial\Delta \cap \mathbb{Z}^2) - 1$. Let r be an non-negative integer such that $s - 2r \geq 0$, and choose a collection $\omega_r = \{p_1, \dots, p_{s-r}\}$ of $s-r$ points in $\mathbb{R}^2$. We should think of ω_r as the image under the map $(z, w) \mapsto (\log|z|, \log|w|)$ of a configuration $\{q_1, \dots, q_{s-2r}, q_{s-2r+1}, \overline{q_{s-2r+1}}, \dots, q_{s-r}, \overline{q_{s-r}}\}$ of s points in $(\mathbb{C}^*)^2$, where $\overline{z}$ is the complex conjugated of z. Hence, points p_i with $s - 2r + 1 \leq i \leq s - r$ represent pairs of complex conjugated points. Denote by $\mathbb{R}\mathcal{C}(\omega_r)$ the set of irreducible real parameterized tropical curves with s marked points $\widetilde{C} = (C, x_1, \dots, x_s, f, \phi)$ of genus 0 and Newton polygon Δ satisfying the following conditions

- for any $1 \leq i \leq s - 2r$, $f(x_i) = p_i$,
- for any $1 \leq i \leq r$, $f(x_{s-2r+2i-1}) = f(x_{s-2r+2i}) = p_{s-2r+i}$,
- if $1 \leq i \leq r$ and if $x_{s-2r+2i-1} = x_{s-2r+2i}$, then $x_{s-2r+2i}$ is a vertex of C,
- any edge in $\Re(\widetilde{C})$ has an odd weight.

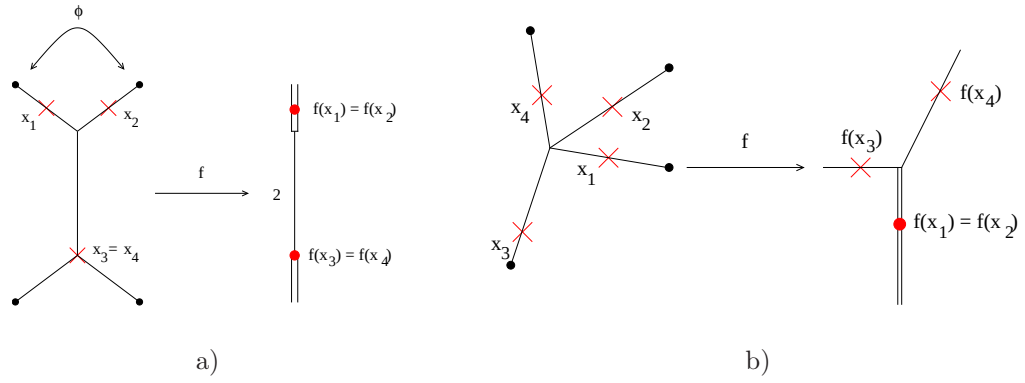


FIGURE 8. Real tropical curves

Proposition 4.13. *For a generic configuration of points ω_r, the set $\mathbb{RC}(\omega_r)$ is finite. Moreover, for any real parameterized curve $\widetilde{C} = (C, x_1, \dots, x_s, f, \phi)$ in $\mathbb{RC}(\omega_r)$, the curve C has only 1, 3 or 4 valent vertices, any neighborhood of any 3 or 4-valent vertex of C is never mapped to a segment by f, any 4-valent vertex of C is adjacent to 2 edges in $\Im(\widetilde{C})$ and 2 edges in $\Re(\widetilde{C})$, and any leaf of C is of weight 1.*

Proof. Let $\widetilde{C}$ be an element of $\mathbb{RC}(\omega_r)$. Passing through $s-r$ points in $\mathbb{R}^2$ in general position imposes $2(s-r)$ independent conditions on a tropical curve. Since all tropical maps are piecewise-linear, to prove the proposition it suffices to show that the dimension of the space of all real parameterized tropical curves with the same combinatorial type as $\widetilde{C}$ has dimension $2(s-r)$, and that any curve with this combinatorial type satisfies the proposition.

Recall that the space of all parameterized irreducible tropical curves (C, f) of genus 0 with x leaves and of a given combinatorial type is a polyhedral complex of dimension

$$x - 1 - \sum_{v \in \mathrm{Vert}(C) \setminus \mathrm{End}(C)} (\mathrm{val}(v) - 3) - n_c$$

where $\mathrm{val}(v)$ is the valence of a vertex v, and n_c is the number of edges of C contracted by f (see [Mik05]). Let $\widetilde{C} = (C, x_1, \dots, x_s, f, \phi)$ be an element of $\mathbb{RC}(\omega_r)$. We may prepare two auxiliary tropical curves $f^r : C^r \to \mathbb{R}^2$ and $f^i : C^i \to \mathbb{R}^2$ from $f : C \to \mathbb{R}^2$. We say that $v \in C$ is a *junction* vertex if any small neighborhood of v intersects both $\Re(\widetilde{C})$ and $\Im(\widetilde{C})$. We denote by J the number of junction vertices of C. Since any edge of $\Re(\widetilde{C})$ has odd weight, a junction vertex is at least 4-valent.

We define C^r to be the result of adding to $\Re(\widetilde{C})$ an infinite ray at each junction vertex of C. We define f^r so that it coincides with f on $\Re(\widetilde{C})$. The values of f^r at the new rays are determined by the balancing condition.

Connected components of $\Im(\widetilde{C})$ are naturally coupled in pairs exchanged by the map ϕ. To define C^i, we take $\Im(\widetilde{C})/\phi$ and replace all edges adjacent to a junction vertex with an infinite ray. We let $f^i : C^i \to \mathbb{R}^2$ to be the tropical map that agrees with f on $\Im(\widetilde{C})/\phi$. We denote by n^i the number of connected components of C^i. Note that $n^i \geq J$, and that equality holds if and only if each junction vertex is 4-valent.

We denote by x^r (resp. x^i) the number of leaves of C which are also leaves of C^r (resp. C^i). Since the curve C has genus 0, the curve C^r is connected and each component of C^i is adjacent to exactly one junction vertex. Hence, the space of parameterized tropical curves with the same combinatorial type as (C^r, f^r) has dimension

$$x^r + J - 1 - \sum_{v \in \mathrm{Vert}(C^r) \setminus \mathrm{End}(C^r)} (\mathrm{val}(v) - 3) - n_{c^r}$$

and the space of parameterized tropical curves with the same combinatorial type as (C^i, f^i) has dimension

$$x^i - \sum_{v \in \mathrm{Vert}(C^i) \setminus \mathrm{End}(C^i)} (\mathrm{val}(v) - 3) - n_{c^i}$$

To get f from f^r and f^i these maps must agree at each junction. Thus each connected component of C^i imposes one condition, and the space of real parameterized tropical curves with the same combinatorial type as (C, f) has dimension

$$x^r + x^i + J - n^i - 1 - \sum_{v \in \mathrm{Vert}(C^r) \setminus \mathrm{End}(C^r)} (\mathrm{val}(v) - 3) - \sum_{v \in \mathrm{Vert}(C^i) \setminus \mathrm{End}(C^i)} (\mathrm{val}(v) - 3) - n_{c^r} - n_{c^i}$$

If we consider, in addition, a configuration of s points on C then our dimension increases by s. Recall though that our points are constrained by the condition that the last $2r$ points are split into pairs invariant with respect to the involution ϕ. Furthermore, recall that if such a pair consists of the same point taken twice then it must be a vertex of C.

Denote with p the number of pairs of distinct points in C invariant with respect to ϕ and with q the number of pairs made from the vertices of C. Clearly we have $p + q = r$, and the dimension of allowed configurations is $s - p - 2q$. Since $f^i(C^i)$ passes through p generic points in $\mathbb{R}^2$, we have $x_i \geq p$, and since $x^r + 2x^i \leq s + 1$ we have $x^r + x^i \leq s + 1 - p$. Hence, the space of real parameterized tropical curves with s marked points with the same combinatorial type as $\widetilde{C}$ has dimension at most

$$2(s - r) + J - n^i - \sum_{v \in \mathrm{Vert}(C^r) \setminus \mathrm{End}(C^r)} (\mathrm{val}(v) - 3) - \sum_{v \in \mathrm{Vert}(C^i) \setminus \mathrm{End}(C^i)} (\mathrm{val}(v) - 3) - n_{c^r} - n_{c^i}$$

Since $\widetilde{C}$ is in $\mathbb{RC}(\omega_r)$, its space of deformation must have dimension at least $2(s - r)$. Hence the curve C has exactly $s + 1$ leaves, $n^i = J$, and any vertex of C which is not an end or a junction vertex is trivalent. □

For a generic configuration ω_r and $(C, x_1, \dots, x_s, f, \phi)$ in $\mathbb{RC}(\omega_r)$, Proposition 4.13 implies that the real structure ϕ on C is uniquely determined by the marked parameterized tropical curve $(C, x_1, \dots, x_s, f)$. Hence we will often omit to precise the map ϕ for elements of $\mathbb{RC}(\omega_r)$. Moreover, $\Im(\widetilde{C})/\phi$ is a (possibly disconnected) non-compact graph, and a vertex v (resp. edge) inside $\Im(\widetilde{C})/\phi$ has a natural complex multiplicity $\mu^{\mathbb{C}}(v, f)$ (resp. weight) induced by the corresponding multiplicity of vertices (resp. edges) of C. If v is a 4-valent vertex of C, then by Proposition 4.13, there exist an edge $e_1 \in \Re(\widetilde{C})$ and an edge $e_2 \in \Im(\widetilde{C})$ adjacent to v. Define $\mu_r^{\mathbb{R}}(v, f) = w_{f,e_1} w_{f,e_2} |det(v_{f,e_1}, v_{f,e_2})|$. Define the integer $o_r^{\mathbb{R}}$ to be the number of vertices v in $\Re(\widetilde{C})$ satisfying one of the following conditions

- v is 3-valent and $\mu^{\mathbb{C}}(v, f) = 3 \ mod \ 4$,
- v is 4-valent adjacent to an edge $e \in \Im(\widetilde{C})$, and $\mu^{\mathbb{R}}(v, f) = w_{f,e} + 1 \ mod \ 2$.

Finally, define the integer $o_r^{\mathbb{C}}$ to be the number of vertices v of $\Im(\widetilde{C})/\phi$ with odd $\mu^{\mathbb{C}}(v, f)$.

Definition 4.14. *The r-real multiplicity of an element $\widetilde{C}$ of $\mathbb{RC}(\omega_r)$, denoted by $\mu_r^{\mathbb{R}}(\widetilde{C})$, is defined as*

$$\mu_r^{\mathbb{R}}(\widetilde{C}) = (-1)^{o_r^{\mathbb{R}} + o_r^{\mathbb{C}}} \prod_{v \in Vert(\Im(\widetilde{C})/\phi)} \mu^{\mathbb{C}}(v, f) \prod_{\substack{v \in Vert(\widetilde{C}), \\ f(v) \in \omega_r}} \mu^{\mathbb{C}}(v, f) \prod_{\substack{v \in Vert(\widetilde{C}), \\ v \ is \ 4\text{-}valent}} \mu^{\mathbb{R}}(v, f)$$

The tropical curves and their multiplicity we are considering here differ slightly from the one in [Shu06]. This difference comes from the fact that we are dealing with parameterization of tropical curves, and that Shustin deals with the cycles resulting as the images of the curves rather than parameterized curves.

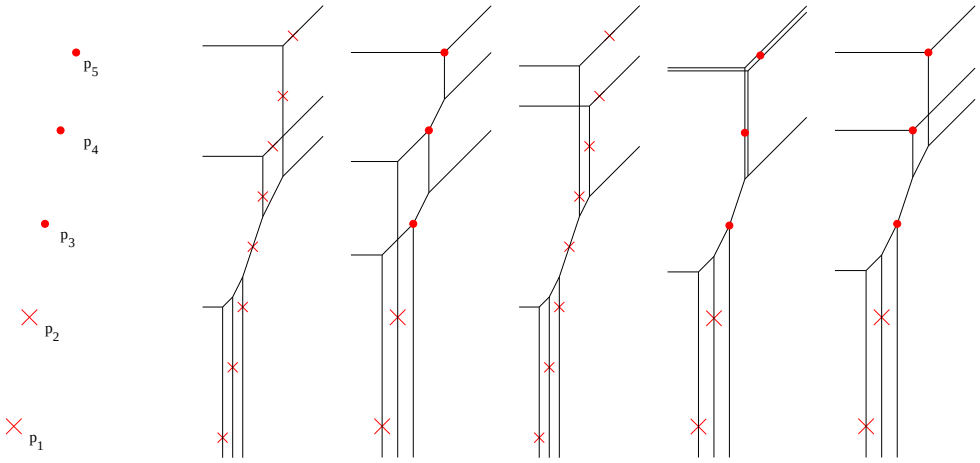
Remark 4.15. If $r = 0$, then for any real parameterized curve $(C, x_1, \dots, x_s, f, \phi)$ in $\mathbb{RC}(\omega_r)$, we have $\phi = Id$, and the map $(C, x_1, \dots, x_s, f, \phi) \mapsto (C, x_1, \dots, x_s, f)$ is a bijection from the set $\mathbb{RC}(\omega_r)$ to the set of elements of $\mathcal{C}(\omega_r)$ with odd complex multiplicity.

Theorem 4.16 (Mikhalkin, [Mik05], Shustin, [Shu06])**.** *Let Δ be a lattice polygon such that Welschinger invariants are defined for the corresponding toric surface $Tor(\Delta)$ equipped with its tautological real structure. Then for any integer r such that $s - 2r \geq 0$, and any generic configuration ω_r of $s - r$ points in $\mathbb{R}^2$, one has*

$$W(\Delta, r) = \sum_{\widetilde{C} \in \mathbb{RC}(\omega_r)} \mu_r^{\mathbb{R}}(\widetilde{C})$$

Remark 4.17. Theorem 4.16 implies that the right hand side of last equality does not depend on ω_r for smooth Del Pezzo toric surfaces $Tor(\Delta)$. However, this is not true in general and one can easily check that the sum of r-real multiplicities over all tropical curves in $\mathbb{RC}(\omega_r)$ in the case of $r > 0$ does not have to be invariant if $Tor(\Delta)$ is singular (see also [ABLdM, Section 7.2]).

Example 4.18. If $\omega_3 = \{p_1, p_2, p_3, p_4, p_5\}$ is the configuration depicted in Figure 9a, then images of all parameterized tropical curves of genus 0 and Newton polygon Δ_3 in $\mathbb{RC}(\omega_3)$ are depicted in Figures 9b, c, d, e, and f (compare with Table 1). Figure 9e) represents the image of 2 distinct marked parameterized tropical curves in $\mathbb{RC}(\omega_3)$, depending on the position of marked points on the connected components of $\Im(\widetilde{C})$. Hence we verify that $W(\Delta_3, 3) = 2$.



a) b) $\mu_3^{\mathbb{R}} = 1$ c) $\mu_3^{\mathbb{R}} = 1$ d) $\mu_3^{\mathbb{R}} = 1$ e) $\mu_3^{\mathbb{R}} = -1$ 2 choices f) $\mu_3^{\mathbb{R}} = 1$

FIGURE 9. $W(\Delta_3, 3) = 2$

5. Proof of Theorems 3.6 and 3.9

Theorems 3.6 and 3.9 are obtained by applying Theorems 4.9 and 4.16 to configurations ω which are stretched in the direction $(0, 1)$.

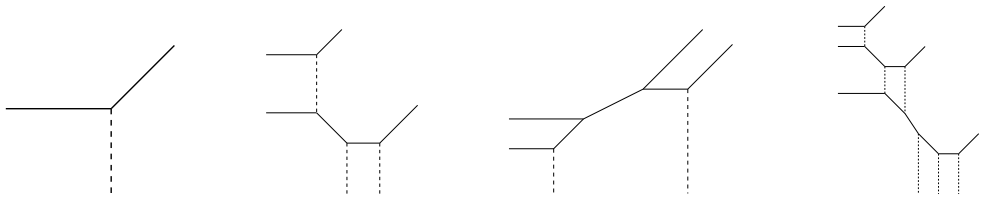
5.1. Floors of a parameterized tropical curve

As we fixed a preferred direction in $\mathbb{R}^2$, it is natural to distinguish between edges of parameterized tropical curves which are mapped parallely to this direction from the others.

Definition 5.1. *An elevator of a parameterized tropical curve (C, f) is an edge e of C with $u_{f,e} = \pm(0, 1)$. The set of elevators of (C, f) is denoted by $\mathcal{E}(f)$. If an elevator e is not a leaf of C, then e is said to be bounded. A floor of a parameterized tropical curve (C, f) is a connected component of the topological closure of $C \setminus (\mathcal{E}(f) \cup End(C))$.*

Naturally, a floor of a parameterized marked tropical curve is a floor of the underlying parameterized tropical curve.

Example 5.2. In Figure 10 are depicted some images of parameterized tropical curves. Elevators are depicted in dotted lines.



a) One floor b) Two floors c) One floor d) Three floors

FIGURE 10. Floors of tropical curves

Let us fix a h-transverse polygon Δ, and a non-negative integer number g. Define $s = Card(\partial\Delta \cap \mathbb{Z}^2) - 1 + g$, and choose a generic configuration ω of s points in $\mathbb{R}^2$. If moreover $g = 0$, choose r an non-negative integer such that $s - 2r \geq 0$, and choose a collection ω_r of $s - r$ points in $\mathbb{R}^2$.

Proposition 5.3. *Let $I = [a; b]$ be a bounded interval of $\mathbb{R}$. Then, if ω (resp. ω_r) is a subset of $I \times \mathbb{R}$, then any vertex of any curve in $\mathcal{C}(\omega)$ (resp. $\mathbb{R}\mathcal{C}(\omega_r)$) is mapped to $I \times \mathbb{R}$.*

Proof. Suppose that there exists an element $(C, x_1, \dots, x_s, f)$ in $\mathcal{C}(\omega)$ or $\mathbb{R}\mathcal{C}(\omega_r)$, and a vertex v of C such that $f(v) = (x_v, y_v)$ with $x_v < a$. Choose v such that no vertex of C is mapped by f to the half-plane $\{(x, y) \mid x < x_v\}$. Suppose that v is a trivalent vertex of C, and denote by e_1, e_2 and e_3 the three edges of C adjacent to v. For $1 \leq i \leq 3$, choose the vector u_{e_i} pointing away from v (see section 4.1). By assumption on v, this vertex is adjacent to a leaf of C, for example e_1, and since Δ is h-transverse we have $u_{f,e_1} = (-1, \alpha)$. Moreover, according to Propositions 4.7 and 4.13, we have $w_{f,e_1} = 1$. By the balancing condition, up to exchanging e_2 and e_3, we have $u_{f,e_2} = (-\beta, \gamma)$ with $\beta \geq 0$, and $u_{f,e_3} = (\delta, \varepsilon)$ with $\delta > 0$. Moreover, as no vertex of C is mapped to the half-plane $\{(x, y) \mid x < x_v\}$, the edge $f(e_2)$ is a leaf of C if $\beta > 0$. Then, by translating the vertex $f(v)$ (resp. and possibly $\phi(v)$) in the direction u_{f,e_3}, we construct a 1-parameter family of parameterized tropical curves in $\mathcal{C}(\omega)$ (resp. $\mathbb{R}\mathcal{C}(\omega_r)$), as depicted in two examples in Figure 11. This contradicts Propositions 4.7 and 4.13. If v is a 4-valent vertex of C, then we construct analogously a 1-parameter family of parameterized tropical curves in $\mathbb{R}\mathcal{C}(\omega_r)$. Alternatively, the contradiction may be derived from [Mik05, Lemma 4.17]. Hence, no vertex of C is mapped by f in the half-plane $\{(x, y) \mid x < a\}$.

The case where there exists an element $(C, x_1, \dots, x_s, f)$ in $\mathcal{C}(\omega)$ or $\mathbb{R}\mathcal{C}(\omega_r)$, and a vertex v of C such that $f(v) = (x_v, y_v)$ with $x_v > b$ works analogously. □

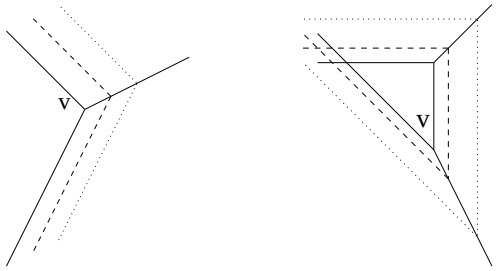


FIGURE 11. 1-parameter family of tropical curves

Corollary 5.4. *Let I be a bounded interval of $\mathbb{R}$. If ω (resp. ω_r) is a subset of $I \times \mathbb{R}$ and if the points of ω (resp. ω_r) are far enough one from the others, then any floor of any curve in $\mathcal{C}(\omega)$ (resp. $\mathbb{R}\mathcal{C}(\omega_r)$) can not contain more than one (resp. two) distinct marked point. If a floor of an element $\widetilde{C}$ in $\mathbb{R}\mathcal{C}(\omega_r)$ contains two distinct marked points, then they are contained in $\Im(\widetilde{C})$.*

Proof. Let $(C, x_1, \dots, x_s, f)$ be an element of $\mathcal{C}(\omega)$ or $\mathbb{R}\mathcal{C}(\omega_r)$ and choose a path γ in $C \setminus \mathcal{E}(C)$. The number of edges of C is bounded from above by a number which depends only on Δ and g, and according to the tropical Bézout Theorem, absolute value of the coordinates of the vector $w_{f,e}u_{f,e}$ for any edge e of C is bounded from above by a number which depends only on Δ. According to Proposition 5.3, all vertices of C are mapped by f to the strip $I \times \mathbb{R}$, so the length (for the Euclidean metric in $\mathbb{R}^2$) of $f(\gamma)$ is bounded from above by a number $l_{max}(\Delta, g)$ which depends only on Δ and g. Hence, if the distance between the points p_i is greater than $l_{max}(\Delta, g)$, two distinct marked points x_i which are not mapped to the same p_j cannot be on the same floor of C. □

For the remaining of this section, let us fix a bounded interval I of $\mathbb{R}$, a configuration $\omega = \{p_1, \dots, p_s\}$, or possibly a configuration $\omega_r = \{p_1, \dots, p_{s-r}\}$, such that the point p_i is very much higher than the points p_j if $j < i$. Here, *very much higher* means that we can apply Corollary 5.4. Actually, we prove in next corollary that any floor of any curve in $\mathcal{C}(\omega)$ or $\mathbb{R}\mathcal{C}(\omega_r)$ contains *exactly* one marked point. More precisely, we have the following statement.

Corollary 5.5. *Let $\widetilde{C}$ be an element of $\mathcal{C}(\omega)$ or $\mathbb{R}\mathcal{C}(\omega_r)$. Then, any floor of $\widetilde{C}$ contains exactly one marked point. Moreover, the curve $\widetilde{C}$ has exactly $Card(d_l(\Delta))$ floors and $Card(d_l(\Delta)) + g + d_-(\Delta) + d_+(\Delta) - 1$ elevators.*

Proof. Let us denote by f_i (resp b_i, $\widetilde{d}_i$) the number of floors (resp. bounded elevators, elevators) of $\widetilde{C}$ containing i marked points. According to Corollary 5.4, $f_i = 0$ as soon as $i \geq 3$, and since the points p_i are in general position, we have $b_i = \widetilde{d}_i = 0$ as soon as

$i \geq 2$. What we have to prove is that $f_0 = f_2 = b_0 = \widetilde{d}_0 = 0$. We have two expressions for the number s which gives us the equation

$$f_1 + 2f_2 + \widetilde{d}_1 = d_+(\Delta) + d_-(\Delta) + 2Card(d_l(\Delta)) - 1 + g \tag{3}$$

According to tropical Bézout Theorem and Corollary 5.4, if a floor of C contains two marked points, then the intersection number of this floor with a generic tropical line is at least 2. Hence we have

$$f_0 + f_1 + 2f_2 \leq Card(d_l(\Delta)) \tag{4}$$

According to Propositions 4.7 and 4.13, we have $d_+ + d_-$ leaves of C which are elevators, thus

$$b_0 + b_1 = \widetilde{d}_0 + \widetilde{d}_1 - (d_+(\Delta) + d_-(\Delta)) \tag{5}$$

An Euler characteristic computation shows us that

$$f_0 + f_1 + f_2 - b_0 - b_1 \geq 1 - g \tag{6}$$

Combining Equations (3) with (4), then with Equation (5), and finally with Equation (6), we obtain

$$f_1 + f_2 \geq Card(d_l(\Delta))$$

which is compatible with Equation (4) if an only if $f_0 = f_2 = 0$. Moreover, in this case inequalities (6) and (4) are actually equalities, which implies $b_0 = \widetilde{d}_0 = 0$. □

5.2. From tropical curves to floor diagrams

To a parameterized tropical curve (C, f), we associate the following oriented weighted graph, denoted by $\mathcal{F}(C, f)$: vertices of $\mathcal{F}(C, f)$ correspond to floors of (C, f), and edges of $\mathcal{F}(C, f)$ correspond to elevators of (C, f). Edges of $\mathcal{F}(C, f)$ inherit a natural weight from weight of (C, f). Moreover, $\mathbb{R}$ is naturally oriented, and edges of $\mathcal{F}(C, f)$ inherit this orientation, since they are all parallel to the coordinate axis $\{0\} \times \mathbb{R}$. Note that we do *not* consider the graph $\mathcal{F}(C, f)$ as a metric graph and that some leaves are non-compact.

Example 5.6. The graphs corresponding to parameterized tropical curves depicted in Figure 10 are depicted in Figure 12. Floors are depicted by ellipses, and elevators by segments. As all elevators have weight 1, we do not precise them on the picture. Orientation is implicitly from down to up.

Let $\widetilde{C}$ be a parameterized tropical curve in $\mathcal{C}(\omega)$ or $\mathbb{R}\mathcal{C}(\omega_r)$. Since $\widetilde{C}$ has exactly $Card(d_l(\Delta))$ floors, any floor ε of $\widetilde{C}$ has a unique leaf e with $u_{f,e} = (-1, -\alpha)$ where $u_{f,e}$ points to infinity. Hence the following map is well defined

$$\begin{array}{rccc} \theta : & \mathrm{Vert}(\mathcal{F}(\widetilde{C})) & \longrightarrow & \mathbb{Z} \\ & \varepsilon & \longmapsto & \alpha \end{array}$$

The following lemma follows directly from Corollary 5.5 and Definition 3.1 of a floor diagram.

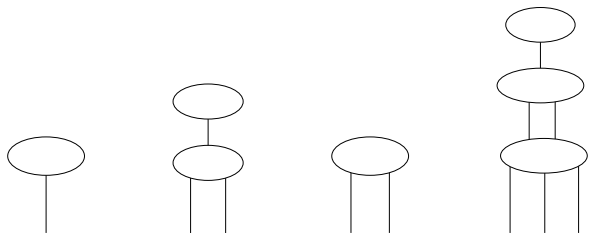


FIGURE 12. Graphs associated to tropical curves

Lemma 5.7. *The graph $\mathcal{F}(\widetilde{C})$ equipped with the map θ is a floor diagram of genus g and Newton polygon Δ.*

Let us denote by $\mathcal{D}(\widetilde{C})$ this floor diagram. Finally we associate to a parameterized tropical curve with n marked points $\widetilde{C} = (C, x_1, \dots, x_n, f)$ in $\mathcal{C}(\omega)$ or $\mathbb{R}\mathcal{C}(\omega_r)$ a marking m of the floor diagram $\mathcal{D}(\widetilde{C})$. The natural idea is to map the points i to the floor or elevator of C containing x_i. However, it can happen if $\widetilde{C}$ is in $\mathbb{R}\mathcal{C}(\omega_r)$ that $x_i = x_{i+1}$ is a vertex v of C. In this case, according to Proposition 4.13 and Corollary 5.5, v is on a floor ε and is adjacent to an elevator e of $\mathcal{D}(\widetilde{C})$. If $u_{f,e} = (0, 1)$ points away from v (resp. to v), then we define $m(i) = \varepsilon$ and $m(i+1) \in e$ (resp. $m(i+1) = \varepsilon$ and $m(i) \in e$). If x_i is not a vertex of C, then we define $m(i)$ as the floor or a point on the edge of C which contains x_i.

The map $m : \{1, \dots, Card(\partial\Delta \cap \mathbb{Z}) - 1 + g\} \to \mathcal{D}(\widetilde{C})$ is clearly an increasing map, hence it is a marking of the floor diagram $\mathcal{D}(\widetilde{C})$. In other words, we have a map $\Phi : \widetilde{C} \mapsto (\mathcal{D}(\widetilde{C}), m)$ from the set $\mathcal{C}(\omega)$ (resp. $\mathbb{R}\mathcal{C}(\omega_r)$) to the set of marked floor diagrams (resp. r-real marked floor diagrams with non-null r-real multiplicity) of genus g and Newton polygon Δ.

Example 5.8. All marked floor diagrams with a non-null complex multiplicity (resp. 3-real multiplicity) in Table 1 correspond exactly to parameterized tropical curves whose image in $\mathbb{R}^2$ are depicted in Figure 7 (resp. 9).

Theorems 3.6 and 3.9 are now a corollary of the next proposition.

Proposition 5.9. *The map Φ is a bijection. Moreover, for any element $\widetilde{C}$ in $\mathcal{C}(\omega)$ (resp. $\mathbb{R}\mathcal{C}(\omega_r)$), one has $\mu^{\mathbb{C}}(\widetilde{C}) = \mu^{\mathbb{C}}(\Phi(\widetilde{C}))$ (resp. $\mu_r^{\mathbb{R}}(\widetilde{C}) = \mu_r^{\mathbb{R}}(\Phi(\widetilde{C}))$).*

Proof. The fact that Φ is a bijection is clear when $Card(d_l(\Delta)) = 1$. Hence the map Φ is always a bijection since an element of $\mathcal{C}(\omega)$ (resp. $\mathbb{R}\mathcal{C}(\omega_r)$) is obtained by gluing, along elevators, tropical curves with a single floor which are uniquely determined by the points p_i they pass through.

Let $\widetilde{C} = (C, x_1, \dots, x_s, f)$ be an element of $\mathcal{C}(\omega)$, and v a vertex of C. According to Corollary 5.4 and Corollary 5.5, v is adjacent to an elevator of weight w and to an edge e on a floor with $u_{f,e} = (\pm 1, \alpha)$ and $w_{f,e} = 1$. Hence, $\mu^{\mathbb{C}}(v, f) = w$. Since any leaf of C is of weight 1, it follows that $\mu^{\mathbb{C}}(\widetilde{C})$ is the product of the square of the multiplicity of all elevators of C, that is equal to $\mu^{\mathbb{C}}(\Phi(\widetilde{C}))$.

Let $\widetilde{C} = (C, x_1, \dots, x_s, f, \phi)$ be an element of $\mathbb{R}\mathcal{C}(\omega_r)$. The same argument as before shows that $\mu_r^{\mathbb{R}}(\widetilde{C})$ and $\mu_r^{\mathbb{R}}(\Phi(\widetilde{C}))$ have equal absolute values. It remains us to prove that both signs coincide, and the only thing to check is that the number $o_r^{\mathbb{R}}$ is even. If v is a 4-valent vertex of C adjacent to an edge e in $\Im(\widetilde{C})$, then $\mu^{\mathbb{R}}(v) = w_{f,e}$. If v is a 3-valent vertex in $\Re(\widetilde{C})$ adjacent to an elevator e, then $\mu^{\mathbb{C}}(v) = w_{f,e}$. So if $\mu^{\mathbb{C}}(v) = 3 \ mod\ 4$, then e is bounded and the other vertex v' adjacent to e satisfy also $\mu^{\mathbb{C}}(v) = 3 \ mod\ 4$. Hence the number $o_r^{\mathbb{R}}$ is even as announced. □

6. Some applications

Here we use floor diagrams to confirm some results in classical enumerative geometry.

6.1. Degree of the discriminant hypersurface of the space of plane curves

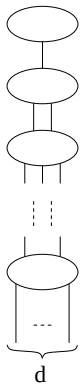


FIGURE 13. Unique floor diagram of maximal genus and Newton polygon Δ_d

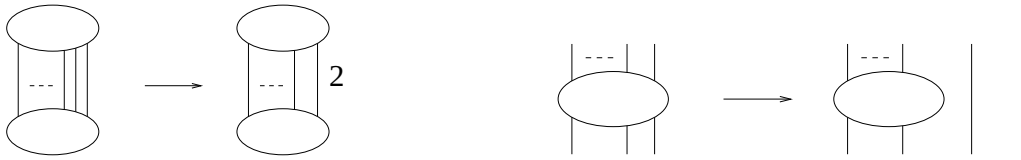
Proposition 6.1. *For any $d \geq 3$, one has*

$$N(\Delta_d, \frac{(d-1)(d-2)}{2} - 1) = 3(d-1)^2$$

Proof. We see easily that the unique floor diagram $\mathcal{D}_{max}$ of genus $\frac{(d-1)(d-2)}{2}$ and Newton polygon Δ_d is the one depicted in Figure 13. Moreover, all floor diagrams of genus $\frac{(d-1)(d-2)}{2}-1$ and Newton polygon Δ_d are obtained by decreasing the genus of $\mathcal{D}_{max}$ via one of the 2 moves depicted in Figure 14. There are $i-1$ different markings of the floor diagram obtained via the move of Figure 14a, and $2i+1$ different markings of the floor diagram obtained via the move of Figure 14b. Then we get

$$\begin{aligned} N(d, \tfrac{(d-1)(d-2)}{2}-1) &= \textstyle\sum_{i=2}^{d-1} 4(i-1) + \sum_{i=2}^{d}(2i-1) \\ &= 3(d-1)^2 \end{aligned}$$

□



a) i edges $\to i-1$ edges
$\mu^{\mathbb{C}} = 4$

b) i outgoing edges $\to i-1$ outgoing edges
$\mu^{\mathbb{C}} = 1$

FIGURE 14. Decrease by 1 the genus of the floor diagram of maximal genus

6.2. Asymptotic of Welschinger invariants

In [Mik05], a combinatorial algorithm in terms of *lattice paths* has been given to enumerate complex and real curves in toric surfaces. The idea is that when we consider (the right number of) points which are sufficiently far one from the other but on the same line L with irrational slope, then all tropical curves passing through these points can be recovered inductively. Hence, if L is the line with equation $x+\varepsilon y$ with y a very small irrational number, then lattice paths and floor diagrams are two ways to encode the same tropical curves. However, in our opinion, floor diagrams are much easier to deal with. In particular, one does not have to consider reducible curves using floor diagrams.

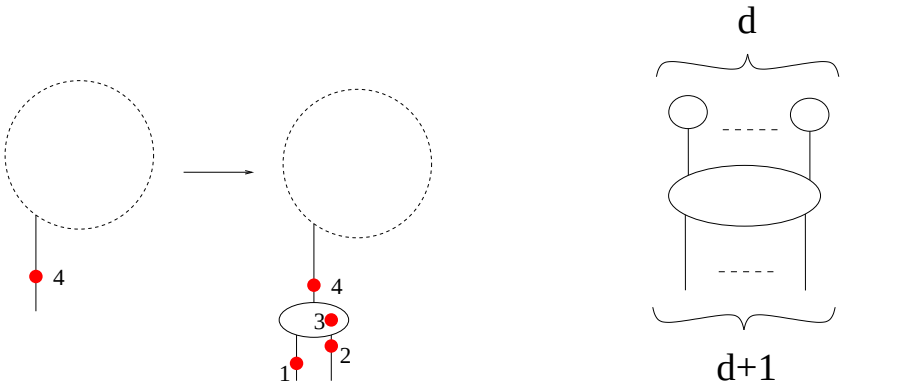
As an example, we give a floor diagram proof of the following theorem that was initially proved with the help of the lattice paths.

Theorem 6.2 (Itenberg, Kharlamov, Shustin [IKS03] [IKS04])**.** *The sequence* $(W(\Delta_d, 0))_{d\geq 1}$ *satisfies the following properties:*

- *it is a sequence of positive numbers,*
- *it is an increasing sequence, and strictly increasing starting from* $d=2$*,*
- *one has* $\ln W(\Delta_d, 0) \sim \ln N(\Delta_d, 0) \sim 3d \ln d$ *when* d *goes to infinity.*

Proof. As we have $\mu_0^{\mathbb{R}} = 1$ for any floor diagram, the numbers $W(\Delta_d, 0)$ are all non-negative. Moreover, we have $W(\Delta_1, 0) = 1$ so the positivity of these numbers will follow from the increasingness of the sequence $(W(\Delta_d, 0))_{d\geq 1}$.

Let $(\mathcal{D}_0, m_0)$ be a marked floor diagram of genus 0 and Newton polygon Δ_d. For convenience we use marking $m_0 : \{4, \dots, 3d+2\} \to D_0$ (instead of the "usual" marking $\{1, \dots, 3d-1\} \to \mathcal{D}_0$). Note that the point 4 has to be mapped to an edge in $\mathrm{Edge}^{-\infty}(\mathcal{D}_0)$. Out of $\mathcal{D}_0$, we can construct a new marked floor diagram $\mathcal{D}$ of genus 0 and Newton polygon Δ_{d+1} as indicated in Figure 15a. Both real multiplicities $\mu_0^{\mathbb{R}}(\mathcal{D}_0)$ and $\mu_0^{\mathbb{R}}(\mathcal{D})$ are the same, and two distinct marked floor diagrams $\mathcal{D}_0$ and $\mathcal{D}_0'$ give rise to two distinct marked floor diagrams $\mathcal{D}$ and $\mathcal{D}'$. Hence, we have $W(\Delta_{d+1}, 0) \geq W(\Delta_d, 0)$ for all $d \geq 1$. Moreover, if $d \geq 2$ then there exist marked floor diagrams with Newton polygon Δ_{d+1} which are not obtained out of a marked floor diagrams with Newton polygon Δ_d as described above. An example is given in Figure 15b, hence $W(\Delta_{d+1}, 0) > W(\Delta_d, 0)$ if $d \geq 2$.



a) From Δ_d to Δ_{d+1} b) Not obtained from Δ_d

FIGURE 15. The numbers $W(\Delta_d, 0)$ are increasing

We study now the logarithmic asymptotic of the sequence $(W(\Delta_d, 0))_{d\geq 1}$. Let $(\mathcal{D}_d)_{d\geq 1}$ be the sequence of floor diagrams constructed inductively in the following way: $\mathcal{D}_1$ is the floor diagram with Newton polygon Δ_1, and $\mathcal{D}_d$ is obtained out of $\mathcal{D}_{d-1}$ by gluing to each edge in $\mathrm{Edge}^{-\infty}(\mathcal{D}_{d-1})$ the piece depicted in Figure 16a. Floor diagrams $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$, and $\mathcal{D}_4$ are depicted in Figures 16b, c, d et e. The floor diagram $\mathcal{D}_d$ is of degree 2^{d-1} and we have $\mu_0^{\mathbb{R}}(\mathcal{D}_d) = 1$. If $\nu(\mathcal{D}_d)$ denotes the number of distinct markings of $\mathcal{D}_d$, then we have

$$\begin{aligned} \forall d \geq 2 \qquad \nu(\mathcal{D}_d) &= \tfrac{\nu(\mathcal{D}_{d-1})^2}{2} C^{3\times 2^{d-2}-2}_{3\times 2^{d-1}-4} \\ &= \frac{(3\times 2^{d-1}-4)!}{2^{2^{d-1}-1}} \prod_{i=2}^{d} \frac{1}{((3\times 2^{d-i}-2)(3\times 2^{d-i}-3))^{2^{i-1}}} \end{aligned}$$

Hence we get

$$\frac{(3\times 2^{d-1}-4)!}{2^{2^d}\prod_{i=1}^{d}(3\times 2^{d-i})^{2^i}} \leq \nu(\mathcal{D}_d) \leq (3\times 2^{d-1}-4)!$$

The Stirling Formula implies that $\ln d! \sim d\ln d$, and we see easily that both right and left hand side of the inequality have the same logarithmic asymptotic, namely $3\times 2^{d-1}\ln(2^{d-1})$. As we have $\ln\nu(\mathcal{D}_d) \leq \ln W(\Delta_{2^{d-1}},0) \leq \ln N(\Delta_{2^{d-1}},0)$, the result follows from the increasingness of the sequence $(W(\Delta_d,0))_{d\geq 1}$ and from the equivalence $\ln(N(\Delta_d,0)) \sim 3d\ln d$ proved in [DFI95]. □

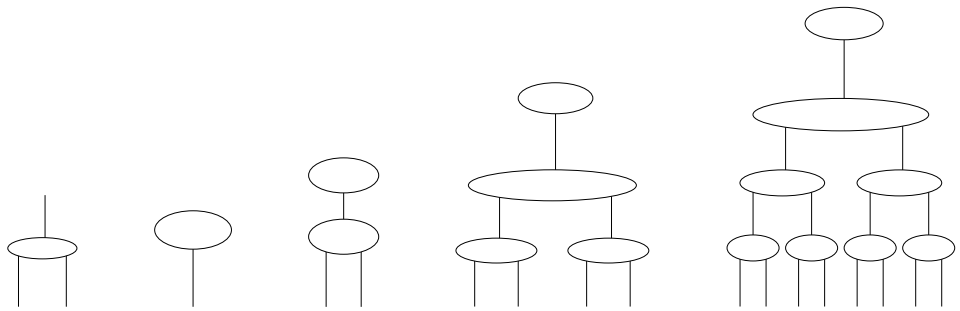


FIGURE 16. Asymptotic of the numbers $W(\Delta_d,0)$

6.3. Recursive formulas

Floor diagrams allow one to write down easily recursive formulas in a Caporaso-Harris style (see [CH98]) for both complex and real enumerative invariants. The recipe to extract such formulas is explained in [ABLdM] in the particular case of the numbers $W(\Delta_d,r)$.

As an example we briefly outline here how to reconstruct Vakil's formula [Vak00], which relates some enumerative invariants of Hirzebruch surfaces.

The Hirzebruch surface $\mathbb{F}_n$ of degree n, with $n \geq 0$, is the compactification of the line bundle over $B = \mathbb{C}P^1$ with first Chern class n. If $\mathbb{F}_n \supset F \approx \mathbb{C}P^1$ denotes the

compactification of a fiber, then the second homology group of $\mathbb{F}_n$ is the free abelian group generated by B and F. In a suitable coordinate system, a generic algebraic curve in $\mathbb{F}_n$ of class $aB+bF$, with $a,b\geq 0$, has the h-transverse Newton polygon $\Delta_{n,a,b}$ with vertices $(0,0)$, $(na+b,0)$, $(0,a)$, and (b,a) (see [Bea83] for more details about Hirzebruch surfaces).

Before stating the theorem, we need to introduce some notations. In the following, $\alpha=(\alpha_1,\alpha_2,\ldots)$ denotes a sequence of non-negative integers, and we set

$$|\alpha|=\sum_{i=1}^{\infty}\alpha_i,\qquad I\alpha=\sum_{i=1}^{\infty}i\alpha_i,\qquad I^{\alpha}=\prod_{i=1}^{\infty}i^{\alpha_i}$$

If a and b are two integer numbers, $\left(\begin{array}{c} a \\ b \end{array}\right)$ denotes the binomial coefficient. If a and $b_1,b_2,\ldots,b_k$ are integer numbers then $\left(\begin{array}{c} a \\ b_1,\ldots,b_k \end{array}\right)$ denotes the multinomial coefficient, i.e.

$$\left(\begin{array}{c} a \\ b_1,\ldots,b_k \end{array}\right)=\prod_{i=1}^{k}\left(\begin{array}{c} a-\sum_{j=1}^{i-1}b_j \\ b_i \end{array}\right)$$

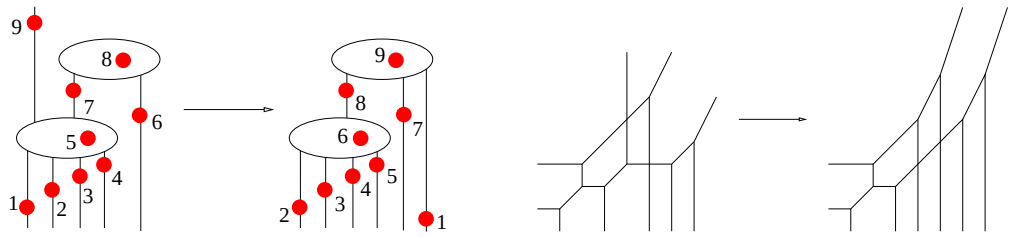
Theorem 6.3 (Vakil, [Vak00])**.** *For any $n\geq 0$, any $g\geq 0$, and any $b\geq 1$, one has*

$$N(\Delta_{n,2,b},g)=N(\Delta_{n+1,2,b-1},g)+\sum_{\substack{I\beta\leq n\\ |\beta|=g+1}}\left(\begin{array}{c} 2n+2b+g+2 \\ n-I\beta \end{array}\right)\left(\begin{array}{c} \beta_1+b \\ b \end{array}\right)\left(\begin{array}{c} |\beta|+b \\ \beta_1+b,\beta_2,\beta_3,\ldots \end{array}\right)I^{2\beta}$$

Proof. We want to enumerate marked floor diagrams of genus g and Newton polygon $\Delta_{n,2,b}$. As these floor diagrams have only two floors, our task is easy. Let $\mathcal{D}$ be such a marked floor diagrams of genus g and Newton polygon $\Delta_{n,2,b}$. Then, the marking m is defined on the set $\{1,\ldots,s\}$ where $s=2(n+2)+2b-1+g$.

Suppose that $m(s)$ is a floor of $\mathcal{D}$. These marked floor diagrams are easy to enumerate, their contribution to the number $N(\Delta_{n,2,b},g)$ is the second term on the right hand side of the equality.

Suppose that $m(s)$ is on an edge e in $\mathrm{Edge}^{+\infty}(\mathcal{D})$. Define a new floor diagram $\mathcal{D}'$ as follows: $\mathrm{Vert}(\mathcal{D}')=\mathrm{Vert}(\mathcal{D}')$, $\mathrm{Edge}(\mathcal{D}')=(\mathrm{Edge}(\mathcal{D})\setminus\{e\})\cup\{e'\}$, where e' is in $\mathrm{Edge}^{-\infty}(\mathcal{D})$ and is adjacent to the other floor than e. Define a marking m' on $\mathcal{D}'$ as follows: $m'(i)=m(i-1)$ if $i\geq 2$ and $m(1)\in e'$. Now, the marked floor diagram $(\mathcal{D}',m')$ is of genus g and Newton polygon $\Delta_{n+1,2,b-1}$ (see Figure 17a). Moreover, we obtain in this way a bijection between the set of marked floor diagrams of genus g and Newton polygon $\Delta_{n,2,b}$ such that $m(s)\in\mathrm{Edge}^{+\infty}(\mathcal{D})$, and marked floor diagrams of genus g and Newton polygon $\Delta_{n+1,2,b-1}$. Hence, we get the first term of the right hand side of the equality, and the theorem is proved. □



a) b)

FIGURE 17. From $\mathbb{F}_n$ to $\mathbb{F}_{n+1}$

Remark 6.4. Our proof of Theorem 6.3 is a combinatorial game on marked floor diagrams that can be obtained as the translation to the floor diagram language of Vakil's original proof: take the highest point p of the configuration, and specialize it to the exceptional section E. Then, either a curve C we are counting breaks into 2 irreducible components, which give the second term, or C has now a prescribed point on E. Blowing up this point and blowing down the strict transform of the fiber, the curve is transformed to a curve in $\mathbb{F}_{n+1}$ with a prescribed point on B (which is the image under the blow down of the second intersection point of C with the fiber).

The effect of such a blow up and down in tropical geometry can be easily seen, since intersection points with E correspond to leaves going up, and intersection with B correspond to leaves going down. An example is given in Figure 17b which correspond to the operation on marked floor diagram depicted in Figure 17a.

7. Further computations

One can adapt the technics of this paper to compute other real and complex enumerative invariants of algebraic varieties. In addition to genus 0 Gromov-Witten invariants and Welschinger invariants of higher dimensional spaces, as announced in [BM07], one can compute in this way characteristic numbers of the projective plane (at least in genus 0 and 1), as well as Gromov-Witten and Welschinger invariants of the blown up projective plane. Details will appear soon.

References

[ABLdM] A. Arroyo, E. Brugallé, and L. Lopez de Medrano, *Recursive formula for Welschinger invariants*, Available at arXiv:0809.1541.

[Bea83] A. Beauville, Complex algebraic surfaces, volume 68 of *London Mathematical Society Lecture Note Series*, Cambridge University Press, Cambridge, 1983.

[BM07] E. Brugallé and G. Mikhalkin, *Enumeration of curves via floor diagrams*, Comptes Rendus de l'Académie des Sciences de Paris, série I, 345, 2007.

[CH98] L. Caporaso and J. Harris, *Counting plane curves of any genus*, Invent. Math., 131(2):345–392, 1998.

[DFI95] P. Di Francesco and C. Itzykson, *Quantum intersection rings*, In The moduli space of curves (Texel Island, 1994), volume 129 of Progr. Math., pages 81–148. Birkhäuser Boston, Boston, MA, 1995.

[FM] S. Fomin and G. Mikhalkin, *Labelled floor diagrams*, To appear.

[GM07] A. Gathmann and H. Markwig, *The numbers of tropical plane curves through points in general position*, Journal für die reine und angewandte Mathematik (Crelle's Journal), 602:155–177, 2007.

[IKS03] I. Itenberg, V. Kharlamov, and E. Shustin, *Welschinger invariant and enumeration of real rational curves*, Int. Math. Research Notices, 49:2639–2653, 2003.

[IKS04] I. Itenberg, V. Kharlamov, and E. Shustin, *Logarithmic equivalence of Welschinger and Gromov-Witten invariants*, Uspehi Mat. Nauk, 59(6):85–110, 2004 (in Russian), English version: arXiv: math.AG/0407188.

[KM94] M. Kontsevich and Yu. Manin, *Gromov-Witten classes, quantum cohomology, and enumerative geometry*, Comm. Math. Phys., 164(3):525–562, 1994.

[Mik] G. Mikhalkin, *Phase-tropical curves in* $\mathbb{R}^n$, In preparation.

[Mik05] G. Mikhalkin, *Enumerative tropical algebraic geometry in* $\mathbb{R}^2$, J. Amer. Math. Soc., 18(2):313–377, 2005.

[Mik07] G. Mikhalkin, *Moduli spaces of rational tropical curves*, Proceedings of Gökova Geometry/Topology conference 2006, pages 39–51, 2007.

[Shu06] E. Shustin, *A tropical calculation of the Welschinger invariants of real toric Del Pezzo surfaces*, J. Algebraic Geom., 15:285–322, 2006. Corrected version available at arXiv:math/0406099.

[Vak00] R. Vakil, *Counting curves on rational surfaces*, Manuscripta Math., 102:53–84, 2000.

[Wel05] J. Y. Welschinger, *Invariants of real symplectic 4-manifolds and lower bounds in real enumerative geometry*, Invent. Math., 162(1):195–234, 2005.

Université Pierre et Marie Curie, Paris 6, 175 rue du Chevaleret, 75 013 Paris, France
E-mail address: brugalle@math.jussieu.fr

Université de Genève, 2-4 rue du Lièvre, Case postale 64, 1211 Genève, Suisse
E-mail address: grigory.mikhalkin@unige.ch

Proceedings of 15th Gökova
Geometry-Topology Conference
pp. 91–119

Slicing planar grid diagrams: a gentle introduction to bordered Heegaard Floer homology

Robert Lipshitz, Peter Ozsváth, Dylan Thurston

Abstract. We describe some of the algebra underlying the decomposition of planar grid diagrams. This provides a useful toy model for an extension of Heegaard Floer homology to 3-manifolds with parametrized boundary. This paper is meant to serve as a gentle introduction to the subject, and does not itself have immediate topological applications.

Contents

1. Introduction 92
2. Background on knot Floer homology and grid diagrams 93
2.1. Planar Floer Homology 95
3. Slicing planar grid diagrams 97
4. Motivating the answer 98
5. The algebra associated to a slicing 102
6. The Type D module 105
7. The Type A module 108
8. The pairing theorem 111
9. Bimodules 112
9.1. Freezing 112
9.2. Type A to Type D 114
9.3. Remarks on Type D to Type A 116
10. How the real world is harder 116
10.1. Complications for $\widehat{HF}$ of 3-manifolds 116
10.2. Complications for toroidal grid diagrams 117
References 118

RL was supported by an NSF Mathematical Sciences Postdoctoral Research Fellowship.
PO was supported by NSF grants number DMS-0505811 and FRG-0244663.
DPT was supported by a Sloan Research Fellowship.

1. Introduction

The Heegaard Floer homology groups of Ozsváth and Szabó are defined in terms of holomorphic curves in Heegaard diagrams. In [7], Heegaard Floer homology is extended to three-manifolds with (parameterized) boundary, by studying holomorphic curves in pieces of Heegaard diagrams. The resulting invariant, *bordered Heegaard Floer homology*, has the following form. To an oriented surface F (together with an appropriate Morse function on F), bordered Heegaard Floer associates a differential graded algebra $\mathcal{A}(F)$. To a three-manifold Y together with a homeomorphism $F \to \partial Y$, bordered Heegaard Floer associates a right ($\mathcal{A}_\infty$) module $\widehat{CFA}(Y)$ over $\mathcal{A}(F)$ and a left (differential graded) module $\widehat{CFD}(Y)$ over $\mathcal{A}(-F)$. (Here, $-F$ denotes F with its orientation reversed.) These modules, which are well-defined up to homotopy equivalence, relate to the closed Heegaard Floer homology group $\widehat{HF}$ via the following *pairing theorem*:

Theorem 1 ([7])**.** *Suppose that* $Y = Y_1 \cup_F Y_2$*. Then* $\widehat{CF}(Y) \simeq \widehat{CFA}(Y_1) \widetilde{\otimes}_{\mathcal{A}(F)} \widehat{CFD}(Y_2)$.

(Recall that $\widehat{CF}(Y)$ is the chain complex underlying the Floer homology group $\widehat{HF}(Y)$. The notation $\widetilde{\otimes}$ denotes the derived tensor product, and the symbol $\simeq$ denotes quasi-isomorphism.)

The definitions of the invariants $\widehat{CFA}$ and $\widehat{CFD}$ are, unfortunately, somewhat involved. There are two kinds of complications which obscure the basic ideas involved:

- **Analytic complications.** The definitions of the invariants $\widehat{CFA}$ and $\widehat{CFD}$ involve counting pseudo-holomorphic curves. In spite of much progress over the last decades, holomorphic curve techniques remain somewhat technical, and often require seemingly unnatural contortions. To make matters worse, the analytic set up is, by necessity, somewhat nonstandard; in particular, it involves counting curves in a manifold with "two kinds of infinities."
- **Algebraic complications.** The invariant $\widehat{CFA}$ is, in general, not an honest module but only an $\mathcal{A}_\infty$-module. While the subject of $\mathcal{A}_\infty$ algebra is increasingly mainstream, it still adds a layer of obfuscation to the study of bordered Heegaard Floer homology. Further exacerbating the situation is a somewhat novel kind of grading.

In developing bordered Heegaard Floer homology we found it useful to study a toy model, in terms of planar grid diagrams, in which these complications are absent. It is the aim of the present paper to present this toy model. We hope that doing so will make the definition of bordered Heegaard Floer homology in [7] more palatable.

We emphasize up front that the main objects of study in this paper do *not* give topological invariants. Still, the algebra involved is reminiscent of well-known objects from representation theory—in particular, the nilCoxeter algebra—so this paper may be of further interest.

Throughout this paper, $\mathbb{F}$ will denote the field with two elements and $\mathbb{A}$ will denote $\mathbb{F}[U_1, \dots, U_N]$ (for whichever N is in play at the time).

Acknowledgements. The first author thanks the organizers of the Gökova Geometry Topology Conference for inviting him to participate in this stimulating event. He also thanks C. Douglas for interesting conversations in the summer of 2006. The authors also thank M. Khovanov for pointing out the relationship of the algebra $\mathcal{A}$ with the nilCoxeter algebra.

2. Background on knot Floer homology and grid diagrams

We start by recalling the combinatorial definition of Manolescu-Ozsváth-Sarkar [8] of the knot Floer homology groups.

Let K be an oriented knot in S^3. Choose a knot diagram D for K such that

- D is composed entirely of horizontal and vertical segments,
- no two horizontal segments of D have the same y-coordinate, and no two vertical segments of D have the same x-coordinate, and
- at each crossing, the vertical segment crosses over the horizontal segment.

(Every knot admits such a diagram; see Figure 1.) The only data in such a diagram are the endpoints of the segments, which we record by placing X's and O's at these endpoints, alternately around the knot, and so that the knot is oriented from X to O along vertical segments. Notice that no two X's (respectively O's) lie on the same horizontal or vertical line.

Let $\mathbb{X} = \{X_i\}_{i=1}^N$ and $\mathbb{O} = \{O_i\}_{i=1}^N$ denote the set of X's and O's, respectively. Up to isotopy of the knot (and renumbering of the X_i), we may assume that the coordinates of X_i are $\left(i - \frac{1}{2}, \sigma_{\mathbb{X}}(i) - \frac{1}{2}\right)$ for some permutation $\sigma_{\mathbb{X}} \in S_N$. Then (after renumbering), the coordinates of O_i are $\left(i - \frac{1}{2}, \sigma_{\mathbb{O}}(i) - \frac{1}{2}\right)$ for some permutation $\sigma_{\mathbb{O}} \in S_N$. The data $(\mathbb{R}^2, \mathbb{X}, \mathbb{O})$ is a *planar grid diagram* for the knot K.

We can also view $\mathbb{X}$ and $\mathbb{O}$ as subsets of the torus $T = \mathbb{R}^2/\langle (N,0), (0,N)\rangle$. The data $(T, \mathbb{X}, \mathbb{O})$ is a *toroidal grid diagram* for the knot K. It is easy to recover the knot K (up to isotopy) from the toroidal grid diagram $(T, \mathbb{X}, \mathbb{O})$. We call the process of passing from a planar grid diagram to a toroidal grid diagram *wrapping*. The inverse operation of passing from a toroidal grid diagram to a planar grid diagram, which depends on a choice of two circles in T, we call *unwrapping*.

The $N+1$ lines $\alpha_i = \{y = i\} \subset \mathbb{R}^2$, $i = 0, \dots, N$, descend to N disjoint circles $\overline{\alpha}_i$ in the torus T, with $\overline{\alpha}_0 = \overline{\alpha}_N$. Similarly, the $N+1$ lines $\beta_i = \{x = i\} \subset \mathbb{R}^2$, $i = 0, \dots, N$, descend to N disjoint circles $\overline{\beta}_i$ in T. Notice that each α_j (respectively $\overline{\alpha}_j$) intersects each β_i (respectively $\overline{\beta}_j$) in a single point. Set $\boldsymbol{\alpha} = \bigcup_{i=0}^N \alpha_i$, $\overline{\boldsymbol{\alpha}} = \bigcup_{i=1}^N \overline{\alpha}_i$, $\boldsymbol{\beta} = \bigcup_{i=0}^N \beta_i$ and $\overline{\boldsymbol{\beta}} = \bigcup_{i=1}^N \overline{\beta}_i$. We view the $\overline{\alpha}_i$ as "horizontal" and the $\overline{\beta}_i$ as "vertical". This means that components of $T \setminus \left(\overline{\boldsymbol{\alpha}} \cup \overline{\boldsymbol{\beta}}\right)$ (little rectangles) have, for instance, *lower left* corners, *lower right* corners, and so on.

We define the knot Floer chain complex $CFK^-(K)$ as follows. Let $\mathbb{A} = \mathbb{F}[U_1, \dots, U_N]$. By a *toroidal generator* we mean an N-tuple of points $\mathbf{x} = \{x_i \in \overline{\alpha}_{\sigma(i)} \cap \overline{\beta}_i\}$, one on each $\overline{\alpha}$-circle and one on each $\overline{\beta}$-circle. Generators, then, are in bijection with the permutation group S_N—but this bijection depends on a choice of unwrapping. Let $\mathfrak{S}(T, \mathbb{X}, \mathbb{O})$ denote

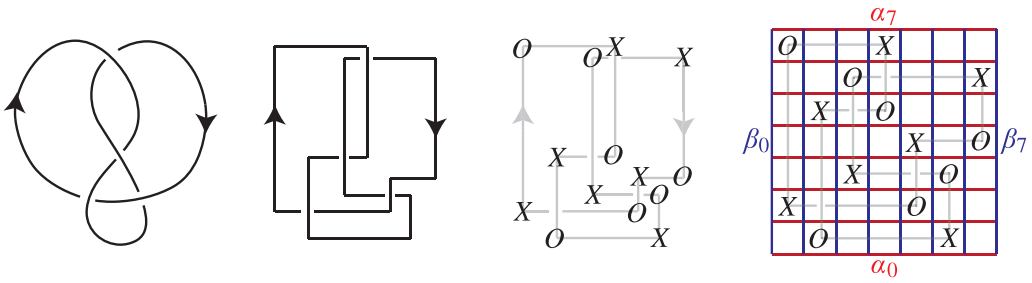


FIGURE 1. **Representing a knot by a grid diagram.** Starting with a knot diagram D, one approximates D using horizontal and vertical segments, so that crossings are always vertical over horizontal. Perturb the result so that no segments lie on the same horizontal or vertical line, and mark the endpoints alternately with X's and O's. The data of the knot is entirely encoded in these X's and O's, which we can see as sitting in the middle of squares on a piece of graph paper.

the set of generators. The knot Floer complex $CFK^-(K)$ is freely generated over $\mathbb{A}$ by $\mathfrak{S}(T,\mathbb{X},\mathbb{O})$.

For two generators $\mathbf{x}=\{x_i\}$ and $\mathbf{y}=\{y_i\}$, we define a set $\overline{\mathrm{Rect}}(\mathbf{x},\mathbf{y})$. The set $\overline{\mathrm{Rect}}(\mathbf{x},\mathbf{y})$ is empty unless all but two of the x_i agree with corresponding y_i. In that case, let $\{i,j\}=\{k\mid x_k\neq y_k\}$; then $\overline{\mathrm{Rect}}(\mathbf{x},\mathbf{y})$ is the set of embedded rectangles R in T with boundary on $\boldsymbol{\alpha}\cup\boldsymbol{\beta}$, and such that x_i and x_j are the lower-left and upper-right corners of R (in either order), and y_i and y_j are the upper-left and lower-right corners of R (in either order). (Consequently, $\overline{\mathrm{Rect}}(\mathbf{x},\mathbf{y})$ always has either zero or two elements.) Call a rectangle $R\in\overline{\mathrm{Rect}}(\mathbf{x},\mathbf{y})$ *empty* if the interior of R contains no point in $\mathbf{x}$, and define $\overline{\mathrm{Rect}}^{\circ}(\mathbf{x},\mathbf{y})$ to be the set of empty rectangles in $\overline{\mathrm{Rect}}(\mathbf{x},\mathbf{y})$. Given a rectangle R, define $O_i(R)$ to be 1 if O_i lies in the interior of R_i and zero otherwise. Define $X_i(R)$ similarly, and set $\mathbb{O}(R)=\sum_{i=1}^{N}O_i(R)$ and $\mathbb{X}(R)=\sum_{i=1}^{N}X_i(R)$. Set $U(R)=\prod_i U_i^{O_i(R)}$.

Now, define

$$\partial\mathbf{x}=\sum_{\mathbf{y}\in\mathfrak{S}(T,\mathbb{X},\mathbb{O})}\sum_{\substack{R\in\overline{\mathrm{Rect}}^{\circ}(\mathbf{x},\mathbf{y})\\ \mathbb{X}(R)=0}}U(R)\cdot\mathbf{y}. \tag{2.1}$$

Lemma 2.2. *Formula (2.1) defines a differential, i.e., $\partial^2=0$.*

This is not hard to prove [9, Proposition 2.8]. See Figure 2 for some of the cases.

By composing rectangles, we get more complicated regions in T, called *domains*. By a *domain connecting* $\mathbf{x}$ *to* $\mathbf{y}$ we mean a cellular two-chain B in $(T,\overline{\boldsymbol{\alpha}}\cup\overline{\boldsymbol{\beta}})$ with the following

property. Let $\partial_\alpha B$ denote the intersection of ∂B with $\boldsymbol{\alpha}$. Then we require $\partial(\partial_\alpha B) = \mathbf{y}-\mathbf{x}$. We can define $O_i(B)$, $X_i(B)$, $\mathbb{O}(B)$, $\mathbb{X}(B)$, and $U(B)$ in the same way as for rectangles.

There are two $\mathbb{Z}$-gradings on $CFK^-(K)$, the *Maslov* or *homological* grading, denoted μ, and the *Alexander grading*, denoted A. These have the property that ∂ preserves A and lowers μ by 1. We give the combinatorial characterization of A and μ from [9], up to an overall shift. First, some notation. Given sets E and F in $\mathbb{R}^2$, let $\mathcal{I}(E,F)$ denote the number of pairs $(e,f) \in E \times F$ such that e lies to the lower left of f (i.e., the number of pairs $e = (e_1, e_2) \in \mathbb{R}^2$ and $f = (f_1, f_2) \in \mathbb{R}^2$, such that $e_1 < f_1$ and $e_2 < f_2$).

Now, fix an unwrapping $(\mathbb{R}^2, \mathbb{X}, \mathbb{O})$ of the diagram $(T, \mathbb{X}, \mathbb{O})$, so a generator $\mathbf{x} \in \mathfrak{S}(T, \mathbb{X}, \mathbb{O})$ corresponds to a N-tuple of points $u(\mathbf{x})$ in $\mathbb{R}^2$. Then, for some constants C_A and C_μ depending on the diagram and the unwrapping (but not on $\mathbf{x}$),

$$A(\mathbf{x}) = \mathcal{I}(\mathbb{X}, \mathbf{x}) - \mathcal{I}(\mathbb{O}, \mathbf{x}) + C_A$$
$$\mu(\mathbf{x}) = \mathcal{I}(\mathbf{x}, \mathbf{x}) - 2\mathcal{I}(\mathbb{O}, \mathbf{x}) + C_\mu,$$

cf. [9, Formulas (1) and (2)], bearing in mind that $I(\mathbb{X}, \mathbf{x})$ differs from $I(\mathbf{x}, \mathbb{X})$ by a constant. Together with the property that $A(U_i) = -1$ and $\mu(U_i) = -2$ this characterizes A and μ up to overall additive constants.

A fundamental result of Manolescu, Ozsváth, and Sarkar [8] states that the complex $CFK^-(K)$ defined above is bi-graded homotopy equivalent to the complex $CFK^-(K)$ defined by Ozsváth and Szabó [10] and also by Rasmussen [11]. It follows, in particular, that the homotopy type of $CFK^-(K)$ is independent of the toroidal grid diagram for K. The fact that the homotopy type of $CFK^-(K)$ depends only on the knot K can also be proved combinatorially [9].

2.1. Planar Floer Homology

In this paper we will study a modification of the grid diagram construction of CFK^-, which we call the *planar Floer homology* and denote CP^-, obtained by replacing toroidal grid diagrams by planar grid diagrams throughout the definition of CFK^-. In the planar setting, when we have N different X's we will have $N+1$ different α- (respectively β-) lines: we view the process of wrapping the diagram as identifying α_0 with α_N, and β_0 with β_N. Thus, a generator over $\mathbb{A}$ of the complex $CP^-(\mathbb{X}, \mathbb{O})$ is an $(N+1)$-tuple of points $\mathbf{x} = \{x_i \in \alpha_{\sigma(i)} \cap \beta_i\}_{i=0}^N$. The set $\mathfrak{S}(\mathbb{R}^2, \mathbb{X}, \mathbb{O})$ is in canonical bijection with the symmetric group S_{N+1}.

Given generators $\mathbf{x}$ and $\mathbf{y}$ in $\mathfrak{S}(\mathbb{R}^2, \mathbb{X}, \mathbb{O})$, let $\mathrm{Rect}^\circ(\mathbf{x}, \mathbf{y})$ denote the set of empty rectangles in $\mathbb{R}^2$ connecting $\mathbf{x}$ to $\mathbf{y}$; for each $\mathbf{x}$ and $\mathbf{y}$ the set $\mathrm{Rect}^\circ(\mathbf{x}, \mathbf{y})$ is either empty or has a single element. The differential on CP^- is defined analogously to Formula (2.1):

$$\partial \mathbf{x} = \sum_{\mathbf{y} \in \mathfrak{S}(\mathbb{R}^2, \mathbb{X}, \mathbb{O})} \sum_{\substack{R \in \mathrm{Rect}^\circ(\mathbf{x}, \mathbf{y}) \\ \mathbb{X}(R) = 0}} U(R) \cdot \mathbf{y}. \tag{2.3}$$

Lemma 2.4. *Formula (2.3) defines a differential, i.e., $\partial^2 = 0$.*

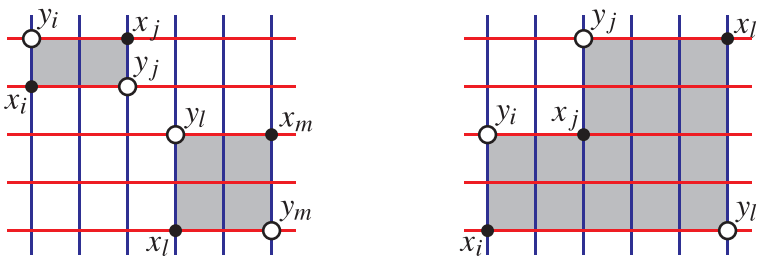


FIGURE 2. **Illustration of why $\partial^2 = 0$ for planar Floer homology.** Left: The contributions to the coefficient of $\mathbf{y}$ from taking the two shaded rectangles in the two orders cancel. Right: This "L"-shaped domain can be decomposed into two rectangles in two different ways, by making either a horizontal cut or a vertical cut. These two contributions cancel.

The proof, which is a strict sub-proof of the proof for toroidal grid diagrams, is illustrated in Figure 2.

The complex $CP^-(\mathbb{X}, \mathbb{O})$ has Alexander and Maslov gradings A and μ, defined exactly as they were for $CFK^-(K)$. We fix the additive constants by setting

$$A(\mathbf{x}) = \mathcal{I}(\mathbb{X}, \mathbf{x}) - \mathcal{I}(\mathbb{O}, \mathbf{x})$$
$$\mu(\mathbf{x}) = \mathcal{I}(\mathbf{x}, \mathbf{x}) - 2\mathcal{I}(\mathbb{O}, \mathbf{x}).$$

***Warning:* The homotopy type of the complex $CP^-(\mathbb{X}, \mathbb{O})$ is *not* an invariant of the underlying knot K.** This is illustrated in Example 2.5. The results of this paper, thus, do not directly give new topological invariants.

Example 2.5. Consider the planar grid diagrams for the unknot shown in Figure 3. The diagram on the left has $N = 1$. The complex has two generators over $\mathbb{F}[U_1]$, which we label with the permutations $[1\,2]$ and $[2\,1]$ in one-line notation. (Here the one-line notation $[2\,3\,1]$, for instance, means the permutation $\{1 \mapsto 2, 2 \mapsto 3, 3 \mapsto 1\}$.) The differential is trivial, so the homology of the complex is $\mathbb{F}[U_1]^{\oplus 2}$.

The diagram on the right has $N = 2$. The complex has six generators. The differential is given by

$$\partial[2\,3\,1] = U_1[3\,2\,1]$$
$$\partial[3\,1\,2] = U_2[3\,2\,1]$$
$$\partial[3\,2\,1] = \partial[1\,2\,3] = \partial[1\,3\,2] = \partial[2\,1\,3] = 0.$$

The homology of the complex is

$$\mathbb{F}\left\langle [3\,2\,1] \right\rangle \oplus \mathbb{F}[U_1, U_2] \left\langle [1\,2\,3], [1\,3\,2], [2\,1\,3], U_2[2\,3\,1] + U_1[3\,1\,2] \right\rangle.$$

This is certainly not the same as $\mathbb{F}[U_1]^{\oplus 2}$.

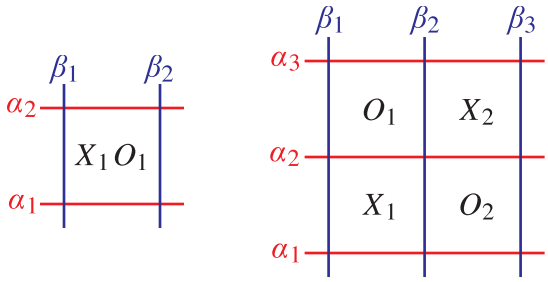


FIGURE 3. **Planar grid diagrams for the unknot.** Left: a diagram with $N = 1$. Right: a diagram with $N = 2$. The corresponding planar Floer complexes are not homotopy equivalent (Example 2.5).

3. Slicing planar grid diagrams

Fix a planar grid diagram $\mathcal{H} = (\mathbb{R}^2, \mathbb{X}, \mathbb{O})$. The goal of this paper is to compute the complex $CP^-(\mathcal{H})$ by cutting the diagram vertically into pieces. (For now, we consider only cutting $\mathcal{H}$ into two pieces; we will consider more general cuttings in Section 9.1.) We want to associate something (ultimately, it will be a differential module) to each side, and something (ultimately, it will be a differential graded algebra) to the interface between the two sides. We want these to contain enough information to reconstruct $CP^-(\mathcal{H})$—but as little information as possible beyond that, so as to be computable.

So, let Z be the vertical line $\{x = k - 1/4\}$ and consider what each side of Z looks like. To the left of Z we have k vertical lines $\beta_0, \dots, \beta_{k-1}$, as well as two injective maps $\mathbb{X}^A \colon \{1, \dots, k\} \to \{1, \dots, N\}$ and $\mathbb{O}^A \colon \{1, \dots, k\} \to \{1, \dots, N\}$. Similarly, to the right of Z we have $N + 1 - k$ vertical lines $\beta_k, \dots, \beta_N$, as well as two injective maps $\mathbb{X}^D \colon \{k+1, \dots, N\} \to \{1, \dots, N\}$ and $\mathbb{O}^D \colon \{k+1, \dots, N\} \to \{1, \dots, N\}$. There are also $N + 1$ α-lines, which intersect both sides of the diagram. Finally, at the interface Z we see $N + 1$ points $\{(i, k - 1/4)\}_{i=0}^N$ where the α_i intersect Z. See Figure 4.

Let H^A denote the half-plane to the left of Z, and H^D the half-plane to the right of Z. We will call the data $\mathcal{H}^A = (H^A, \mathbb{X}^A, \mathbb{O}^A)$ or $\mathcal{H}^D = (H^D, \mathbb{X}^D, \mathbb{O}^D)$ a *partial planar grid diagram*. If we view Z as oriented upwards then there is a distinction between $\mathcal{H}^A$ and $\mathcal{H}^D$: for $\mathcal{H}^A$ the induced orientation of Z agrees with the given one, while for $\mathcal{H}^D$ the induced orientation differs. We will call the first case "type A" and the second case "type D." We say that $\mathcal{H}^A$ has *height* $N + 1$ and *width* k, and $\mathcal{H}^D$ has *height* $N + 1$ and *width* $N + 1 - k$.

Finally, a generator $\mathbf{x} = \{x_i\}_{i=0}^N$ corresponds to k points $\mathbf{x}^A = \{x_i \in \alpha_{\sigma^A(i)} \cap \beta_i\}_{i=0}^{k-1}$ to the left of Z and $N+1-k$ points $\mathbf{x}^D = \{x_i \in \alpha_{\sigma^D(i)} \cap \beta_i\}_{i=k}^N$ to the right of Z. Here, σ^A is an injection $\{0, \dots, k-1\} \to \{0, \dots, N\}$ and σ^D is an injection $\{k, \dots, N\} \to \{0, \dots, N\}$. For use later, let $\mathfrak{S}(\mathcal{H}^A)$ denote the set of k-tuples $\mathbf{x}^A = \{x_i \in \alpha_{\sigma^A(i)} \cap \beta_i\}_{i=0}^{k-1}$ where σ^A is

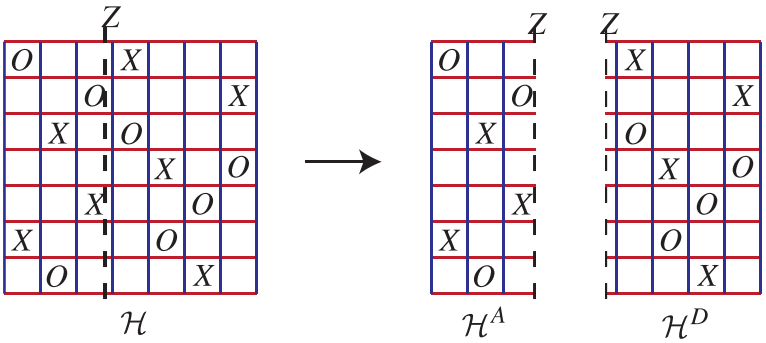


FIGURE 4. **Cutting a planar grid diagram.** The resulting diagrams $\mathcal{H}^A$ and $\mathcal{H}^D$ have $N = 7$ and $k = 3$.

an injection $\{0,\dots,k-1\} \to \{0,\dots,N\}$. Let $\mathfrak{S}(\mathcal{H}^D)$ denote the set of $(N+1-k)$-tuples $\mathbf{x}^D = \{x_i \in \alpha_{\sigma^D(i)} \cap \beta_i\}_{i=k}^N$ where σ^D is an injection $\{k,\dots,N\} \to \{0,\dots,N\}$.

4. Motivating the answer

The purpose of this section is to motivate the answers which will be described in later sections; thus, it can be skipped by the impatient reader without sacrificing mathematical content.

We want to associate some kind of object, which with hindsight we will call $CPA^-(\mathcal{H}^A)$ to $\mathcal{H}^A$, and some other kind of object $CPD^-(\mathcal{H}^D)$ to $\mathcal{H}^D$. These should be objects in some (algebraic) categories $\mathcal{C}^A$ and $\mathcal{C}^D$ associated to the interface Z (together perhaps with a little additional data). We would also like a pairing map P from $\mathcal{C}^A \times \mathcal{C}^D$ to $\mathcal{D}^b(\mathbb{A}-\mathsf{Mod})$, the derived category of complexes over the ground ring $\mathbb{A}$, so that $CP^-(\mathcal{H}) = P(\mathcal{H}^A, \mathcal{H}^D)$. The (derived) category of chain complexes of (right/left) $\mathcal{A}$-modules for any $\mathbb{A}$-algebra $\mathcal{A}$ admit such a pairing map, so this seems like a reasonable example to keep in mind. (That is also how the story goes in Khovanov homology [3], which is encouraging.)

Since a generator $\mathbf{x}$ of $CP^-(\mathcal{H})$ decomposes as a pair $(\mathbf{x}^A, \mathbf{x}^D)$, it seems reasonable that $CPA^-(\mathcal{H}^A)$ would be generated—in some sense to be determined—by $\mathfrak{S}(\mathcal{H}^A)$ and that $CPD^-(\mathcal{H}^D)$ would be generated by $\mathfrak{S}(\mathcal{H}^D)$.

Not every pair $(\mathbf{x}^A, \mathbf{x}^D) \in \mathfrak{S}(\mathcal{H}^A) \times \mathfrak{S}(\mathcal{H}^D)$ corresponds to a generator in $\mathfrak{S}(\mathcal{H})$: the necessary and sufficient condition is that the images of the injections σ^A and σ^D be disjoint. It seems reasonable that our putative $\mathcal{A}$ would remember this—that if σ_1^A and σ_2^A have different images then corresponding generators $\mathbf{x}_1^A$ and $\mathbf{x}_2^A$ would "live over" different "objects" in $\mathcal{A}$. In the language of differential graded categories (see, e.g., [2]), this makes sense; for algebras this can be encoded via idempotents. That is, suppose $\mathcal{A}$ has $\binom{N+1}{k}$ different primitive idempotents I_S, one for each k-element subset S of $\{0,\dots,N\}$. Then we could say $\mathbf{x}^A I_S = \mathbf{x}^A$ if and only if $S = \mathrm{Im}(\sigma^A)$, and $I_S \mathbf{x}^D = \mathbf{x}^D$ if and only if $S \cap \mathrm{Im}(\sigma^D) = \emptyset$; otherwise these products are 0. It then follows that an expression of the

form $\mathbf{x}^A \otimes_{\mathcal{A}} \mathbf{x}^D$ is nonzero if and only if $(\mathbf{x}^A, \mathbf{x}^D)$ actually corresponds to a generator in $\mathfrak{S}(\mathcal{H})$. We will write $S(\mathbf{x}^A)$ to denote $\mathrm{Im}(\sigma^A)$, and $S(\mathbf{x}^D)$ to denote $\{0,\dots,N\}\setminus \mathrm{Im}(\sigma^D)$.

There are three kinds of rectangles which contribute to the differential on $CP^-(\mathcal{H})$:

- Rectangles contained entirely in $\mathcal{H}^A$. It seems reasonable that these should contribute to a differential on $CPA^-(\mathcal{H}^A)$, and there is an obvious way for them to do so.
- Rectangles contained entirely in $\mathcal{H}^D$. Again, it seems reasonable to let these contribute to a differential on $CPD^-(\mathcal{H}^D)$.
- Rectangles which cross through the interface Z. It is somewhat less clear how to count these.

Let R be a rectangle crossing through Z. Each of $CPA^-(\mathcal{H}^A)$ and $CPD^-(\mathcal{H}^D)$ see Z as a half strip, and these half strips should somehow be involved in the definitions of $CPA^-(\mathcal{H}^A)$ and $CPD^-(\mathcal{H}^D)$. The rectangle R intersects Z in a segment running from some α_i to some α_j (with $i<j$ by convention). If R is in $\mathrm{Rect}(\mathbf{x},\mathbf{y})$, with $\mathbf{x}=(\mathbf{x}^A,\mathbf{x}^D)$ and $\mathbf{y}=(\mathbf{y}^A,\mathbf{y}^D)$, then the objects (idempotents) associated to $\mathbf{x}^A$ and $\mathbf{y}^A$ differ: $S(\mathbf{y}^A)=\big(S(\mathbf{x}^A)\setminus i\big)\cup j$. The objects $S(\mathbf{x}^D)$ and $S(\mathbf{y}^D)$ differ in the same way. So, we could view $R\cap Z$ as an "arrow" from $S(\mathbf{x}^A)$ to $S(\mathbf{y}^A)$ or, in the algebra language, as an element ρ of $\mathcal{A}$ for which $I_{S(\mathbf{x}^A)}\cdot\rho\cdot I_{S(\mathbf{y}^A)}=\rho$.

Actually, since a single rectangle in $\mathcal{H}$ can be in $\mathrm{Rect}(\mathbf{x},\mathbf{y})$ for many different $\mathbf{x}$ and $\mathbf{y}$, the chord $R\cap Z$ gives many arrows. More specifically, for any set S with $i\in S$ and $j\notin S$, $R\cap Z$ gives an arrow $\rho_{S,i,j}$, with the property that $I_S\cdot\rho_{S,i,j}\cdot I_T=\rho_{S,i,j}$, where $T=(S\setminus i)\cup j$. We can view these as coming from a single element $\rho_{i,j}=\sum_S \rho_{S,i,j}$ by multiplying with an idempotent. In some sense, $\rho_{i,j}$ "is" $R\cap Z$.

With this in mind, there are two ways we can think of the effect of the rectangle R on one of the sides:

- It could start at Z, as the element $\rho_{i,j}$, and then come in to act on the module, moving one of the dots in the generator $\mathbf{x}$ to get the new generator $\mathbf{y}$ (if not blocked). This is the point of view we will take for CPA^-.
- It could originate inside the partial diagram, and then propagate out to the boundary (if not blocked), leaving a residue $\rho_{i,j}$ in $\mathcal{A}$ when it reaches the boundary. This is the point of view we will take for CPD^-.

The two perspectives fit naturally with the pairing theorem: each rectangle crossing the boundary starts in $\mathcal{H}^D$, propagates out to the boundary, and then propagates through to $\mathcal{H}^A$.

More precisely, define $CPA^-(\mathcal{H}^A)$ to be generated *over the base ring* $\mathbb{A}$ by $\mathfrak{S}(\mathcal{H}^A)$. We have already defined an action of the idempotents of $\mathcal{A}$ on CPA^-. Define a right action of $\mathcal{A}$ on CPA^- by setting $\mathbf{x}^A\cdot\rho_{i,j}=U(H)\cdot\mathbf{y}^A$ if there is an empty half strip H connecting $\mathbf{x}^A$ and $\mathbf{y}^A$ with rightmost edge equal to $\rho_{i,j}$ (and not crossing any X_k). (Here $U(H)$ is the obvious extension of the earlier notation to domains with boundary on Z.) Define the product to be zero otherwise. Define the differential on CPA^- to count rectangles entirely contained in $\mathcal{H}^A$, in the obvious way.

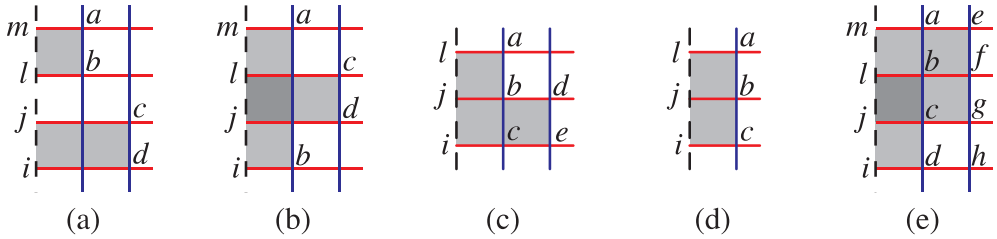


FIGURE 5. **Domains in $\mathcal{H}^D$ forcing relations and a differential $\mathcal{A}$.** Part (a) forces $\rho_{i,j}$ and $\rho_{l,m}$ to commute. Part (b) forces $\rho_{i,m}$ and $\rho_{j,l}$ to commute. Part (c) forces $\rho_{i,j} \cdot \rho_{j,l} = \rho_{i,l}$. Part (d) forces the algebra to have a differential, and part (e) forces the product $\rho_{i,l} \cdot \rho_{j,m}$ to vanish.

Define $CPD^-(\mathcal{H}^D)$ to be "freely" generated as a left $\mathcal{A}$-module by $\mathfrak{S}(\mathcal{H}^D)$. (More precisely, CPD^- is as free as possible given the action of the idempotents we have already defined. It is a direct sum of elementary modules, one for each element of $\mathfrak{S}(\mathcal{H}^D)$.) Thus, the module structure on CPD^- is rather dull. Define the differential on CPD^- as follows: given generators $\mathbf{x}^D, \mathbf{y}^D \in \mathfrak{S}(\mathcal{H}^D)$, define $\mathrm{Half}^\circ(\rho_{i,j}; \mathbf{x}^D, \mathbf{y}^D)$ to be the set of empty half strips connecting $\mathbf{x}^D$ to $\mathbf{y}^D$ with boundary $\rho_{i,j}$; see Figure 9. (The set $\mathrm{Half}^\circ(\rho_{i,j}; \mathbf{x}^D, \mathbf{y}^D)$ is either empty or has a single element.) Define

$$\partial \mathbf{x}^D = \sum_{\mathbf{y}^D} \sum_{\substack{R \in \mathrm{Rect}^\circ(\mathbf{x}^D, \mathbf{y}^D) \\ \mathbb{X}(R)=0}} U(R) \cdot \mathbf{y} + \sum_{\mathbf{y}^D} \sum_{\rho_{i,j}} \sum_{\substack{H \in \mathrm{Half}^\circ(\rho_{i,j}; \mathbf{x}^D, \mathbf{y}^D) \\ \mathbb{X}(H)=0}} U(H) \cdot \rho_{i,j} \mathbf{y}.$$

Remark 4.1. The A in CPA^- is a mnemonic for the fact that the half-strips are included in the *a*lgebra *a*ction on CPA^-. The D in CPD^- is a mnemonic for the fact that the half-strips are included in the *d*ifferential on CPD^-.

It is fairly clear that $CPA^-(\mathcal{H}^A) \otimes_{\mathcal{A}} CPD^-(\mathcal{H}^D) = CP^-(\mathcal{H})$. All rectangles not crossing the interface are obviously accounted for. If $R \in \mathrm{Rect}(\mathbf{x}, \mathbf{y})$ is a rectangle crossing the interface, with $R \cap Z = \rho_{i,j}$, then

$$\partial(\mathbf{x}^A \otimes \mathbf{x}^D) = \mathbf{x}^A \otimes (\partial \mathbf{x}^D) + \cdots = \mathbf{x}^A \otimes \rho_{i,j} \cdot \mathbf{y}^D + \cdots = \mathbf{x}^A \rho_{i,j} \otimes \mathbf{y}^D + \cdots = \mathbf{y}^A \otimes \mathbf{y}^D + \cdots,$$

as desired.

What is not clear—and, *a priori*, not true—is that CPA^- and CPD^- are, in fact, chain complexes (differential modules) over $\mathcal{A}$. Indeed, trying to make CPD^- into a module forces certain relations—and a differential—on the algebra $\mathcal{A}$.

Consider the module $CPD^-(\mathcal{H}^D)$. In Part (a) of Figure 5 is a plausible piece of $\mathcal{H}^D$. One sees here several generators; we single out $\{a, c\}$, $\{a, d\}$, $\{b, c\}$ and $\{b, d\}$. Parts of

the shaded region contribute to the differential as follows:

$$\begin{aligned}\partial\{a,c\} &= \rho_{l,m}\{b,c\} + \rho_{i,j}\{a,d\} + \cdots \\ \partial\{b,c\} &= \rho_{i,j}\{b,d\} + \cdots \\ \partial\{a,d\} &= \rho_{l,m}\{b,d\} + \cdots .\end{aligned}$$

(Here, the dots indicate contributions from regions of the diagram other than the shaded one. The philosophy is that cancellation should be local in $\mathcal{H}^D$.)

Thus, one has

$$\partial^2\{a,c\} = (\rho_{l,m}\cdot\rho_{i,j} + \rho_{i,j}\cdot\rho_{l,m})\{b,d\} + \cdots .$$

So, in order to have $\partial^2 = 0$ we should require that $\rho_{i,j}$ and $\rho_{l,m}$ commute.

Similarly, one sees by examining the shaded region in Part (b) of Figure 5 that $\rho_{i,m}$ and $\rho_{j,l}$ should commute.

In Part (c), consider the differentials

$$\begin{aligned}\partial\{a,d\} &= \rho_{i,j}\{a,e\} + \rho_{i,l}\{c,d\} + \cdots \\ \partial\{a,e\} &= \rho_{j,l}\{b,e\} + \cdots \\ \partial\{c,d\} &= \{b,e\} + \cdots .\end{aligned}$$

Here,

$$\partial^2\{a,d\} = \rho_{i,j}\cdot\rho_{j,l}\{b,e\} + \rho_{i,l}\{b,e\} + \cdots .$$

Thus, we should set $\rho_{i,j}\cdot\rho_{j,l} = \rho_{i,l}$—a relation which looks rather reasonable in its own right.

Part (d) is a little trickier. Considering the generators $\{a\}$, $\{b\}$ and $\{c\}$ we have

$$\begin{aligned}\partial\{a\} &= \rho_{j,l}\{b\} + \rho_{i,l}\{c\} + \cdots \\ \partial\{b\} &= \rho_{i,j}\{c\} + \cdots \\ \partial\{c\} &= 0 + \cdots .\end{aligned}$$

Thus, it seems we have $\partial^2\{a\} = \rho_{j,l}\cdot\rho_{i,j}\{c\}$. One might try setting $\rho_{j,l}\cdot\rho_{i,j} = 0$, but it turns out this is inconsistent with CPA^-. Instead, we set (in this case)

$$\partial\rho_{i,l} = \rho_{j,l}\cdot\rho_{i,j}.$$

Then it follows that $\partial^2\{a\} = 0$. Thus, we were forced to introduce a differential on our algebra $\mathcal{A}$.

Note that, in our example, $j \in S(\{a\})$. In general, we define

$$\partial(\rho_{S,i,l}) = \sum_{\substack{j\in S \\ i<j<l}} \rho_{S,j,l}\cdot\rho_{i,j}.$$

This takes care of the example discussed above. The Leibniz rule extends ∂ to all of $\mathcal{A}$.

Part (e) is the most complicated. We will consider $\partial^2\{b,e\}$. We compute

$$\begin{aligned}
\partial\{b,e\} &= \{a,f\} + \rho_{j,l}\{c,e\} + \rho_{i,l}\{d,e\}\\
\partial\{a,f\} &= \rho_{j,m}\{c,f\} + \rho_{i,m}\{d,f\} + \rho_{j,l}\{a,g\} + \rho_{i,l}\{a,h\}.\\
\partial\left(\rho_{j,l}\{c,e\}\right) &= \rho_{j,l}\{a,g\} + \rho_{j,m}\{c,f\} + \rho_{j,l}\cdot\rho_{i,j}\{d,e\}\\
\partial\left(\rho_{i,l}\{d,e\}\right) &= \rho_{j,l}\cdot\rho_{i,j}\{d,e\} + \rho_{i,l}\{a,h\} + \rho_{i,m}\{d,f\} + \rho_{i,l}\cdot\rho_{j,m}\{d,g\}.
\end{aligned}$$

Most of the terms in $\partial^2\{b,e\}$ cancel, but the term $\rho_{i,l}\cdot\rho_{j,m}\{d,g\}$ does not. The offending domain is shaded.

To resolve this difficulty, we impose the relation $\rho_{i,l}\cdot\rho_{j,m}=0$ whenever $i<j<l<m$.

These are essentially all of the cases to check for CPD^-; we will verify this more carefully in Section 6.

Finally, consider the module $CPA^-(\mathcal{H}^A)$. One must check that the relations we imposed on $\mathcal{A}$ are compatible with the action of $\mathcal{A}$ on $CPA^-(\mathcal{H}^A)$; roughly, this follows by rotating the pictures from Figure 5 by 180 degrees. We will discuss this more thoroughly in Section 7.

These are the only relations we will need to impose on the algebra $\mathcal{A}$. It turns out—we will see this next—that this algebra has a clean description in terms of *strand diagrams*.

5. The algebra associated to a slicing

Fix integers $N+1$ and k, representing the height and width respectively of a partial planar grid diagram $\mathcal{H}^A$. We will define an algebra $\mathcal{A}_{N,k}$. We indicated, in a somewhat roundabout manner, generators and relations for $\mathcal{A}_{N,k}$ in Section 4. We start by giving that definition in a more orderly manner and then move on to a description in terms of strand diagrams.

The algebra $\mathcal{A}_{N,k}$ is free as an $\mathbb{A}$-module. For each k-element subset S of $\{0,\dots,N\}$ there is a primitive idempotent I_S, so that

$$I_S\cdot I_T=\begin{cases} I_S & \text{if } S=T,\\ 0 & \text{otherwise.}\end{cases}$$

The algebra $\mathcal{A}_{N,k}$ is generated as an $\mathbb{A}$-algebra by a set of elements $\rho_{S,i,j}$ (together with the idempotents). Here, $0\le i<j\le N$ and S is a k-element subset of $\{0,\dots,N\}$ such that $i\in S$ and $j\notin S$. The relations with the idempotents are as follows:

$$I_T\cdot\rho_{S,i,j}=\begin{cases}\rho_{S,i,j} & \text{if } S=T\\ 0 & \text{otherwise}\end{cases}$$

$$\rho_{S,i,j}\cdot I_T=\begin{cases}\rho_{S,i,j} & \text{if } T=(S\setminus i)\cup j\\ 0 & \text{otherwise.}\end{cases}$$

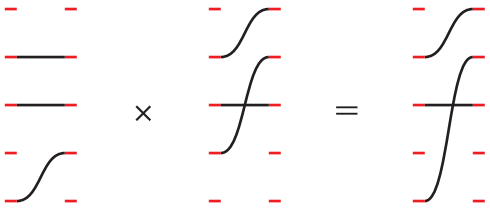


FIGURE 6. **The product on $\mathcal{A}_{4,3}$.** Two examples of upward-veering strand diagrams on 3 strands and 5 positions are shown left and center, and their product on the right.

Set $\rho_{i,j} = \sum_S \rho_{S,i,j}$, so $\rho_{S,i,j} = I_S \rho_{i,j}$. The relations we impose on $\mathcal{A}_{N,k}$ are:

$$\rho_{i,j} \cdot \rho_{l,m} = \rho_{l,m} \cdot \rho_{i,j} \qquad \text{for } j < l \text{ or } i < l < m < j \tag{5.1}$$

$$\rho_{i,j} \cdot \rho_{l,m} = 0 \qquad \text{for } i < l < j < m \tag{5.2}$$

$$\rho_{S,i,j} \cdot \rho_{j,l} = \rho_{S,i,l} \qquad \text{for } j \notin S. \tag{5.3}$$

We also define a differential on $\mathcal{A}_{N,k}$ by setting

$$\partial(\rho_{S,i,j}) = \sum_{\substack{l \in S \\ i<l<j}} \rho_{l,j} \cdot \rho_{i,l}$$

and extending by the Leibniz rule.

Let $\mathcal{I}_{N,k}$ denote the subalgebra of $\mathcal{A}_{N,k}$ generated by the idempotents.

We will check that $\partial^2 = 0$ and that ∂ has a consistent extension to all of $\mathcal{A}_{N,k}$, but first we reinterpret this algebra graphically, and introduce a grading.

Let $kI = \coprod_{i=1}^{k} [0,1]$, $\partial_- kI = \coprod_{i=1}^{k} \{0\}$ and $\partial_+ kI = \coprod_{i=1}^{k} \{1\}$. By an *upward-veering strand diagram on k strands and $N+1$ positions* we mean a class $[\rho]$ of smooth maps

$$\rho \colon (kI, \partial_- kI, \partial_+ kI) \to ([0,1] \times [0,N], \{0\} \times \{0,\dots,N\}, \{1\} \times \{0,\dots,N\})$$

such that $\rho'(t) \geq 0$ for all $t \in kI$, and such that the restrictions $\rho|_{\partial_- kI}$ and $\rho|_{\partial_+ kI}$ are injective, modulo homotopy and reordering of the strands. (See Figure 6 for an illustration.) Let $\mathcal{B}(N,k)$ denote the set of upward-veering strand diagrams on k strands and $N+1$ positions.

Given an element $[\rho] \in \mathcal{B}(N,k)$, let $\mathrm{cr}([\rho])$ denote the minimum number of crossings (double points) of any representative ρ of $[\rho]$.

If $[\rho_1], [\rho_2] \in \mathcal{B}(N,k)$ are such that $\partial_+[\rho_1] = \partial_-[\rho_2]$ then we can concatenate ρ_1 and ρ_2 to obtain a new upward-veering strand diagram $\rho_1\rho_2$. Note that $\mathrm{cr}([\rho_1\rho_2]) \leq \mathrm{cr}([\rho_1]) + \mathrm{cr}([\rho_2])$. Let $\tilde{\mathcal{A}}_{N,k}$ denote the free $\mathbb{A}$-module on $\mathcal{B}(N,k)$, and extend the concatenation operation to a product on $\tilde{\mathcal{A}}_{N,k}$ by setting

$$[\rho_1] \cdot [\rho_2] = \begin{cases} [\rho_1\rho_2] & \text{if } \partial_+[\rho_1] = \partial_-[\rho_2] \text{ and } \mathrm{cr}([\rho_1\rho_2]) = \mathrm{cr}([\rho_1]) + \mathrm{cr}([\rho_2]) \\ 0 & \text{otherwise.} \end{cases}$$

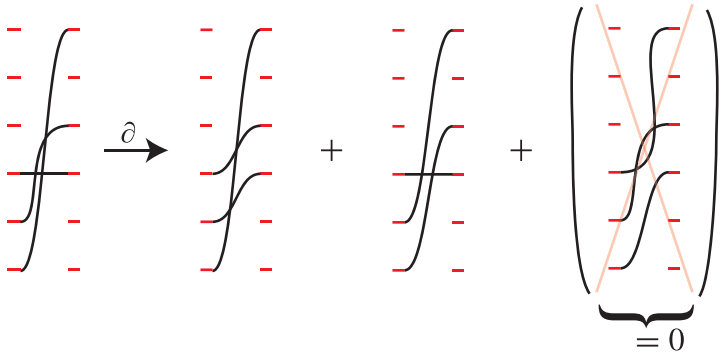


FIGURE 7. **The differential on $\mathcal{A}_{5,3}$.** Note that the term on the far right is not included in the differential of the term on the left because of the condition on the number of crossings cr.

This operation is obviously associative. The idempotents of $\tilde{\mathcal{A}}_{N,k}$ are braids consisting of k horizontal strands, and as such are in bijection with the set of k-element subsets of $\{1,\dots,N\}$.

We define a differential ∂ on $\tilde{\mathcal{A}}_{N,k}$. Given $[\rho] \in \mathcal{B}(N,k)$, with representative ρ, let $\mathrm{smooth}(\rho)$ denote the multiset of strand diagrams obtained by smoothing a single crossing in ρ. Then define

$$\partial[\rho] = \sum_{\substack{\rho' \in \mathrm{smooth}(\rho) \\ \mathrm{cr}([\rho'])=\mathrm{cr}([\rho])-1}} [\rho'].$$

See Figure 7.

Lemma 5.4. *The algebra $\mathcal{A}_{N,k}$ is isomorphic to the algebra $\tilde{\mathcal{A}}_{N,k}$, via an isomorphism identifying the differentials.*

Proof. This is easy to check; see Figure 8 for a convincing illustration that the relations agree. That the differentials agree is similarly straightforward. □

Provisionally, we define a grading on $\mathcal{A}_{N,k}$ by setting $\mathrm{gr}([\rho]) = \mathrm{cr}([\rho])$.

Proposition 5.5. *The algebra $\mathcal{A}_{N,k}$ is a differential graded algebra. That is:*

(1) *The differential satisfies $\partial^2 = 0$.*
(2) *The differential satisfies the Leibniz rule $\partial(ab) = (\partial a)b + a(\partial b)$.*
(3) *Multiplication has degree 0.*
(4) *The differential has degree -1.*

Proof. All four parts are obvious from the description in terms of strand diagrams. □

Remark 5.6. We have given two different definitions of $\mathcal{A}_{N,k}$. We could give a third, closely related to permutations: the algebra is generated over $\mathbb{A}$ by bijective maps $f\colon S \to T$

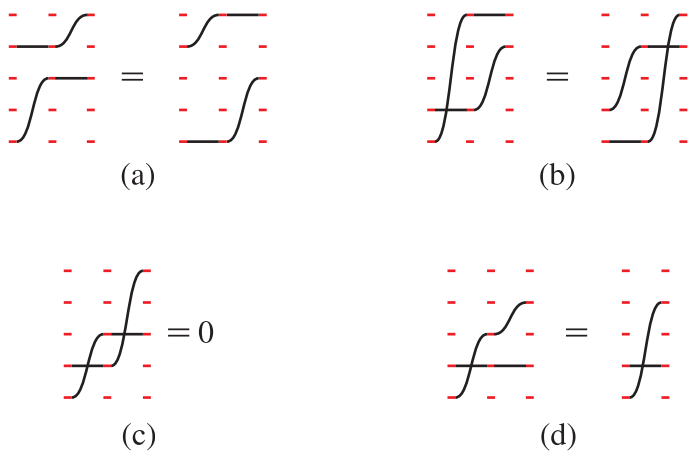


FIGURE 8. **The relations on $\mathcal{A}_{4,2}$.** Parts (a) and (b) correspond to relation (5.1). Part (c) corresponds to relation (5.2). Part (d) corresponds to relation (5.3).

between k-element subsets of $\{0,\dots,N\}$, such that for all $i \in S$, $f(i) \geq i$. The function cr is then the number of inversions of the map (i.e., the number of pairs of integers $i < j$ for which $f(i) > f(j)$), and the multiplication is composition if it is defined and preserves cr and zero otherwise. See [7, Section 3.1.1] for further discussion.

The homological (Maslov) grading we want is not the same as gr. In fact, both the Maslov and Alexander gradings on $\mathcal{A}_{N,k}$ depend not just on N and k but also on which rows contain X's and O's to the left of Z.

More precisely, fix k-element subsets L_X and L_O of $\{1/2,\dots,N-1/2\}$, which are the y-coordinates of the X_i's and O_i's contained in $\mathcal{H}^A$ (including X_k). Given an algebra element a, viewed as a strand diagram, let $L_X(a)$ denote the intersection number of a with the lines $y = \ell$ for $\ell \in L_X$. (Equivalently, define $L_X(\rho_{i,j}) = \#\{\ell \in L_X \mid i < \ell < j\}$ and extend to all of $\mathcal{A}_{N,k}$.) Define $L_O(a)$ similarly.

For $a \in \mathcal{A}_{N,k}$, define gradings A and μ by

$$A(a) = L_X(a) - L_O(a)$$
$$\mu(a) = \mathrm{cr}(a) - 2L_O(a).$$

It is clear that A is preserved by multiplication and the differential, and that multiplication preserves μ while the differential drops μ by 1.

6. The Type D module

Fix a partial planar grid diagram $\mathcal{H}^D$ of height $N+1$ and width $N+1-k$. We will associate to $\mathcal{H}^D$ a differential $\mathcal{A}_{N,k}$-module.

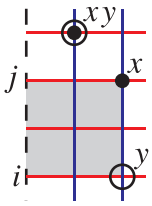


FIGURE 9. **An element (shaded) of** $\mathrm{Half}(\rho_{i,j};\mathbf{x},\mathbf{y})$**.** In fact, the region pictured lies in $\mathrm{Half}^\circ(\rho_{i,j};\mathbf{x},\mathbf{y})$. It is also permitted for there to be some O_i or X_i in the domain (though in the latter case we will not, in fact, count the domain for the theory under discussion).

We define a left action of the idempotents $\mathcal{I}_{N,k}$ on $\mathbb{A}\langle\mathfrak{S}(\mathcal{H}^D)\rangle$, the free $\mathbb{A}$-module generated by the generators in $\mathcal{H}^D$ (see Section 3). Recall that a generator $\mathbf{x}^D\in\mathfrak{S}(\mathcal{H}^D)$ corresponds to an injection $\sigma_{\mathbf{x}}\colon\{k,\dots,N+1\}\to\{0,\dots,N\}$. So, set

$$I_S\mathbf{x}^D=\begin{cases}\mathbf{x}^D & \text{if } S\cap\mathrm{Im}(\sigma_{\mathbf{x}^D})=\emptyset\\ 0 & \text{otherwise.}\end{cases}$$

As an $\mathcal{A}_{N,k}$-module, let

$$CPD^-(\mathcal{H}^D)=\mathcal{A}_{N,k}\otimes_{\mathcal{I}_{N,k}}\mathbb{A}\langle\mathfrak{S}(\mathcal{H}^D)\rangle.$$

That is, the module $CPD^-(\mathcal{H}^D)$ is a direct sum of elementary $\mathcal{A}_{N,k}$-modules, one for each generator in $\mathfrak{S}(\mathcal{H}^D)$.

We next define the differential on $CPD^-(\mathcal{H}^D)$. For generators $\mathbf{x}^D$ and $\mathbf{y}^D$, define $\mathrm{Rect}^\circ(\mathbf{x}^D,\mathbf{y}^D)$ exactly as in Section 2. Given generators $\mathbf{x}^D$, $\mathbf{y}^D$ and a segment $\rho_{i,j}$ in Z, we define a set $\mathrm{Half}(\rho_{i,j};\mathbf{x},\mathbf{y})$, as follows. Define $\mathrm{Half}(\rho_{i,j};\mathbf{x},\mathbf{y})$ to be empty unless $x_i=y_i$ for all but one i. If $x_i=y_i$ for $i\neq j$ and the y-coordinate of x_j is (strictly) greater than the y-coordinate of y_j, then let $\mathrm{Half}(\rho_{i,j};\mathbf{x},\mathbf{y})$ be the singleton set containing the rectangle (or "half-strip") H with upper right corner x_j, and lower right corner y_j, and left edge along the interface Z, where it is the segment from $y=i$ to $y=j$. See Figure 9. Call a half strip $H\in\mathrm{Half}(\rho_{i,j};\mathbf{x}^D,\mathbf{y}^D)$ *empty* if the interior of H is disjoint from $\mathbf{x}^D$ (or equivalently from $\mathbf{y}^D$). Let $\mathrm{Half}^\circ(\rho_{i,j};\mathbf{x}^D,\mathbf{y}^D)$ denote the set of empty half strips in $\mathrm{Half}(\rho_{i,j};\mathbf{x}^D,\mathbf{y}^D)$; this set has at most one element.

Now, for $\mathbf{x}^D$ a generator, define

$$\partial\mathbf{x}^D=\sum_{\mathbf{y}^D\in\mathfrak{S}(\mathcal{H}^D)}\sum_{\substack{R\in\mathrm{Rect}^\circ(\mathbf{x}^D,\mathbf{y}^D)\\ \mathbb{X}(R)=0}}U(R)\cdot\mathbf{y}+\sum_{\mathbf{y}^D}\sum_{\rho_{i,j}}\sum_{\substack{H\in\mathrm{Half}^\circ(\rho_{i,j};\mathbf{x}^D,\mathbf{y}^D)\\ \mathbb{X}(R)=0}}U(H)\cdot\rho_{i,j}\mathbf{y}.$$

We extend the definition via the Leibniz rule to all of $CPD^-(\mathcal{H}^D)$.

Proposition 6.1. *The module (CPD^-,∂) is a differential module. That is, $\partial^2=0$.*

Proof. Since

$$\begin{aligned}\partial^2(a\mathbf{w}^D) &= \partial\left((\partial a)\mathbf{w}^D + a(\partial\mathbf{w}^D)\right)\\ &= \left(\partial^2 a\right)\mathbf{w}^D + 2(\partial a)(\partial\mathbf{w}^D) + a\left(\partial^2\mathbf{w}^D\right)\\ &= a\left(\partial^2\mathbf{w}^D\right),\end{aligned}$$

it suffices to show that the coefficient of $\mathbf{y}^D$ in $\partial^2\mathbf{w}^D$ is zero for any $\mathbf{w}^D, \mathbf{y}^D \in \mathfrak{S}(\mathcal{H}^D)$.

The remainder of the proof is similar to the combinatorial proof in the closed case [9, Proposition 2.8]. Let $a_{\mathbf{w}^D,\mathbf{x}^D}$ denote the coefficient of $\mathbf{x}^D$ in $\partial\mathbf{w}^D$. Then the coefficient of $\mathbf{y}^D$ in $\partial^2\mathbf{w}^D$ is

$$\left(\sum_{\mathbf{x}^D} a_{\mathbf{w}^D,\mathbf{x}^D}\cdot a_{\mathbf{x}^D,\mathbf{y}^D}\right) + \partial a_{\mathbf{w}^D,\mathbf{y}^D}. \tag{6.2}$$

The first term in Formula (6.2) is a sum of terms coming from pairs (A, B) where one of the following cases holds.

(1) $A \in \mathrm{Rect}^\circ(\mathbf{w}^D, \mathbf{x}^D)$ and $B \in \mathrm{Rect}^\circ(\mathbf{x}^D, \mathbf{y}^D)$ (for some $\mathbf{x}^D$). These contributions cancel in pairs exactly as in [9, Proposition 2.8]; see Figure 2.

(2) $A \in \mathrm{Rect}^\circ(\mathbf{w}^D, \mathbf{x}^D)$ and $B \in \mathrm{Half}^\circ(\rho_{i,j}; \mathbf{x}^D, \mathbf{y}^D)$ (for some $\mathbf{x}^D$ and $\rho_{i,j}$). There are several cases here, the most interesting of which is illustrated in Figure 5(c). In this case, the relation $\rho_{S,i,j}\cdot\rho_{j,m} = \rho_{S,i,m}$ implies this term cancels with a pair of half strips (A', B') obtained by cutting the domain horizontally instead of vertically.

(3) $A \in \mathrm{Half}^\circ(\rho_{i,j}; \mathbf{w}^D, \mathbf{x}^D)$ and $B \in \mathrm{Half}^\circ(\rho_{l,m}; \mathbf{x}^D, \mathbf{y}^D)$ (for some $\mathbf{x}^D$, $\rho_{i,j}$, and $\rho_{l,m}$). Again, there are several cases. The two half-strips may be disjoint (Figure 5(a)), or they may form a sideways "T" (Figure 5(b)); in these two cases, relation (5.1) implies the contributions from taking the two strips in the two different orders cancel. The two half-strips may abut top to bottom, in an "L"-shape (Figure 5(c)); this cancels with one of the cases from Item (2).

Another possibility is that the upper right corner of B is the lower right corner of A, as in Figure 5(d). This configuration contributes a coefficient of $\rho_{j,l}\cdot\rho_{i,j}$ (times some U-power). There is also a half-strip, $A\cup B$, which contributes $\rho_{i,l}$ to $\partial\mathbf{w}$; since $\partial\rho_{i,l} = \rho_{j,l}\cdot\rho_{i,j}$ in this case, these terms cancel.

Finally, the half strips may overlap as in Figure 5(e). But in this case the coefficient contributed is $\rho_{i,l}\cdot\rho_{j,m}$ which is 0.

Note that all terms in $\partial a_{\mathbf{w}^D,\mathbf{y}^D}$ cancelled against terms in Part (3). This completes the proof. □

Finally, we turn to the gradings on $CPD^-(\mathcal{H}^D)$. Fix any planar grid diagram $\mathcal{H} = (\mathbb{R}^2, \mathbb{X}, \mathbb{O})$ such that $\mathcal{H}^D = (H^D, \mathbb{X}^D, \mathbb{O}^D)$ can be obtained by cutting $\mathcal{H}$. Then, for a generator $\mathbf{x}^D \in \mathfrak{S}(\mathcal{H}^D)$, there are numbers $\mathcal{I}(\mathbb{X}, \mathbf{x}^D)$ and $\mathcal{I}(\mathbb{O}, \mathbf{x}^D)$, as in Section 2. These numbers obviously do not depend on the choice of $\mathcal{H}$. Further, fix any generator $\mathbf{x} \in \mathfrak{S}(\mathcal{H})$ extending $\mathbf{x}^D$. Then we have a number $\mathcal{I}(\mathbf{x}, \mathbf{x}^D)$, which again does not depend

on the choice of $\mathcal{H}$ or $\mathbf{x}$. Now, define the gradings of $\mathbf{x}^D$ by

$$A(\mathbf{x}^D) = \mathcal{I}(\mathbb{X}, \mathbf{x}^D) - \mathcal{I}(\mathbb{O}, \mathbf{x}^D)$$
$$\mu(\mathbf{x}^D) = \mathcal{I}(\mathbf{x}, \mathbf{x}^D) - 2\mathcal{I}(\mathbb{O}, \mathbf{x}^D).$$

Extend these definitions to all of $CPD^-(\mathcal{H}^D)$ by setting $A(a\mathbf{x}^D) = A(a) + A(\mathbf{x}^D)$ and $\mu(a\mathbf{x}^D) = \mu(a) + \mu(\mathbf{x}^D)$ for $a \in \mathcal{A}_{N,k}$.

Proposition 6.3. *The gradings A and μ make $CPD^-(\mathcal{H}^D)$ into a graded module over $\mathcal{A}_{N,k}$. The differential ∂ on $CPD^-(\mathcal{H}^D)$ drops μ by 1 while preserving A.*

(When assigning gradings to the algebra, we let L_X denote the set of $i - 1/2$ which are *not* y-coordinates of points in $\mathbb{X}^D$, and similarly for L_O.)

Proof. The first statement is trivial. To verify that the differential drops μ by 1, write $\mathbf{x} = (\mathbf{x}^A, \mathbf{x}^D)$. Suppose that $(\prod_\ell U_\ell^{n_\ell})\, \rho_{i,j} \cdot \mathbf{y}^D$ occurs in $\partial \mathbf{x}^D$. Then

$$\mathcal{I}(\mathbf{x}, \mathbf{x}^D) - \mathcal{I}(\mathbf{y}, \mathbf{y}^D) = 1 + \#\{(r,s) \in \mathbf{x}^A \mid i < s < j\}.$$

This is exactly $1 + \mathrm{cr}(\rho_{S,i,j})$, where $S = \{0, \dots, N\} \setminus \mathrm{Im}(\sigma_{\mathbf{x}^D})$. Also,

$$\mathcal{I}(\mathbb{O}, \mathbf{x}^D) - \mathcal{I}(\mathbb{O}, \mathbf{y}^D) = \left(\sum_\ell n_\ell\right) + L_O(\rho_{i,j}).$$

This implies that the differential decreases μ by 1, as desired. That the differential preserves A is similar but easier. □

7. The Type A module

The module CPA^- is much smaller than CPD^-. Fix a partial planar grid diagram $\mathcal{H}^A$ with width k and height $N+1$. The module $CPA^-(\mathcal{H}^A)$ is freely generated over $\mathbb{A}$ by $\mathfrak{S}(\mathcal{H}^A)$. There is a differential ∂ on $CPA^-(\mathcal{H}^A)$ defined by

$$\partial \mathbf{x}^A = \sum_{\mathbf{y}^A \in \mathfrak{S}(\mathcal{H}^A)} \sum_{\substack{R \in \mathrm{Rect}^\circ(\mathbf{x}^A, \mathbf{y}^A) \\ \mathbb{X}(R) = 0}} U(R) \cdot \mathbf{y}^A.$$

It remains to define an action of $\mathcal{A}_{N,k}$ on $CPA^-(\mathcal{H}^A)$.

Given a generator $\mathbf{x}^A \in \mathfrak{S}(\mathcal{H}^A)$, let $\sigma_{\mathbf{x}^A}$ denote the corresponding map $\{0, \dots, k-1\} \to \{0, \dots, N\}$. We define an action of the idempotents $\mathcal{I}_{N,k}$ by

$$\mathbf{x}^A I_S = \begin{cases} \mathbf{x}^A & \text{if } S = \mathrm{Im}(\sigma_{\mathbf{x}^A}) \\ 0 & \text{otherwise.} \end{cases}$$

This is, in some sense, exactly the opposite of the action of the idempotents on CPD^-.

Given generators $\mathbf{x}^A$ and $\mathbf{y}^A$ in $\mathfrak{S}(\mathcal{H}^A)$ and a generator $\rho_{i,j}$ of $\mathcal{A}_{N,k}$ (which we view as a chord in Z from $y = i$ to $y = j$) define $\mathrm{Half}(\mathbf{x}, \mathbf{y}; \rho_{i,j})$ to be empty unless $x_k = y_k$ for all but one k, and in this case let it be the singleton set containing the rectangle (or "half-strip") H with lower left corner x_k and upper left corner y_k, and right edge

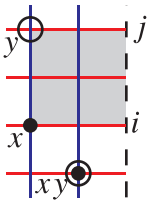


FIGURE 10. **An element of** $\mathrm{Half}(\mathbf{x},\mathbf{y};\rho_{i,j})$**.** The definition is essentially the same as the definition for CPD^-, only rotated by 180 degrees.

$\rho_{i,j}$ if such a rectangle exists, and empty otherwise. See Figure 10. Call a half strip $H \in \mathrm{Half}(\mathbf{x}^A,\mathbf{y}^A;\rho_{i,j})$ *empty* if the interior of H is disjoint from $\mathbf{x}^A$. Let $\mathrm{Half}^\circ(\mathbf{x}^A,\mathbf{y}^A;\rho_{i,j})$ denote the set of empty half strips in $\mathrm{Half}(\mathbf{x}^A,\mathbf{y}^A;\rho_{i,j})$.

We define an action by the generators $\rho_{i,j}$ of $\mathcal{A}_{N,k}$ by

$$\mathbf{x}^A\rho_{i,j} = \sum_{\substack{\mathbf{y}^A\in\mathfrak{S}(\Sigma)\\ H\in\mathrm{Half}^\circ(\mathbf{x}^A,\mathbf{y}^A;\rho_{i,j})\\ \mathbb{X}(H)=0}} U(H)\cdot\mathbf{y}^A.$$

(The sum contains at most one term.)

Proposition 7.1. *The module $CPA^-(\mathcal{H}^A)$ is a differential $\mathcal{A}_{N,k}$-module. That is:*

(1) *The action of the $\rho_{i,j}$ defined above respects the relations in $\mathcal{A}_{N,k}$.*
(2) *The action satisfies the Leibniz rule.*
(3) *The differential ∂ satisfies $\partial^2 = 0$.*

Proof. (The reader may wish to compare this with the proof of Proposition 6.1: the pictures are almost the same, but their interpretations are different.)

That the $\mathcal{A}_{N,k}$-action respects the three relations (5.1), (5.2) and (5.3) follow from the cases illustrated in Figure 11. In parts (a) and (b), we have

$$(\{a,c\}\rho_{i,j})\,\rho_{l,m} = (\{a,c\}\rho_{l,m})\,\rho_{i,j} = \{b,d\}.$$

so relation (5.1) is respected. (We suppress the U-powers, but since these depend only on the domains they, too, agree.)

In part (c) of Figure 11,

$$(\{a\}\rho_{i,j})\rho_{j,l} = \{b\}\rho_{j,l} = \{c\} = \{a\}\rho_{i,l},$$

so relation (5.2) is respected.

In part (d) of Figure 11 we have

$$(\{a,f\}\rho_{i,l})\,\rho_{j,m} = \{c,f\}\rho_{j,m} = 0$$

since the corresponding half-strip is not empty. So, relation (5.3) is respected. (This is only one of the two pictures we need to check in this case, but the other is similar.)

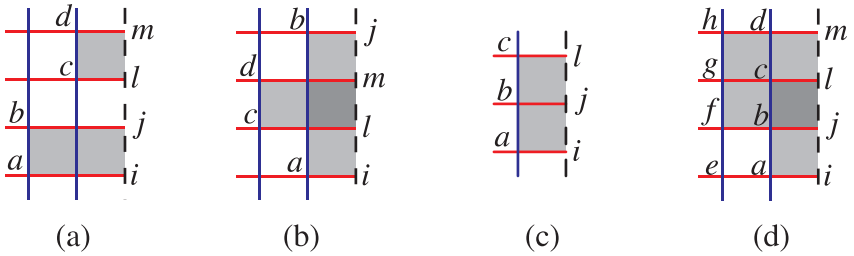


FIGURE 11. **The $\mathcal{A}_{N,k}$-action on CPA^- respects the relations on the algebra.** Parts (a) and (b) correspond to relation (5.1). Part (c) corresponds to relation (5.2). Part (d) corresponds to relation (5.3).

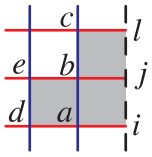


FIGURE 12. **The Leibniz rule for CPA^-.** The domain shown can be decomposed in two ways: as a rectangle followed by a half-strip, or as two half strips. These correspond to $(\partial\{b,d\})\,\rho_{i,l}$ and $\{b,d\}\,(\partial\rho_{i,l})$ respectively.

This proves Part (1).

Part (2) follows from Figure 12. More precisely, it suffices to show that for any i, j,

$$\partial\left(\mathbf{x}^A\rho_{i,j}\right) = \left(\partial\mathbf{x}^A\right)\rho_{i,j} + \mathbf{x}^A\left(\partial\rho_{i,j}\right).$$

Both $\partial\left(\mathbf{x}^A\rho_{i,j}\right)$ and $\left(\partial\mathbf{x}^A\right)\rho_{i,j}$ correspond to a domain which is a union of a rectangle and a half-strip. The most interesting case is when these abut to form an "L"-shape, as in Figure 12. There, for $\mathbf{x}^A = \{b,d\}$ we have

$$\begin{aligned}\partial\{b,d\} &= \{a,e\}\\ \{a,e\}\rho_{i,l} &= \{c,e\}\\ \{b,d\}\rho_{j,l}\rho_{i,j} &= \{c,e\}\\ \{b,d\}\rho_{i,l} &= 0,\end{aligned}$$

so

$$\partial\left(\{b,d\}\rho_{i,l}\right) = 0 = \left(\partial\{b,d\}\right)\rho_{i,l} + \{b,d\}\left(\partial\rho_{i,l}\right).$$

(The other interesting but similar case is obtained by flipping Figure 12 vertically.) This proves Part 2.

Part (3) follows from the same argument as in the closed case [9, Proposition 2.8]; see also Figure 2. □

Finally, we turn to the gradings on $CPA^-(\mathcal{H}^A)$. Define

$$A(\mathbf{x}^A) = \mathcal{I}(\mathbb{X}^A, \mathbf{x}^A) - \mathcal{I}(\mathbb{O}^A, \mathbf{x}^A)$$
$$\mu(\mathbf{x}^D) = \mathcal{I}(\mathbf{x}^A, \mathbf{x}^A) - 2\mathcal{I}(\mathbb{O}^A, \mathbf{x}^A).$$

Proposition 7.2. *These gradings make $CPA^-(\mathcal{H}^A)$ into a graded $\mathcal{A}_{N,k}$-module. The differential on $CPA^-(\mathcal{H}^A)$ preserves the Alexander grading A and drops the Maslov grading μ by 1.*

(When assigning gradings to the algebra, we let L_X denote the set of $i - 1/2$ which are y-coordinates of points in $\mathbb{X}^A$, and similarly for L_O.)

Proof. First we check that multiplication preserves the A grading. Suppose that $\mathbf{x}^A \rho_{i,j} = (\prod_\ell U_\ell^{n_\ell})\, \mathbf{y}^A$. Then

$$\mathcal{I}(\mathbb{X}^A, \mathbf{x}^A) = \mathcal{I}(\mathbb{X}^A, \mathbf{y}^A) - L_X(\rho_{i,j})$$
$$\mathcal{I}(\mathbb{O}^A, \mathbf{x}^A) = \mathcal{I}(\mathbb{O}^A, \mathbf{y}^A) - L_O(\rho_{i,j}) + \sum_\ell n_\ell.$$

The result follows.

That multiplication preserves μ is similar; see also the proof of Proposition 6.3 That the differential preserves A and drops μ by 1 is straightforward. □

Remark 7.3. The definition of CPA^- is somewhat different in spirit from the definition of $\widehat{CFA}$ for bordered three-manifolds in [7, Section 7]: there the product $\mathbf{x}^A a$ is defined directly for any algebra element a. In our setting, we could do this by counting more complicated domains than rectangles.

8. The pairing theorem

Theorem 2. *Let $\mathcal{H}$ be a planar grid diagram, decomposed as $\mathcal{H}^A \cup_Z \mathcal{H}^D$, where $\mathcal{H}^A$ (respectively $\mathcal{H}^D$) is a partial planar grid diagram with width k (respectively $N + 1 - k$) and height $N + 1$. Then*

$$CP^-(\mathcal{H}) \cong CPA^-(\mathcal{H}^A) \otimes_{\mathcal{A}_{N,k}} CPD^-(\mathcal{H}^D),$$

as $(\mathbb{Z} \oplus \mathbb{Z})$-graded chain complexes over $\mathbb{A}$.

Proof. There is an obvious identification between the generators of $CP^-(\mathcal{H})$ and the generators of $CPA^-(\mathcal{H}^A) \otimes_{\mathcal{A}_{N,k}} CPD^-(\mathcal{H}^D)$. It follows from their definitions that this identification respects the A and μ gradings.

The rest of the proof is essentially trivial, so we write it with formulas to make it seem complicated. Given a generator $\mathbf{x}$ of $CP^-(\mathcal{H})$, we split $\partial\mathbf{x}$ into three pieces, according to

whether the domain rectangle is entirely to the left of the dividing line $\{x = k - 1/4\}$, crosses the dividing line, or is entirely to the right of the dividing line:

$$\partial \mathbf{x} = \partial_L \mathbf{x} + \partial_M \mathbf{x} + \partial_R \mathbf{x}.$$

Then if $\mathbf{x}$ is identified with $\mathbf{x}^A \otimes \mathbf{x}^D$, we have

$$\begin{aligned}
\partial(\mathbf{x}^A \otimes \mathbf{x}^D) &= (\partial \mathbf{x}^A) \otimes \mathbf{x}^D + \mathbf{x}^A \otimes (\partial \mathbf{x}^D) \\
&= \partial_L \mathbf{x} + \partial_R \mathbf{x} + \sum_{\mathbf{y}^D \in \mathfrak{S}(\mathcal{H}^D)} \sum_{\substack{H \in \mathrm{Half}^\circ(\rho_{i,j}; \mathbf{x}^D, \mathbf{y}^D) \\ \mathbb{X}(H) = 0}} U(H)(\mathbf{x}^A \otimes \rho_{i,j} \mathbf{y}^D) \\
&= \partial_L \mathbf{x} + \partial_R \mathbf{x} + \sum_{\mathbf{y}^D \in \mathfrak{S}(\mathcal{H}^D)} \sum_{\substack{H \in \mathrm{Half}^\circ(\rho_{i,j}; \mathbf{x}^D, \mathbf{y}^D) \\ \mathbb{X}(H) = 0}} U(H)(\mathbf{x}^A \rho_{i,j} \otimes \mathbf{y}^D) \\
&= \partial_L \mathbf{x} + \partial_R \mathbf{x} \\
&\quad + \sum_{\substack{\mathbf{y}^D \in \mathfrak{S}(\mathcal{H}^D) \\ \mathbf{y}^A \in \mathfrak{S}(\mathcal{H}^A) \\ \rho_{i,j}}} \sum_{\substack{H \in \mathrm{Half}^\circ(\rho_{i,j}; \mathbf{x}^D, \mathbf{y}^D) \\ \mathbb{X}(H) = 0}} \sum_{\substack{H' \in \mathrm{Half}^\circ(\mathbf{x}^A, \mathbf{y}^A; \rho_{i,j}) \\ \mathbb{X}(H') = 0}} U(H' \cup H)(\mathbf{y}^A \otimes \mathbf{y}^D) \\
&= \partial_L \mathbf{x} + \partial_R \mathbf{x} + \partial_M \mathbf{x} \\
&= \partial \mathbf{x},
\end{aligned}$$

as desired. □

Remark 8.1. More useful is the fact that $CP^-(\mathcal{H})$ is quasi-isomorphic to the derived tensor product $CPA^-(\mathcal{H}^A) \widetilde{\otimes}_{\mathcal{A}_{N,k}} CPD^-(\mathcal{H}^D)$. For instance, this allows one to simplify the complexes CPA^- and CPD^- more dramatically before taking the tensor product. In fact, the $\mathcal{A}_{N,k}$-module $CPD^-(\mathcal{H}^D)$ is projective (hence flat), as one can see by imitating an argument from Bernstein and Lunts [1, Proposition 10.12.2.6]. It follows that the derived tensor product agrees with the ordinary one.

9. Bimodules

At this point we have encountered left and right modules over $\mathcal{A}_{N,k}$. We will now see that bimodules also have several important roles to play. (The material in this section is analogous to material in [6].)

9.1. Freezing

We have studied how to take a planar grid diagram and make a single vertical cut. In the spirit of factoring a braid into generators, however, we might want to make several different vertical cuts. In this section we will see that the correct objects to assign to slices in the middle are $(\mathcal{A}_{N,k}, \mathcal{A}_{N,l})$-bimodules.

That is, consider the result of slicing a planar grid diagram $\mathcal{H}$ along the lines $Z_1 = \{x = k - 1/4\}$ and $Z_2 = \{x = l - 1/4\}$ (with $l > k$). The result is two partial planar grid diagrams $\mathcal{H}^A = \mathcal{H} \cap \{x < k - 1/4\}$ and $\mathcal{H}^D = \mathcal{H} \cap \{x > l - 1/4\}$, and a

middle partial planar grid diagram $\mathcal{H}^{DA} = \mathcal{H} \cap \{k-1/4 < x < l-1/4\}$. We will associate an $(\mathcal{A}_{N,k}, \mathcal{A}_{N,l})$-bimodule $CPDA^-(\mathcal{H}^{DA})$ to $\mathcal{H}^{DA}$.

A generator for $\mathcal{H}^{DA}$ is an $(l-k)$-tuple of points $\mathbf{x} = \{x_i\}_{i=k}^{l-1}$; a generator $\mathbf{x}$ corresponds to an injection $\sigma_{\mathbf{x}} \colon \{k, \dots, l-1\} \to \{1, \dots, N\}$. (For consistency with earlier notation, we should really write $\mathbf{x}$ as $\mathbf{x}^{DA}$, but the notation becomes too cumbersome.) Let $\mathfrak{S}(\mathcal{H}^{DA})$ denote the set of generators for $\mathcal{H}^{DA}$. Call a generator $\mathbf{x}$ *compatible* with an idempotent $I_S \in \mathcal{A}_{N,k}$ if $\mathrm{Im}(\sigma_{\mathbf{x}}) \cap S = \emptyset$. As a left module, $CPDA^-(\mathcal{H}^{DA})$ is a direct sum of elementary modules,

$$CPDA^-(\mathcal{H}^{DA}) = \bigoplus_{\substack{\mathbf{x} \in \mathfrak{S}(\mathcal{H}^{DA}) \\ S \text{ compatible with } \mathbf{x}}} \mathcal{A}_{N,k} I_S.$$

We will write the generator of the summand $\mathcal{A}_{N,k} I_S$ coming from $\mathbf{x}$ as $I_S \mathbf{x}$. Note that, unlike for CPD^- or CPA^-, the generator $\mathbf{x}$ does not determine the idempotent S.

We next define a differential on $CPDA^-(\mathcal{H}^{DA})$. Given generators $\mathbf{x}, \mathbf{y} \in \mathfrak{S}(\mathcal{H}^{DA})$ such that $x_n = y_n$ for $n \neq m$ (for some m), and $i < j \in \{0, \dots, N\}$, define $\mathrm{Half}(\rho_{i,j}; \mathbf{x}, \mathbf{y})$ to be the set of rectangles with upper right corner at x_m, lower right corner at y_m and left edge the segment $\rho_{i,j}$ in Z_1 from $(k-1/4, i)$ to $(k-1/4, j)$. Define $\mathrm{Half}^\circ(\rho_{i,j}; \mathbf{x}, \mathbf{y})$ to be the subset of $\mathrm{Half}(\rho_{i,j}; \mathbf{x}, \mathbf{y})$ consisting of empty half-strips, i.e., half strips not containing any element of $\mathbf{x}$ in their interiors. Then set

$$\partial(I_S \mathbf{x}) = \sum_{\mathbf{y} \in \mathfrak{S}(\mathcal{H}^{DA})} \sum_{\substack{R \in \mathrm{Rect}^\circ(\mathbf{x}, \mathbf{y}) \\ \mathbb{X}(R) = 0}} U(R) \cdot I_S \mathbf{y} + \sum_{\substack{\mathbf{y} \in \mathfrak{S}(\mathcal{H}^{DA}) \\ i < j \in \{0, \dots, N\}}} \sum_{\substack{H \in \mathrm{Half}^\circ(\rho_{i,j}; \mathbf{x}, \mathbf{y}) \\ \mathbb{X}(H) = 0}} U(H) \cdot I_S \rho_{i,j} \mathbf{y}.$$

Here, the notation $I_S \rho_{i,j} \mathbf{y}$, though suggestive, should be explained. If $i \in S$ and $j \notin S$ then $I_S \rho_{i,j} \mathbf{y}$ denotes $\rho_{i,j} I_T$, where $T = (S \setminus i) \cup j$, if T is compatible with $\mathbf{y}$. Otherwise (i.e., if $i \notin S$, $j \in S$, or T is not compatible with $\mathbf{y}$) we declare $I_S \rho_{i,j} \mathbf{y}$ to be 0.

Finally, we define the right module structure on $CPDA^-(\mathcal{H}^{DA})$. Given a primitive idempotent $I_T \in \mathcal{I}_{N,l}$, define

$$(I_S \mathbf{x}) I_T = \begin{cases} I_S \mathbf{x} & \text{if } (S \cup \mathrm{Im}(\sigma_{\mathbf{x}})) \cap T = \emptyset \\ 0 & \text{otherwise.} \end{cases}$$

Given generators $\mathbf{x}, \mathbf{y} \in \mathfrak{S}(\mathcal{H}^{DA})$ such that $x_\ell = y_\ell$ for $\ell \neq m$ (for some m), and $i < j \in \{0, \dots, N\}$, define $\mathrm{Half}(\mathbf{x}, \mathbf{y}; \rho_{i,j})$ to be the set of rectangles with lower left corner at x_m, upper left corner at y_m and right edge the segment $\rho_{i,j}$ in Z_2 from $(l-1/4, i)$ to $(l-1/4, j)$. Define $\mathrm{Half}^\circ(\mathbf{x}, \mathbf{y}; \rho_{i,j})$ to be the subset of $\mathrm{Half}(\mathbf{x}, \mathbf{y}; \rho_{i,j})$ consisting of empty half-strips, i.e., half strips not containing any element of $\mathbf{x}$ in their interiors.

Given a chord $\rho_{i,j}$ in Z_2 from $(l-1/4, i)$ to $(l-1/4, j)$, define $\mathrm{Strip}(\rho_{i,j})$ to be the horizontal strip with right edge $\rho_{i,j} \subset Z_2$ and left edge $\rho_{i,j} \subset Z_1$. Given $\rho_{i,j}$ and a generator $\mathbf{x} \in \mathfrak{S}(\mathcal{H}^{DA})$, define $\mathrm{Strip}^\circ(\mathbf{x}; \rho_{i,j})$ to be the empty set if $\mathrm{Strip}(\rho_{i,j})$ contains a point in $\mathbf{x}$ (even along its boundary) and the singleton set $\mathrm{Strip}(\rho_{i,j})$ if $\mathrm{Strip}(\rho_{i,j})$ does not contain a point in $\mathbf{x}$.

At last, define

$$(I_S\mathbf{x})\rho_{i,j} = \sum_{\substack{E\in \mathrm{Strip}^\circ(\mathbf{x},\rho_{i,j})\\ \mathbb{X}(E)=0}} U(E)\cdot I_S\rho_{i,j}\mathbf{x} + \sum_{\mathbf{y}\in\mathfrak{S}(\mathcal{H}^{DA})}\sum_{\substack{H\in \mathrm{Half}^\circ(\mathbf{x},\mathbf{y};\rho_{i,j})\\ \mathbb{X}(H)=0}} U(H)\cdot I_S\mathbf{y}.$$

These definitions are, in fact, compatible:

Proposition 9.1. *The module $CPDA^-(\mathcal{H}^{DA})$ is a differential $(\mathcal{A}_{N,k},\mathcal{A}_{N,l})$-bimodule.*

We leave the proof to the interested reader.

As on CPD^-, the grading of a generator $I_S\mathbf{x}$ of $CPDA^-(\mathcal{H}^{DA})$ is given by

$$A(I_S\mathbf{x}) = \mathcal{I}(\mathbb{X},\mathbf{x}) - \mathcal{I}(\mathbb{O},\mathbf{x})$$
$$\mu(I_S\mathbf{x}) = \mathcal{I}(\bar{\mathbf{x}},\mathbf{x}) - 2\mathcal{I}(\mathbb{O},\mathbf{x}),$$

where $\mathcal{H}=(\mathbb{R}^2,\mathbb{X},\mathbb{O})$ is any planar diagram completing $\mathcal{H}^{DA}$, and $\bar{\mathbf{x}}$ is a generator in $\mathfrak{S}(\mathcal{H})$ completing $\mathbf{x}$ and compatible with the idempotent I_S in the obvious sense.

Finally, the module $CPDA^-(\mathcal{H}^{DA})$ satisfies a pairing theorem:

Proposition 9.2. *With notation as above,*

$$CPA^-(\mathcal{H}^A\cup_{Z_1}\mathcal{H}^{DA}) = CPA^-(\mathcal{H}^A)\otimes_{\mathcal{A}_{N,k}} CPDA^-(\mathcal{H}^{DA})$$
$$CPD^-(\mathcal{H}^{DA}\cup_{Z_2}\mathcal{H}^D) = CPDA^-(\mathcal{H}^{DA})\otimes_{\mathcal{A}_{N,l}} CPD^-(\mathcal{H}^D)$$
$$CP^-(\mathcal{H}) = CPA^-(\mathcal{H}^A)\otimes_{\mathcal{A}_{N,k}} CPDA^-(\mathcal{H}^{DA})\otimes_{\mathcal{A}_{N,l}} CPD^-(\mathcal{H}^D).$$

The proof is obvious. The analogous result for cutting along more than two vertical lines is also true.

Remark 9.3. The notation $CPDA^-$ denotes that the module is "Type D" from the left and "Type A" from the right.

9.2. Type A to Type D

The reader might wonder about the relation between CPA^- and CPD^-. One might expect that they are, in some appropriate sense, dual to each other. In the case of bordered Heegaard Floer homology this is true. In this section, we hint at that story by reconstructing CPD^- from CPA^-. In Section 9.3 we will discuss going the other direction. We will suppress both the gradings and the U-variables: our treatment of both has been too naïve to extend properly to the present discussion.

Let $k' = N+1-k$. We construct a $(\mathcal{A}_{N,k},\mathcal{A}_{N,k'})$-bimodule $CPDD^-_{N,k}$ so that $CPD^-(\mathcal{H}^D) = CPA^-(\mathcal{H}^D)\otimes_{\mathcal{A}_{N,k}} CPDD^-_{N,k}$. Actually, unlike a traditional bimodule with a left action and a right action, we will construct the $\mathcal{A}_{N,k}$- and $\mathcal{A}_{N,k'}$-actions as a pair of commuting *left* actions, so the module $CPA^-(\mathcal{H}^D)\otimes_{\mathcal{A}_{N,k}} CPDD^-_{N,k}$ comes equipped with a left action rather than a right action.

The module $CPDD^-_{N,k}$ is easy to describe. Note that there is an obvious isomorphism $\mathcal{I}_{N,k}\to\mathcal{I}_{N,k'}$, taking I_S to $I_{\{0,\dots,N\}\setminus S}$. This makes $\mathcal{A}_{N,k'}$ into a right $\mathcal{I}_{N,k}$-module. The

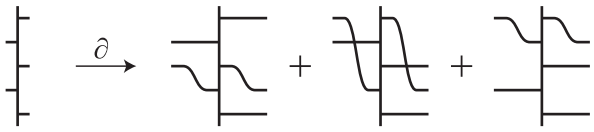


FIGURE 13. **A graphical representation of $CPDD^-_{N,k}$.** The case shown is $N = 4$, $k = 2$. The element I_S, for $S = \{1, 3\}$, is shown on the left. On the right is the differential of $I_S \otimes 1$. This graphical representation treats $CPDD^-_{N,k}$ as a traditional (left,right) bimodule, rather than a (left,left) bimodule; this is the reason that the strands on the right are downward-veering.

module $CPDD^-_{N,k}$ is just

$$\mathcal{A}_{N,k'} \otimes_{\mathcal{I}_{N,k}} \mathcal{A}_{N,k'}$$

where the tensor product identifies the *right* actions of $\mathcal{I}_{N,k}$ on $\mathcal{A}_{N,k'}$ and $\mathcal{A}_{N,k'}$. This module, then, is equipped with two left actions. The differential on $CPDD^-_{N,k}$ is not the one inherited from the tensor product. Rather, for S a k-element subset of $\{0, \dots, N\}$ we define

$$\partial(I_S \otimes 1) = \sum_{\substack{i \in S \\ j > i \\ j \notin S}} \rho^k_{i,j} \rho^{k'}_{i,j} (I_T \otimes 1)$$

where $\rho^k_{i,j}$ denotes the element $\rho_{i,j}$ of $\mathcal{A}_{N,k}$, $\rho^{k'}_{i,j}$ denotes the element $\rho_{i,j}$ of $\mathcal{A}_{N,k'}$, and $T = (S \setminus i) \cup j$. We extend the differential to all of $CPDD^-_{N,k}$ by the Leibniz rule. An example is illustrated in Figure 13.

Lemma 9.4. *The module $CPDD^-_{N,k}$ is a differential $(\mathcal{A}_{N,k}, \mathcal{A}_{N,k'})$-bimodule.*

Proof. This is immediate from the definitions. □

One can view the module $CPDD^-_{N,k}$ as the (Type D, Type D) module associated to a middle partial planar grid diagram with zero β-lines (i.e., in the notation of Section 9.1, $k = l$). The generator corresponds to the empty set in $\boldsymbol{\alpha} \cap \boldsymbol{\beta}$. The differential comes from the strips $\mathrm{Strip}(\rho_{i,j})$ (as in Section 9.1).

As promised, we have the following pairing theorem:

Proposition 9.5. *Fix a partial Heegaard diagram $\mathcal{H}^D$. Then*

$$CPD^-(\mathcal{H}^D) = CPA^-(\mathcal{H}^D) \otimes_{\mathcal{A}_{N,k}} CPDD^-_{N,k}.$$

Proof. The tensor product $CPA^-(\mathcal{H}^D) \otimes_{\mathcal{A}_{N,k}} CPDD^-_{N,k}$ is a direct sum of elementary modules $\mathbf{x}^A \otimes \mathcal{A}_{N,k'} I_S$, one for each generator $\mathbf{x}^A$ of $CPA^-(\mathcal{H}^D)$, where S is $\{0, \dots, N\} \setminus \mathrm{Im}(\sigma_{\mathbf{x}^A})$. The part of the differential on the tensor product coming from the differential on $CPA^-(\mathcal{H}^D)$ counts empty rectangles. The part of the differential coming from the differential on $CPDD^-_{N,k}$ counts empty half strips, exactly as on $CPD^-(\mathcal{H}^D)$. □

9.3. Remarks on Type D to Type A

Turning the module CPD^- into CPA^- is more subtle than turning CPA^- into CPD^-. It is clearly not possible to find a module $CPAA^-_{N,k}$ so that CPA^- is exactly equal to $CPAA^-_{N,k} \otimes_{\mathcal{A}_{N,k'}} CPD^-$: the ranks of these modules over $\mathbb{A}$ prevent this.

There are two approaches one might take. One approach is to use properties of $CPDD^-_{N,k}$ to prove that it induces an equivalence of categories $\mathcal{D}^b(\mathcal{A}_{N,k} - \mathsf{Mod}) \to \mathcal{D}^b(\mathcal{A}_{N,k'} - \mathsf{Mod})$, and then construct a bimodule giving the inverse equivalence of categories.

Another approach is to define $CPAA^-_{N,k}$ as an $\mathcal{A}_\infty$-bimodule. Using the appropriate model for the $\mathcal{A}_\infty$-tensor product (see, e.g., [7, Section 2]), it is then possible for CPA^- to be exactly $CPAA^-_{N,k} \otimes_{\mathcal{A}_{N,k'}} CPD^-$. The generators and first few $\mathcal{A}_\infty$-operations for this $CPAA^-_{N,k}$ are easy to guess. As an $\mathbb{A}$-module, $CPAA^-_{N,k}$ would be just $\mathcal{I}_{N,k} \cong \mathcal{I}_{N,k'}$. The first few $\mathcal{A}_\infty$-relations would be

$$\begin{aligned} m_1(I_S) &= 0 \\ m_2(a, I_S) &= 0 \\ m_2(I_S, a) &= 0 \\ m_3(\rho_{i,j}, I_S, \rho_{i,j}) &= \begin{cases} I_T & \text{if } i \in S,\ j \notin S, \text{ and where } T = (S \setminus i) \cup j \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

(Even though $CPAA^-_{N,k}$ should really have two right actions, for clarity we have written it with one right and one left action.)

Unfortunately, higher $\mathcal{A}_\infty$-relations are harder to guess and, at least in the case of bordered Heegaard Floer homology, depend on some choices. Fortunately, in the case of bordered Heegaard Floer homology, these modules are induced by counts of holomorphic curves, so we need not build them by hand; see [7]. (In particular, it turns out that the choices are induced by a choice of almost complex structure.) The challenge in defining $CPAA^-_{N,k}$, then, becomes counting holomorphic curves.

10. How the real world is harder

In this section, we preview the difficulties involved in using the ideas from this paper to define more useful invariants.

10.1. Complications for $\widehat{HF}$ of 3-manifolds

As discussed in the introduction, applying the ideas of this paper to the case of the Heegaard Floer group $\widehat{HF}(Y)$ gives an invariant of 3-manifolds with boundary; see [7]. The main complications are as follows.

10.1.1. Heegaard diagrams.

Instead of working with grid diagrams, the invariant $\widehat{HF}(Y^3)$ is defined by using a "Heegaard diagram" for Y. One needs, then, an appropriate family of partial Heegaard diagrams. Such a class, called either "Heegaard diagrams with boundary" or "bordered Heegaard diagrams" was presented in [5]; see also [7, Section 4]. These diagrams are induced by a self-indexing Morse function f on a three manifold with boundary $(Y, \partial Y)$ such that ∇f is tangent to ∂Y (and subject to a few more constraints). Bordered Heegaard diagrams specify not just the three-manifold Y but also a parametrization of ∂Y; this is obviously needed for the pairing theorem to make sense.

One incidental effect is that the algebra $\mathcal{A}_{N,k}$ needs to be modified somewhat. In the planar setting, each α-line intersects the interface Z in a single point; in the bordered case (or the toroidal case) this is not true. The solution in the bordered case is to work with a subalgebra of $\mathcal{A}_{N,k}$ which, roughly, remembers how the points $\boldsymbol{\alpha} \cap Z$ are paired-up. (In the toroidal case described below, it is more convenient to remember only half of the points and drop the requirement that strand diagrams be upward-veering.)

10.1.2. Holomorphic curves.

Like the closed Heegaard Floer invariant $\widehat{CF}(Y)$, the definitions of the bordered Heegaard Floer invariants $\widehat{CFA}(Y)$ and $\widehat{CFD}(Y)$ involve counting holomorphic curves. The analytic setup here is somewhat nonstandard, complicating matters.

Like $\widehat{CF}(Y)$, the techniques of Sarkar and Wang [12] allow one to compute $\widehat{CFA}(Y)$ and $\widehat{CFD}(Y)$ combinatorially, by using a particular kind of diagram called a *nice diagram.* Such diagrams also make the pairing theorem as trivial as it was in the planar case. However, there is currently no way to prove invariance for even the closed invariant while staying in the class of nice diagrams; also, working with a nice diagram seems to require super-exponentially more generators in most cases.

10.1.3. $\mathcal{A}_\infty$-structures and noncommutative gradings

For general Heegaard diagrams, associativity fails for $\widehat{CFA}(Y)$. Fortunately, associativity holds up to homotopy, and in fact one can organize the higher associators neatly into the structure of an $\mathcal{A}_\infty$-module. (In the case that the bordered Heegaard diagram is nice, all higher associators vanish, and hence $\widehat{CFA}(Y)$ is an honest module.)

Another algebraic complication is the grading. For boundary of genus at least one, the algebra $\mathcal{A}(F)$ associated to a surface F is not $\mathbb{Z}$-graded but rather is graded by a certain noncommutative group G. (This grading intertwines the homological and spin^c gradings.) The modules associated to bordered 3-manifolds are graded by G-sets.

10.2. Complications for toroidal grid diagrams

One can also try to pursue an analogue of this theory for toroidal grid diagrams. Slicing a toroidal grid diagram yields a representation of a tangle, so this can be viewed

as a theory of tangles. There seem to be two main complications, the second more serious than the first.

10.2.1. Boundary degenerations and matrix factorizations.

For planar grid diagrams, or for bordered Heegaard diagrams, there are no domains with boundary contained entirely in the α-curves (or entirely in the β-curves). This prevents certain degenerations of holomorphic curves (called "boundary degenerations" in [10]). For toroidal grid diagrams, there are such degenerations. Their cancellation, holomorphically [10] or combinatorially [9], is delicate, and not preserved by the slicing operation. The result is that the invariants one must associate to partial toroidal grid diagrams are not differential modules but instead matrix factorizations. (Matrix factorizations also arise in other knot homology theories; see, e.g., [4].) Equivalently, one can deform a suitable version of the algebra $\mathcal{A}_{N,k}$ to an $\mathcal{A}_\infty$-algebra with a nontrivial μ_0.

10.2.2. Derived equivalences.

In this paper, we have not talked at all about invariance, because the planar Floer homology CP^- is itself not an invariant. For the toroidal theory, a partial diagram of height N and width k will result in a module over an algebra $\mathcal{A}^{\mathbb{X},\mathbb{O}}_{N,k}$, a variant of $\mathcal{A}_{N,k}$. One can have diagrams for a tangle with different heights and widths; the "invariants" associated to them, then, are modules over different algebras. In order to even express invariance, then, one would like derived equivalences

$$\mathcal{D}^b(\mathcal{A}^{\mathbb{X},\mathbb{O}}_{N,k} - \mathsf{Mod}) \to \mathcal{D}^b(\mathcal{A}^{\mathbb{X}',\mathbb{O}'}_{N',k'} - \mathsf{Mod})$$

between certain of these algebras. Moreover, these must be compatible with how stabilization acts on the modules. We return to these issues in a future paper [6].

References

[1] Joseph Bernstein and Valery Lunts, *Equivariant sheaves and functors*, Lecture Notes in Mathematics, vol. 1578, Springer-Verlag, Berlin, 1994.

[2] Bernhard Keller, *On differential graded categories*, International Congress of Mathematicians. Vol. II, Eur. Math. Soc., Zürich, 2006, pp. 151–190, arXiv:math.KT/0601185.

[3] Mikhail Khovanov, *A functor-valued invariant of tangles*, Algebr. Geom. Topol. **2** (2002), 665–741.

[4] Mikhail Khovanov and Lev Rozansky, *Matrix factorizations and link homology*, Fund. Math. **199** (2008), no. 1, 1–91.

[5] Robert Lipshitz, *A Heegaard-Floer invariant of bordered 3-manifolds*, Ph.D. thesis, Stanford University, Palo Alto, CA, 2006.

[6] Robert Lipshitz, Peter S. Ozsváth, and Dylan P. Thurston, *Bimodules in bordered Heegaard Floer homology*, in preparation.

[7] ———, *Bordered Heegaard Floer homology: Invariance and pairing*, 2008, arXiv:0810.0687.

[8] Ciprian Manolescu, Peter S. Ozsváth, and Sucharit Sarkar, *A combinatorial description of knot Floer homology*, 2006, arXiv:math.GT/0607691.

[9] Ciprian Manolescu, Peter S. Ozsváth, Zoltán Szabó, and Dylan P. Thurston, *On combinatorial link Floer homology*, Geom. Topol. **11** (2007), 2339–2412, arXiv:math.GT/0610559.

[10] Peter S. Ozsváth and Zoltán Szabó, *Holomorphic disks and link invariants*, 2005, arXiv:math/0512286.

[11] Jacob Rasmussen, *Floer homology and knot complements*, Ph.D. thesis, Harvard University, Cambridge, MA, 2003.
[12] Sucharit Sarkar and Jiajun Wang, *An algorithm for computing some Heegaard Floer homologies*, 2006, arXiv:math/0607777.

Department of Mathematics, Columbia University, New York, NY 10027
E-mail address: lipshitz@math.columbia.edu

Department of Mathematics, Columbia University, New York, NY 10027
E-mail address: petero@math.columbia.edu

Department of Mathematics, Barnard College, Columbia University, New York, NY 10027
E-mail address: dthurston@barnard.edu

Proceedings of 15th Gökova
Geometry-Topology Conference
pp. 120–136

Grid diagrams, braids, and contact geometry

Lenhard Ng and Dylan Thurston

Abstract. We use grid diagrams to present a unified picture of braids, Legendrian knots, and transverse knots.

1. Introduction

Grid diagrams, also known in the literature as arc presentations, are a convenient combinatorial tool for studying knots and links in $\mathbb{R}^3$. Although grid diagrams (or equivalent structures) have been studied for over a century ([Bru, Cro, Dyn]), they have recently regained prominence due to their role in the combinatorial formulation of knot Floer homology ([MOS, MOST]).

It has been known for some time that grid diagrams are closely related to contact geometry as well as to braid theory. Our purpose here is to indicate the extent to which the relationships are similar. Indeed, braids, like the Legendrian and transverse knots in contact geometry, can be viewed as certain equivalence classes of grid diagrams, and we will see that the various equivalences fit into one single description. Furthermore, this description is compatible with the various maps between these objects, like the transverse knot constructed from a braid. Much of the picture we will present has previously appeared, but we believe that the full picture (especially the part concerning braids) is new.

Definition 1. A *grid diagram* with *grid number* n is an $n \times n$ square grid with n X's and n O's placed in distinct squares, such that each row and each column contains exactly one X and one O.

We will employ the word "knot" throughout as shorthand for "oriented knot or oriented link". Then any grid diagram yields a diagram of a knot in a standard way: connect O to X in each row, connect X to O in each column, and have the vertical line segments pass over the horizontal ones (Figure 1). In addition, one can associate to any grid diagram not only a topological knot but also a braid, a Legendrian knot, and a transverse knot.

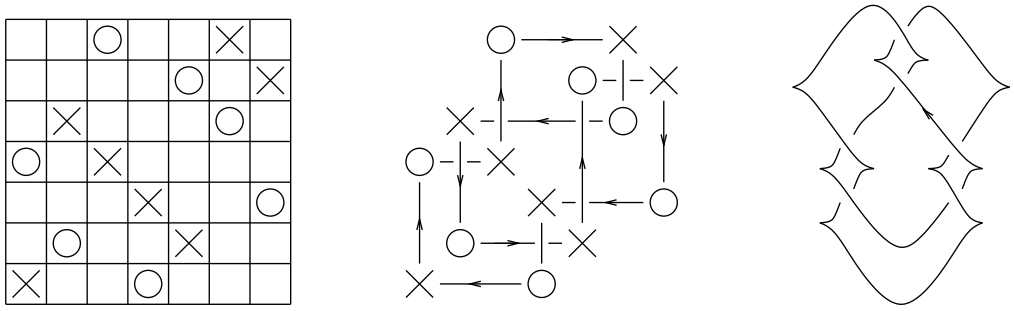


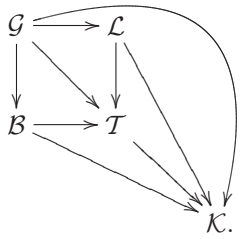
FIGURE 1. A grid diagram and corresponding knot diagram and Legendrian front.

We will use the following notation:

$$\begin{aligned}
\mathcal{G} &= \{\text{grid diagrams}\}\\
\mathcal{K} &= \{\text{isotopy classes of topological knots}\}\\
\mathcal{B} &= \{\text{isotopy classes of braids modulo conjugation and exchange}\}\\
\mathcal{L} &= \{\text{Legendrian isotopy classes of Legendrian knots}\}\\
\mathcal{T} &= \{\text{transverse isotopy classes of transverse knots}\}.
\end{aligned}$$

(For definitions, see Section 2.)

In Section 2, we will review maps between these various sets that fit together into the following commutative diagram:



(1)

Here the map from $\mathcal{G}$ to $\mathcal{K}$ is as described above. For the other maps, see also [Ben, Cro, Dyn, KN, MM, OST].

In [Cro] (see also [Dyn]), Cromwell provides a list of alterations of grid diagrams that do not change topological knot type, the grid-diagram equivalent of Reidemeister moves for knot diagrams. These are collectively known as *Cromwell moves* and consist of translations, commutations, and stabilizations/destabilizations. The last we distinguish into four types, *X:NW*, *X:NE*, *X:SW*, and *X:SE*, following [OST].

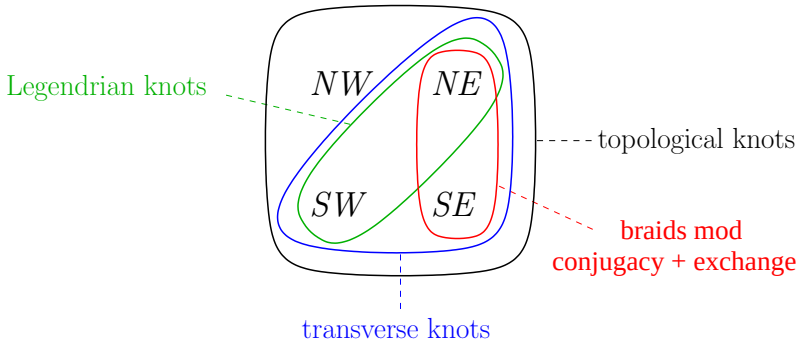


FIGURE 2. Quotienting $\tilde{\mathcal{G}}$, the set of grid-diagram orbits under translation and commutation, by various combinations of X (de)stabilizations yields equivalence classes of braids and various types of knots.

Proposition 1 (Cromwell [Cro]). *The map $\mathcal{G} \to \mathcal{K}$ sending grid diagrams to topological knots induces a bijection*

$$\mathcal{K} \longleftrightarrow \mathcal{G}/(\textit{translation, commutation, (de)stabilization}).$$

We will see that the maps from $\mathcal{G}$ to $\mathcal{B}$, $\mathcal{L}$, and $\mathcal{T}$ can be similarly understood. More precisely, we have the following result.

Proposition 2. *Let $\tilde{\mathcal{G}}$ denote the quotient set $\mathcal{G}/(\textit{translation, commutation})$. The maps $\mathcal{G} \to \mathcal{B}$, $\mathcal{G} \to \mathcal{L}$, and $\mathcal{G} \to \mathcal{T}$ induce bijections*

$$\mathcal{B} \longleftrightarrow \tilde{\mathcal{G}}/(\textit{X:NE,X:SE (de)stabilization})$$
$$\mathcal{L} \longleftrightarrow \tilde{\mathcal{G}}/(\textit{X:NE,X:SW (de)stabilization})$$
$$\mathcal{T} \longleftrightarrow \tilde{\mathcal{G}}/(\textit{X:NE,X:SW,X:SE (de)stabilization}).$$

It follows from this result that the maps between $\mathcal{B}, \mathcal{L}, \mathcal{T}, \mathcal{K}$ can also be interpreted in terms of grid diagrams. For instance, the map $\mathcal{B} \to \mathcal{T}$ is the quotient

$$\tilde{\mathcal{G}}/(\textit{X:NE,X:SE} \text{ (de)stabilization}) \longrightarrow \tilde{\mathcal{G}}/(\textit{X:NE,X:SW,X:SE} \text{ (de)stabilization}).$$

Similarly, the maps $\mathcal{B} \to \mathcal{K}$, $\mathcal{L} \to \mathcal{T}$, $\mathcal{L} \to \mathcal{K}$, $\mathcal{T} \to \mathcal{K}$, in terms of grid diagrams, are quotients by various (de)stabilizations.

Proposition 2 is summarized diagrammatically in Figure 2. The bijections in Proposition 2 involving $\mathcal{L}$ and $\mathcal{T}$ have already been established in [OST]; the new content in this note is the bijection involving $\mathcal{B}$.

We note that stabilization operations on braids and Legendrian and transverse knots can be expressed in terms of Cromwell moves. More precisely, we have the following.

Proposition 3. *Under the identifications of Proposition 2, we have*

$$\begin{aligned}
\textit{positive braid stabilization} &\longleftrightarrow \textit{X:SW stabilization}\\
\textit{negative braid stabilization} &\longleftrightarrow \textit{X:NW stabilization}\\
\textit{positive Legendrian stabilization} &\longleftrightarrow \textit{X:NW stabilization}\\
\textit{negative Legendrian stabilization} &\longleftrightarrow \textit{X:SE stabilization}\\
\textit{transverse stabilization} &\longleftrightarrow \textit{X:NW stabilization.}
\end{aligned}$$

Proposition 3 follows from an inspection of the effect of the various X stabilizations on the corresponding braid or Legendrian or transverse knot. See also the table at the end of Section 2.4.

Propositions 2 and 3 give an alternate proof via grid diagrams of the following result.

Proposition 4 (Transverse Markov Theorem [OSh, Wr])**.** *Two braids represent isotopic transverse knots if and only if they are related by a sequence of conjugations and positive braid stabilizations and destabilizations.*

In the usual formulation of Proposition 4, the map from braids to transverse knots uses a contact-geometric construction of Bennequin [Ben] (cf. Section 2.4), rather than the map we use here; see [KN] for a proof that the two maps coincide.

In Section 2, we recall the various relevant constructions and discuss the effects of grid-diagram symmetries on the maps in Formula (1). We prove our main result, Proposition 2, in Section 3.

2. Definitions and maps

2.1. Grid diagrams

The Cromwell moves on grid diagrams, translation, commutation, and stabilization/destabilization, are illustrated in Figure 3 and defined below. From that figure it is clear that each Cromwell move preserves the topological type of the corresponding knot.

Translation views a grid diagram as lying on a torus by identifying opposite ends of the grid, and changes the diagram by translation in the torus. Any translation is a composition of some number of *vertical translations*, which move the top row of the diagram to the bottom or vice versa, and *horizontal translations*, which move the leftmost column of the diagram to the rightmost or vice versa.

Commutation interchanges two adjacent rows (*vertical commutation*) or two adjacent columns (*horizontal commutation*). These adjacent rows or columns are required to be disjoint or nested in the following sense. For rows, the four X's and O's in the adjacent rows must lie in distinct columns, and the horizontal line segments connecting O and X in each row must be either disjoint or nested (one contained in the other) when projected to a single horizontal line; there is an obvious analogous condition for columns.

An X (resp. O) *destabilization* replaces a 2×2 subgrid containing two X's and one O (resp. two O's and one X) with a single square containing an X (resp. O), eliminating

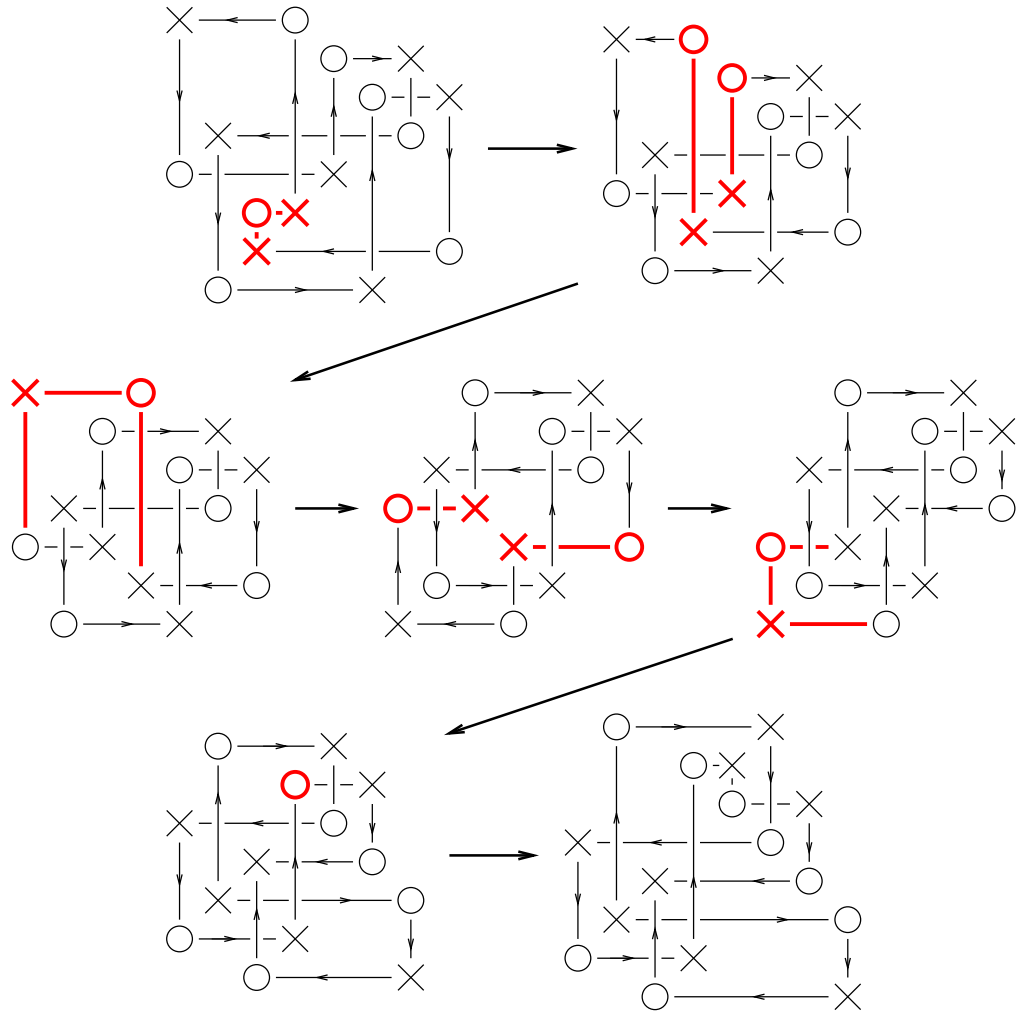


FIGURE 3. Illustration of a sequence of Cromwell moves. In succession: *X:SE* destabilization; horizontal commutation; vertical torus translation; vertical commutation; horizontal torus translation; *O:SW* stabilization. The highlighted sections of each diagram indicate the portion that changes under the following move.

one row and one column in the process. *Stabilization* is the inverse of destabilization. Each (de)stabilization is identified by its type, X or O, along with the corner in the 2×2 subgrid not occupied by a symbol. This yields eight possibilities: *X:NW*, *X:NE*, *X:SW*,

X:SE, *O:NW*, *O:NE*, *O:SW*, *O:SE*. It is easy to check that any *O:NW* (resp. *O:NE*, *O:SW*, *O:SE*) (de)stabilization can be expressed as a composition of translations, commutations, and one *X:SE* (resp. XSW, *X:NE*, *X:NW*) (de)stabilization. Thus we restrict our set of Cromwell moves to include only X (de)stabilizations.

Remark 5. By the argument of [OST, Lemma 4.3], we can instead drop torus translations and keep matching O (de)stabilizations to yield alternate definitions for topological, Legendrian, and transverse knots in terms of grid diagrams. In particular, *X:NE*, *X:SW*, *O:SW*, and *O:NE* (de)stabilizations, combined with commutations, generate all torus translations. The same argument can also be adapted for braids: that is, $\mathcal{B}$ is also $\mathcal{G}$ modulo commutation and *X:NE*, *X:SE*, *O:NW*, and *O:SW* (de)stabilization, as follows. Sequences of moves similar to those from [OST, Lemma 4.3] show that any horizontal torus translation can be achieved by these moves, as can any vertical torus translation where the O appears to the left of the X. But any vertical torus translation can be put into the correct position by horizontal torus translations.

2.2. Braids

As usual, a *braid* of braid index n is an element of the group $\mathcal{B}_n$ generated by $\sigma_1, \ldots, \sigma_{n-1}$ with relations $\sigma_i\sigma_{i+1}\sigma_i = \sigma_{i+1}\sigma_i\sigma_{i+1}$ for $1 \leq i \leq n-2$ and $\sigma_i\sigma_j = \sigma_j\sigma_i$ for $|i-j| \geq 2$. Note the natural inclusion $\mathcal{B}_n \subset \mathcal{B}_{n+1}$ sending σ_i to itself for $i \leq n-1$. The relevant moves to consider on braids are:

- braid conjugation: $B \mapsto B'B(B')^{-1}$ for $B, B' \in \mathcal{B}_n$;
- exchange move [BM]: $B_1\sigma_{n-1}B_2\sigma_{n-1}^{-1} \mapsto B_1\sigma_{n-1}^{-1}B_2\sigma_{n-1}$ on $\mathcal{B}_n$, where B_1, B_2 are in $\mathcal{B}_{n-1} \subset \mathcal{B}_n$;
- braid stabilization: either positive braid stabilization $(B \in \mathcal{B}_n) \mapsto (B\sigma_n \in \mathcal{B}_{n+1})$ or negative braid stabilization $(B \in \mathcal{B}_n) \mapsto (B\sigma_n^{-1} \in \mathcal{B}_{n+1})$; and
- braid destabilization: the inverse of braid stabilization.

In fact, by an observation of Birman and Wrinkle [BW], an exchange move can be expressed as a combination of one positive stabilization, one positive destabilization, and a number of conjugations. (Here the positive stabilization and positive destabilization can equally well be replaced by a negative stabilization and negative destabilization.) For reference, we include the calculation here.

$$\begin{aligned} B_1\sigma_{n-1}B_2\sigma_{n-1}^{-1} &\overset{\text{conj}}{\longmapsto} \sigma_{n-1}B_1\sigma_{n-1}B_2\sigma_{n-1}^{-2} \overset{+\text{ stab}}{\longmapsto} \sigma_{n-1}B_1\sigma_{n-1}B_2\sigma_{n-1}^{-2}\sigma_n \\ &\overset{\text{conj}}{\longmapsto} B_1\sigma_{n-1}B_2\sigma_{n-1}^{-2}\sigma_n\sigma_{n-1} = B_1\sigma_{n-1}B_2\sigma_n\sigma_{n-1}\sigma_n^{-2} \\ &\overset{\text{conj}}{\longmapsto} \sigma_n^{-2}B_1\sigma_{n-1}\sigma_nB_2\sigma_{n-1} = B_1\sigma_{n-1}\sigma_n\sigma_{n-1}^{-2}B_2\sigma_{n-1} \\ &\overset{\text{conj}}{\longmapsto} \sigma_{n-1}^{-2}B_2\sigma_{n-1}B_1\sigma_{n-1}\sigma_n \overset{+\text{ destab}}{\longmapsto} \sigma_{n-1}^{-2}B_2\sigma_{n-1}B_1\sigma_{n-1} \\ &\overset{\text{conj}}{\longmapsto} B_1\sigma_{n-1}^{-1}B_2\sigma_{n-1}. \end{aligned}$$

FIGURE 4. Positive and negative Legendrian stabilizations of the front projection of a Legendrian knot.

We will depict braids horizontally from left to right, with strands numbered from top to bottom; for instance, σ_1 interchanges the top two strands, with the top strand passing over the other as we move from left to right.

2.3. Legendrian and transverse knots

We give a quick description of Legendrian and transverse knots, which occur naturally in contact geometry; see, e.g., [Et] for more details. A *Legendrian knot* is a knot in $\mathbb{R}^3$ along which the standard contact form $dz - y\,dx$ vanishes everywhere; a *transverse knot* is a knot in $\mathbb{R}^3$ along which $dz - y\,dx > 0$ everywhere. (Note for the condition $dz - y\,dx > 0$ that the knot is oriented.) We consider Legendrian (resp. transverse) knots up to *Legendrian isotopy* (resp. *transverse isotopy*), which is simply isotopy through Legendrian (resp. transverse) knots.

One convenient way to depict a Legendrian knot is through its *front projection*, or projection in the xz plane. A generic front projection has three features: it has no vertical tangencies; it is immersed except at cusp singularities; and at all crossings, the strand of larger slope passes underneath the strand of smaller slope. Any front with these features corresponds to a Legendrian knot, with the y coordinate given by $y = dz/dx$.

The knot diagram corresponding to any grid diagram can be viewed as the front projection of a Legendrian knot by rotating it 45° counterclockwise and smoothing out the corners, creating cusps where necessary; see Figure 1 for an example. This yields a map $\mathcal{G} \to \mathcal{L}$ from grid diagrams to isotopy classes of Legendrian knots. Note that our convention differs from the convention of [OST]: the convention there is to reverse all crossings in the grid diagram and then rotate 45° clockwise. See also Section 2.5.

In [OST], it is verified that changing a grid diagram by translation, commutation, or (in our convention) *X:SW*, *X:NE* (de)stabilization does not change the isotopy class of the corresponding Legendrian knot. Changing by *X:NW* (resp. *X:SE*) stabilization does change the Legendrian knot type, by *positive Legendrian stabilization* (resp. *negative Legendrian stabilization*). Legendrian stabilizations can be described in the front projection as adding a zigzag, as shown in Figure 4.

Any Legendrian knot is isotopic to one obtained from some grid diagram. It is shown in [OST] that the set of equivalence classes of Legendrian knots under Legendrian isotopy corresponds precisely to grid diagrams modulo translation, commutation, and *X:NE*, *X:SW* (de)stabilization, as presented in Proposition 2.

A Legendrian knot can be C^0 perturbed to a transverse knot, its positive transverse pushoff. The resulting map $\mathcal{L} \to \mathcal{T}$ is not injective; negative Legendrian stabilization does

not change the transverse isotopy type of the positive transverse pushoff. It is a standard fact in contact geometry [EFM] that this gives a bijection

$$\mathcal{T} \longleftrightarrow \mathcal{L}/(\text{negative Legendrian stabilization}).$$

Since negative Legendrian stabilization corresponds to an *X:SE* Cromwell move, the characterization in Proposition 2 of $\mathcal{T}$ as a quotient of $\mathcal{G}$ holds. Note that positive Legendrian stabilization becomes the "transverse stabilization" operation on transverse knots.

2.4. Maps between $\mathcal{G}, \mathcal{B}, \mathcal{L}, \mathcal{T}, \mathcal{K}$

Here we collect the constructions of the maps in Formula (1). It suffices to define $\mathcal{G} \to \mathcal{L}$, $\mathcal{G} \to \mathcal{B}$, $\mathcal{L} \to \mathcal{T}$, $\mathcal{B} \to \mathcal{T}$, and $\mathcal{T} \to \mathcal{K}$, since the other maps follow by composition. We note that the commutativity of the square

$$\begin{array}{ccc} \mathcal{G} & \longrightarrow & \mathcal{L} \\ \downarrow & & \downarrow \\ \mathcal{B} & \longrightarrow & \mathcal{T} \end{array}$$

was established in [KN], and in fact our description of the maps is essentially identical to the one given there. The maps $\mathcal{G} \to \mathcal{L}$ and $\mathcal{L} \to \mathcal{T}$ have already been discussed; since the map $\mathcal{T} \to \mathcal{K}$ is obvious, we are left with $\mathcal{G} \to \mathcal{B}$ and $\mathcal{B} \to \mathcal{T}$.

We begin with the map $\mathcal{G} \to \mathcal{B}$, as described in [Cro, Dyn]; this is also called a "flip" in [MM]. Any braid in B_n can be viewed as a braid diagram: a tangle diagram of n strands in the strip $[0,1] \times \mathbb{R}$, oriented so that the orientation points rightward at all points, with some collection of n distinct points $x_1, \ldots, x_n \in \mathbb{R}$ for which the braid intersects $\{0\} \times \mathbb{R}$ and $\{1\} \times \mathbb{R}$ in $\{(0,x_1), \ldots, (0,x_n)\}$ and $\{(1,x_1), \ldots, (1,x_n)\}$ respectively. Define a *rectilinear braid diagram* (cf. "braided rectangular diagram" [MM]) to be a tangle diagram in $[0,1] \times \mathbb{R}$ with the same boundary conditions as a braid diagram, but consisting exclusively of horizontal and vertical line segments, satisfying the following properties:

- vertical segments always pass over horizontal segments;
- each strand can be oriented so that every horizontal segment is oriented rightwards.

Any rectilinear braid diagram can be perturbed into a standard braid diagram by perturbing vertical segments slightly to point rightwards, as in Figure 5.

Now given a grid diagram, one obtains a knot diagram as usual by drawing horizontal and vertical lines. Turn this into a rectilinear braid diagram by replacing any horizontal line oriented leftwards from O to X by two horizontal lines, one pointing rightwards from the O, one pointing rightwards to the X, and have these new horizontal lines pass under all vertical line segments as usual. The rectilinear braid diagram corresponds to a braid as described above. This produces the desired map $\mathcal{G} \to \mathcal{B}$.

It remains to define the map $\mathcal{B} \to \mathcal{T}$. The original contact-geometric definition from [Ben] is as follows. Identify ends of B to obtain a knot or link in the solid torus $S^1 \times D^2$.

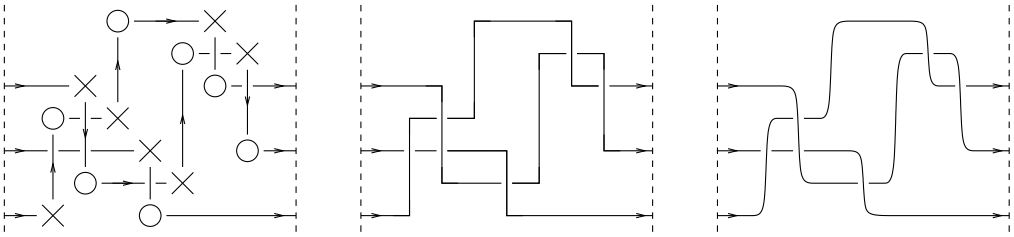


FIGURE 5. Braid version (left) of the grid diagram in Figure 1. Omitting the X's and O's produces a rectilinear braid diagram, which can be perturbed to become a braid, in this case $\sigma_2^{-1}\sigma_1\sigma_2^2\sigma_1^2 \in \mathcal{B}_3$.

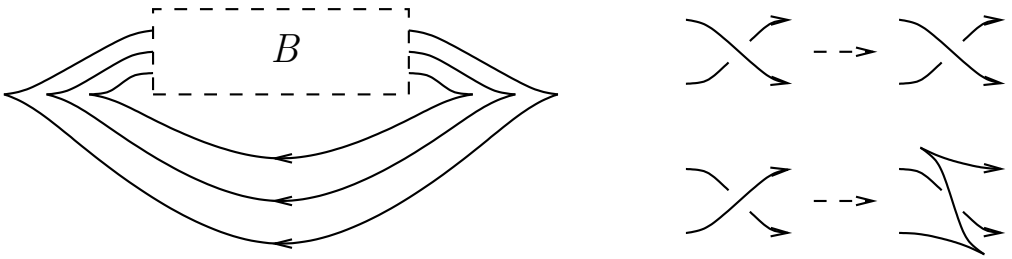


FIGURE 6. A Legendrian front for a braid B.

View the solid torus as a small (framed) tubular neighborhood of the standard transverse unknot in $\mathbb{R}^3$ with self-linking number -1. Then B becomes a transverse knot in a neighborhood of the transverse unknot.

There is also a combinatorial description for the map $\mathcal{B} \to \mathcal{T}$, which we now describe. (This description is proven to coincide with the contact-geometric description in [KN]; see also [MM, OSh]). Create a front by replacing each braid crossing as shown in Figure 6 and joining corresponding braid ends. (Joining ends introduces $2n$ cusps for a braid with n strands; see Figure 6.) This construction produces a Legendrian knot from any braid.

It is an easy exercise in Legendrian Reidemeister moves to show that changing the braid by isotopy changes the Legendrian knot by isotopy and negative Legendrian (de)stabilization; the stabilization is needed when one introduces cancelling terms $\sigma_i\sigma_i^{-1}$ or $\sigma_i^{-1}\sigma_i$ in the braid. Similarly, a conjugation or exchange move on a braid produces a Legendrian isotopy of the Legendrian knot. See Figure 7 for the exchange move.

The map $\mathcal{B} \to \mathcal{T}$ is now given as follows: given a braid, the corresponding Legendrian front is well-defined up to isotopy and negative Legendrian stabilization, and hence its positive transverse pushoff is well-defined. This transverse knot (equivalently, the class of the Legendrian knot modulo negative Legendrian (de)stabilization) is unchanged by braid conjugation and exchange.

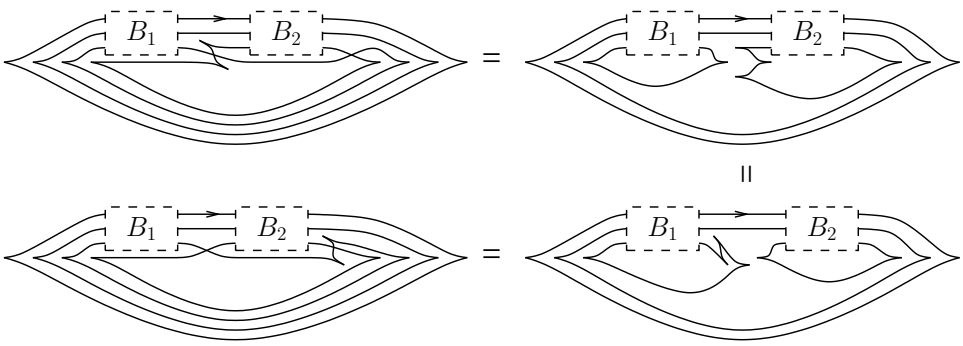


FIGURE 7. A braid exchange move produces a Legendrian-isotopic front. Equality denotes Legendrian isotopy.

Grid diagram	Braid	Legendrian knot	Transverse knot
torus translation	conjugation	Legendrian isotopy	transverse isotopy
vertical commutation	unchanged	Legendrian isotopy	transverse isotopy
horizontal commutation	conj, exchange	Legendrian isotopy	transverse isotopy
X:NE, *O:SW* stab	unchanged	Legendrian isotopy	transverse isotopy
X:SW, *O:NE* stab	conj, + braid stab	Legendrian isotopy	transverse isotopy
X:SE, *O:NW* stab	unchanged	− Legendrian stab	transverse isotopy
X:NW, *O:SE* stab	conj, − braid stab	+ Legendrian stab	transverse stab

TABLE 1. The effect of Cromwell moves on associated topological structures.

Table 1 has a summary of the effect of the Cromwell moves on grid diagrams correspond to changes in the associated braid, Legendrian knot, and transverse knot. The braid column is verified in Section 3, while the Legendrian and transverse columns were established in [OST]. For completeness, the table includes O as well as X stabilizations.

2.5. Symmetries and conventions

Here we discuss various symmetries of grid diagrams and how they relate the conventions for the maps in Formula (1) to other, sometimes conflicting, conventions in the literature. In this section, we will denote the maps $\mathcal{G} \to \mathcal{L}$, $\mathcal{G} \to \mathcal{T}$, $\mathcal{G} \to \mathcal{B}$ described in Section 2.4 by $G \mapsto L(G)$, $G \mapsto T(G)$, $G \mapsto B_{\rightarrow}(G)$, respectively.

Consider the symmetries S_1, S_2, S_3, and S_4 of grid diagrams defined as follows:

- S_1 rotates the grid diagram 180°;
- S_2 reflects the diagram about the NE-SW diagonal and interchanges X's and O's;
- S_3 reflects the diagram across the horizontal axis; and
- S_4 rotates the grid diagram 180° and interchanges X's and O's.

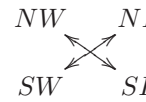
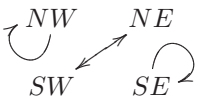
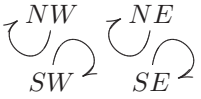
Symm	Knot	Braid	Legendrian	Transverse	X stabilizations
S_1	$K \mapsto K$	$B_{\rightarrow} \mapsto B_{\leftarrow}$	$L \mapsto \mu(L)$	—	
S_2	$K \mapsto K$	$B_{\rightarrow} \mapsto B_{\uparrow}$	$L \mapsto L$	$T \mapsto T$	
S_3	$K \mapsto m(K)$	$B_{\rightarrow} \mapsto m(B_{\rightarrow})$	—	—	
S_4	$K \mapsto -K$	$B_{\rightarrow} \mapsto -B_{\rightarrow}$	$L \mapsto -\mu(L)$	$T \mapsto -\mu(T)$	

TABLE 2. The effect of symmetries of a grid diagram on associated topological structures.

Both S_1 and S_2 preserve topological knot type, while S_3 produces the topological mirror knot $m(K)$ (with reversed orientation on $\mathbb{R}^3$), and S_4 produces the inverse (i.e., orientation-reversed) knot $-K$.

The symmetries descend to the quotient $\tilde{\mathcal{G}}$ of grid diagrams by translation and commutation. On $\tilde{\mathcal{G}}$, it is readily checked that the symmetries permute the four X stabilizations as shown in Table 2. We will use this information to examine the effect of the symmetries on Legendrian and transverse knots and braids, as shown in the table and explained below.

Since S_1 and S_2 send *X:NE*, *X:SW* stabilizations to themselves or each other, Proposition 2 implies that these symmetries descend to maps on $\mathcal{L}$. Indeed, it can be shown (see, e.g., [OST, Lemma 4.6]) that S_2 does not change Legendrian isotopy type: $L \circ S_2(G) = L(G)$. It follows also that $T \circ S_2(G) = T(G)$. On the other hand, we have $L \circ S_1(G) = \mu(L(G))$, where $\mu : \mathcal{L} \to \mathcal{L}$ is the Legendrian mirror operation, which reflects Legendrian front diagrams in the horizontal axis [FT, OST]. In general, the two maps lead to two distinct Legendrian knots [Ng]; note that Legendrian "mirroring" preserves topological type. We remark that S_3 does not descend to a map on $\mathcal{L}$ (there is no Legendrian version of the topological mirror construction), and Legendrian mirrors do not descend to the transverse category.

The map S_4 on Legendrian knots produces the orientation reverse of the Legendrian mirror: $L \mapsto -\mu(L)$. This operation descends to (oriented) transverse knots, in an operation that could be called the transverse mirror.

We next consider braids. Given a grid diagram, there are four equally valid ways to obtain a map $\mathcal{G} \to \mathcal{B}$ that preserves topological knot type. One can require that the braid goes from left to right, as we do in Section 2.4, but one could instead require that the braid go from bottom to top, right to left, or top to bottom. We write the resulting maps as $G \mapsto B_{\rightarrow}(G)$, $G \mapsto B_{\uparrow}(G)$, $G \mapsto B_{\leftarrow}(G)$, and $G \mapsto B_{\downarrow}(G)$, respectively. In general, these maps lead to four distinct braids, related by

$$B_{\rightarrow} \circ S_1(G) = B_{\leftarrow}(G) \qquad B_{\rightarrow} \circ S_2(G) = B_{\uparrow}(G) \qquad B_{\rightarrow} \circ S_1 \circ S_2(G) = B_{\downarrow}(G).$$

As noted in [KN], it follows from $L \circ S_2(G) = L(G)$ that the braids $B_{\rightarrow}(G)$ and $B_{\uparrow}(G)$ represent the same element of $\mathcal{T}$ even though they usually differ in $\mathcal{B}$, and the same is true of the pair $B_{\leftarrow}(G)$ and $B_{\downarrow}(G)$. In addition, if we define operations $B \mapsto m(B)$ and $B \mapsto -B$ on braids, where $m(B)$ replaces every letter in B by its inverse and $-B$ is the braid word B read backwards, then $B_{\rightarrow} \circ S_3(G) = m(B_{\rightarrow}(G))$ and $B_{\rightarrow} \circ S_4(G) = -B_{\rightarrow}(G)$.

All symmetries of the NW-NE-SE-SW square are generated by S_1, S_2, S_3. The following generalization of Proposition 2 is an immediate consequence of the symmetries and Proposition 2.

Corollary 6. *We have bijections*

$$\tilde{\mathcal{G}}/(X{:}NE, X{:}SE) \xrightarrow{B_{\rightarrow}} \mathcal{B} \qquad \tilde{\mathcal{G}}/(X{:}SW, X{:}SE) \xrightarrow{B_{\uparrow}} \mathcal{B}$$

$$\tilde{\mathcal{G}}/(X{:}NW, X{:}SW) \xrightarrow{B_{\leftarrow}} \mathcal{B} \qquad \tilde{\mathcal{G}}/(X{:}NW, X{:}NE) \xrightarrow{B_{\downarrow}} \mathcal{B}$$

$$\tilde{\mathcal{G}}/(X{:}NE, X{:}SW) \xrightarrow{L} \mathcal{L} \qquad \tilde{\mathcal{G}}/(X{:}NW, X{:}SE) \xrightarrow{L \circ S_3} \mathcal{L}$$

$$\tilde{\mathcal{G}}/(X{:}NE, X{:}SW, X{:}SE) \xrightarrow{T} \mathcal{T} \qquad \tilde{\mathcal{G}}/(X{:}NW, X{:}NE, X{:}SW) \xrightarrow{T \circ S_1} \mathcal{T}$$

$$\tilde{\mathcal{G}}/(X{:}NW, X{:}SW, X{:}SE) \xrightarrow{T \circ S_3} \mathcal{T} \qquad \tilde{\mathcal{G}}/(X{:}NW, X{:}NE, X{:}SE) \xrightarrow{T \circ S_3 \circ S_2} \mathcal{T}$$

where L, T are induced from the maps $\mathcal{G} \to \mathcal{L}$, $\mathcal{G} \to \mathcal{T}$ described in Section 2.4.

Note that three of the bijections in Proposition 6 involve S_3 and thus topological mirroring.

We now discuss the conventions used in Section 2.4 in light of symmetries of grid diagrams. Our conventions are chosen to make the maps in Formula (1) always preserve topological knot type. This involves making several choices:

- vertical over horizontal line segments in grid diagrams (vs. horizontal over vertical), and Legendrian fronts obtained by 45° counterclockwise rotation (vs. clockwise);
- transverse knots given by positive pushoffs of Legendrian knots (vs. negative);
- braids going from left to right (vs. bottom to top, right to left, top to bottom).

These choices largely agree with the standard conventions in the literature [Cro, Dyn, EFM, Et, MOS, MOST]. One can obtain different conventions from ours by applying grid-diagram symmetries. For braids, this is discussed above, while for transverse knots, positive pushoffs become negative pushoffs by applying the symmetry S_1: negative pushoffs are transversely isotopic under *X:NW*,*X:NE*,*X:SW* (de)stabilization.

For the knot Floer homology invariant introduced in [OST] and subsequently used in [KN, NOT], a slightly different set of conventions is useful. Here an element λ^+ of combinatorial knot Floer homology HK^- is associated to any grid diagram, and λ^+ is shown to be invariant under translation, commutation, and *X:NW*,*X:SW*,*X:SE* (de)stabilization. (Another element λ^- is also considered in [OST]; in our notation, $\lambda^- = \lambda^+ \circ S_1$.) If we apply symmetry $S_2 \circ S_3$ to a grid diagram G before calculating λ^+, then λ^+ becomes an invariant of the transverse knot $T(G)$.

In [KN, NOT, OST], the map $\mathcal{G} \to \mathcal{L}$ is thus given by $G \mapsto (L \circ S_2 \circ S_3)(G)$ rather than $G \mapsto L(G)$. More explicitly, given a grid diagram, one can use the horizontal-over-vertical convention and 45° clockwise rotation to obtain a Legendrian front, as is done in these papers. (In particular, to translate from our conventions to those of [KN], first apply $S_2 \circ S_3$ to all grid diagrams.) Note that due to the presence of S_3, λ^+ becomes an element of HK^- of the topological *mirror* of the transverse knot.

3. Proof of Proposition 2

Let $B(G)$ ($= B_{\rightarrow}(G)$ from Section 2.5) denote the braid associated to a grid diagram G as described in Section 2. Proposition 2 (or, more precisely, the braid statement of Proposition 2) is a direct consequence of the following stronger result.

Proposition 7. *Let G be a grid diagram.*

(1) *Changing G by torus translation or X:NE,X:SE (de)stabilization changes $B(G)$ by conjugation.*
(2) *Changing G by commutation changes $B(G)$ by a combination of conjugation and exchange moves.*
(3) *The map $G \mapsto B(G)$ induces a bijection between $\mathcal{G}$/(translation, commutation, X:NE, X:SE (de)stabilization) and $\mathcal{B}$/(conjugation, exchange).*

Proof. We first check claims (1) and (2). A quick inspection of braid diagrams reveals that changing a grid diagram G by horizontal commutation or by *X:NE* or *X:SE* stabilization does not change the braid isotopy type of $B(G)$.

Changing G by horizontal torus translation changes $B(G)$ by conjugation; some portion of the beginning of $B(G)$ is moved to the end, or vice versa. See Figure 8.

Next we claim that changing G by vertical torus translation also changes $B(G)$ by conjugation. Indeed, consider moving the topmost column of G to the bottom. By conjugating by a horizontal torus translation if necessary, we may assume that in the relevant row, the O lies to the left of the X. Then moving the column keeps the braid unchanged; see Figure 8 again.

Finally, we claim that changing G by a vertical commutation changes $B(G)$ by conjugation and/or exchange. Indeed, by conjugating with an appropriate torus translation if necessary, we may assume the following: the two relevant rows are the bottom two rows in the grid diagram; the X and O in the bottom row both lie to the right of the X and O in the row above it; and the bottom right corner of the grid diagram is occupied by an

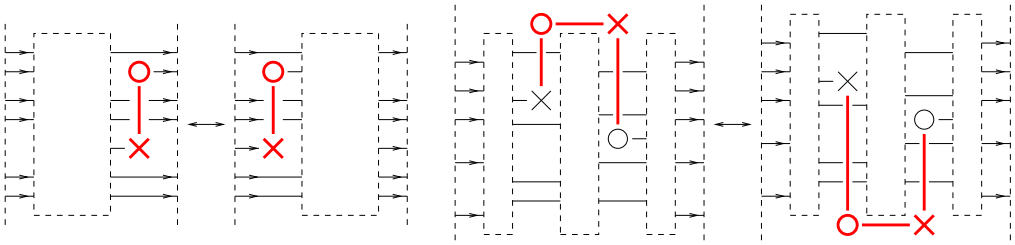


FIGURE 8. The effect on $B(G)$ of changing G by horizontal (left) and vertical (right) torus translation. The bold X and O represent the column/row being moved.

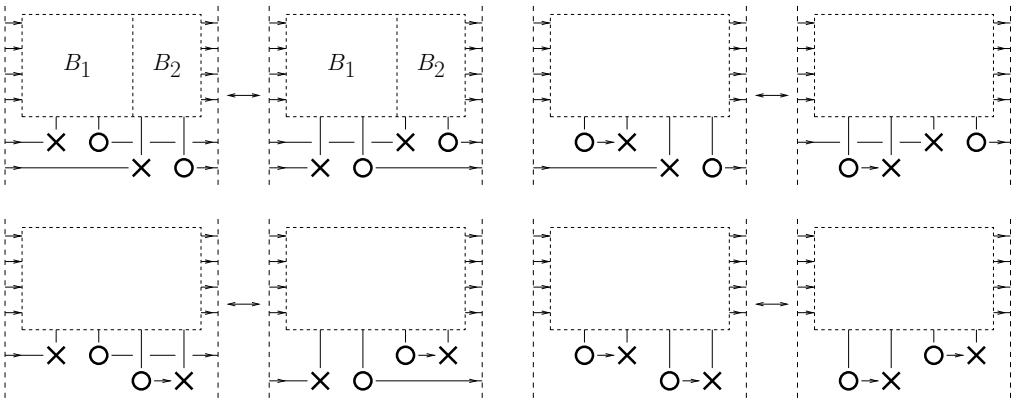


FIGURE 9. The effect on $B(G)$ of changing G by horizontal commutation. In three cases, $B(G)$ is unchanged. In the other case (upper left), the n-strand braid $B(G)$ changes from $B_1\sigma_{n-1}^{-1}B_2\sigma_{n-1}$ to $B_1\sigma_{n-1}B_2\sigma_{n-1}^{-1}$, an exchange move.

X or O. If X lies to the left of O in both rows, then the commutation changes $B(G)$ by exchange; otherwise, it does not change $B(G)$. See Figure 9.

We now establish claim (3). From claims (1) and (2), the map in (3) is well-defined. To prove bijectivity, we construct an inverse. Any braid B can be given a rectilinear braid diagram by replacing each crossing by an appropriate rectilinear version; see Figure 10.

Perturb the resulting rectilinear diagram slightly to another rectilinear diagram for which no vertical line segments have the same x-coordinate (i.e., are collinear), and no horizontal line segments have the same y-coordinate except for those that are identified when the ends of the braid are identified. The perturbed diagram is oriented (from left to right), and each corner can be assigned an X or O in the usual way. The collection of X's and O's forms a grid diagram $G(B)$, and by construction we have $B = B(G(B))$.

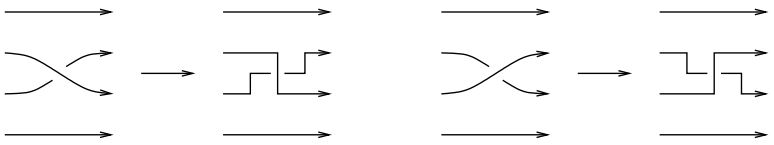


FIGURE 10. Turning a braid diagram into a rectilinear braid diagram.

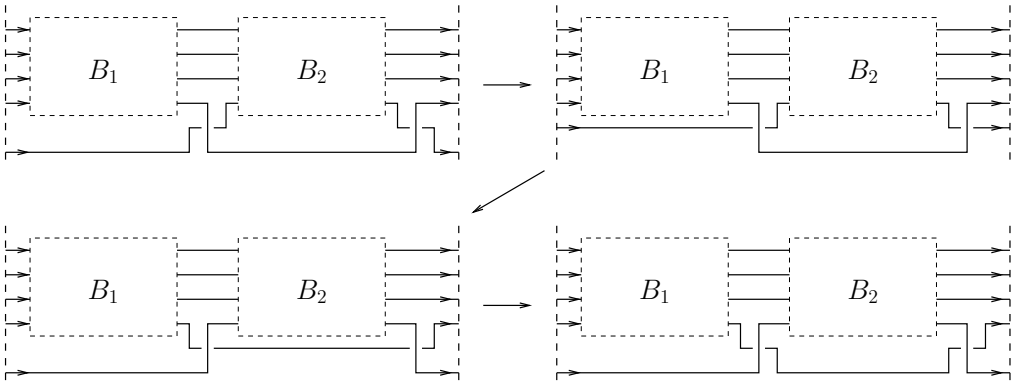


FIGURE 11. Accomplishing an exchange move through a sequence of commutation and (de)stabilization moves. The first arrow is given by commutations, one *X:NE* destabilization, and one *X:SE* destabilization; the second is a horizontal commutation; the third is commutations, one *X:NE* stabilization, and one *X:SE* destabilization. See also Figure 12 for the moves corresponding to the first and third arrows.

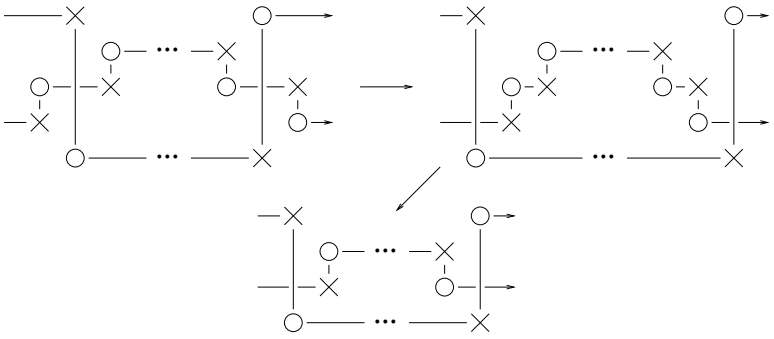


FIGURE 12. Detail of local moves in the first step of Figure 11. A vertical commutation move is followed by *X:NE* and *X:SE* destabilization.

Note that $G(B)$ depends on the choice of perturbation from rectilinear braid diagram to grid diagram, but a different perturbation simply changes $G(B)$ by commutation. In fact, up to commutation and *X:SW*,*X:SE* (de)stabilization, $G(B)$ is well-defined for an isotopy class of braids B. This fact is readily established by examining how $G(B)$ changes when the braid word for B changes by one of the relations $\sigma_i\sigma_i^{-1} = \sigma_i^{-1}\sigma_i = 1$, $\sigma_i\sigma_j = \sigma_j\sigma_i$ for $|i-j| \geq 2$, and $\sigma_i\sigma_{i+1}\sigma_i = \sigma_{i+1}\sigma_i\sigma_{i+1}$. See [Cro] for details.

In addition, changing B by conjugation changes $G(B)$ by horizontal torus translation, while changing B by an exchange move changes $G(B)$ by a combination of horizontal commutations and *X:NE*,*X:SE* (de)stabilizations; see Figures 11 and 12. Thus B induces a map from $\mathcal{B}$/(conjugation, exchange) to $\mathcal{G}$/(translation, commutation, *X:NE*, *X:SE* (de)stabilization).

If we consider G and B as maps between $\mathcal{G}$/(translation, commutation, *X:NE*, *X:SE* (de)stabilization) and $\mathcal{B}$/(conjugation, exchange), then as noted earlier, $B \circ G$ is the identity, and one readily checks that $G \circ B$ is the identity as well. Claim (3) follows, and the proof of Proposition 7 is complete. □

Acknowledgments

LLN thanks the participants of the conference "Knots in Washington XXVI" for useful comments on a preliminary version of the results presented here. DPT thanks Ciprian Manolescu, Peter Ozsváth, and Zoltán Szabó for helpful conversations. LLN was supported by NSF grant DMS-0706777; DPT was supported by a Sloan Research Fellowship.

References

[Ben] D. Bennequin, Entrelacements et équations de Pfaff, *Astérisque* **107–108** (1983), 87–161.

[BM] J. S. Birman and W. M. Menasco, Studying links via closed braids IV: Composite links and split links, *Invent. Math.* **102** (1990), no. 1, 115–139.

[BW] J. S. Birman and N. C. Wrinkle, On transversally simple knots, *J. Differential Geom.* **55** (2000), no. 2, 325–354; `arXiv:math.GT/9910170`.

[Bru] H. Brunn, Über verknotete Kurven, *Verhandlungen des Internationalen Math. Kongresses (Zürich 1897)*, 256–259, 1898.

[Cro] P. R. Cromwell, Embedding knots and links in an open book I: Basic properties, *Topology Appl.* **64** (1995), no. 1, 37–58.

[Dyn] I. A. Dynnikov, Arc-presentations of links: monotonic simplification, *Fund. Math.* **190** (2006), 29–76; `arXiv:math.GT/0208153`.

[EFM] J. Epstein, D. Fuchs, and M. Meyer, Chekanov–Eliashberg invariants and transverse approximations of Legendrian knots, *Pacific J. Math.* **201** (2001), no. 1, 89–106.

[Et] J. B. Etnyre, Legendrian and transversal knots, in *Handbook of knot theory*, 105–185, Elsevier B. V., Amsterdam, 2005; `arXiv:math.SG/0306256`.

[FT] D. Fuchs and S. Tabachnikov, Invariants of Legendrian and transverse knots in the standard contact space, *Topology* **36** (2007), no. 5, 1025–1053.

[KN] T. Khandhawit and L. Ng, A family of transversely nonsimple knots, `arXiv:0806.1887`.

[MM] H. Matsuda and W. Menasco, On rectangular diagrams, Legendrian knots and transverse knots, `arXiv:0708.2406`.

[MOS] C. Manolescu, P. Ozsváth, and S. Sarkar, A combinatorial description of knot Floer homology, `arXiv:math/0607691`.

[MOST] C. Manolescu, P. Ozsváth, Z. Szabó, and D. Thurston, On combinatorial link Floer homology, *Geom. Topol.* **11** (2007), 2339–2412; arXiv:math/0610559.

[Ng] L. Ng, Computable Legendrian invariants, *Topology* **42** (2003), no. 1, 55–82; arXiv:math.GT/0011265.

[NOT] L. Ng, P. Ozsváth, and D. Thurston, Transverse knots distinguished by knot Floer homology, *J. Symplectic Geom.*, to appear; arXiv:math/0703446.

[OSh] S. Yu. Orevkov and V. V. Shevchishin, Markov theorem for transversal links, *J. Knot Theory Ramifications* **12** (2003), no. 7, 905–913; arXiv:math.GT/0112207.

[OST] P. S. Ozsváth, Z. Szabó, and D. P. Thurston, Legendrian knots, transverse knots and combinatorial Floer homology, arXiv:math/0611841.

[Wr] N. C. Wrinkle, The Markov Theorem for transverse knots, arXiv:math.GT/0202055.

Mathematics Department, Duke University, Durham, NC 27708
E-mail address: ng@math.duke.edu

Department of Mathematics, Barnard College, Columbia University, New York, NY 10027
E-mail address: dpt@math.columbia.edu

Proceedings of 15th Gökova
Geometry-Topology Conference
pp. 137–150

Intersection forms, fundamental groups and 4-manifolds

Ian Hambleton

Abstract. This is a short survey of some connections between the intersection form and the fundamental group for smooth and topological 4-manifolds.

1. Introduction

A classical construction of Kervaire [36] shows that any finitely-presented group can be realized as the fundamental group of a closed, oriented smooth 4-manifold M. However, much less is known about other homotopy invariants of 4-manifolds, such as the second homotopy group $\pi_2(M)$, which inherits a $\mathbb{Z}[\pi_1(M,x_0)]$-module structure via the action of the deck transformations on the universal covering $\widetilde{M}$.

Another basic invariant is the *equivariant intersection form* of a 4-manifold M, defined as the triple $(\pi_1(M,x_0), \pi_2(M), s_M)$, where $x_0 \in M$ is a base-point, and

$$s_M \colon \pi_2(M) \otimes_{\mathbb{Z}} \pi_2(M) \to \mathbb{Z}[\pi_1(M,x_0)]$$

is the form defined by counting intersections of immersed 2-spheres (see [59, Chap. 5]). This pairing is Λ-hermitian, in the sense that for all $\lambda \in \Lambda := \mathbb{Z}[\pi_1(M,x_0)]$ we have

$$s_M(\lambda \cdot x, y) = \lambda \cdot s_M(x,y) \quad \text{and} \quad s_M(y,x) = \overline{s_M(x,y)}$$

where $\lambda \mapsto \bar{\lambda}$ is the involution on Λ given by $\bar{g} = g^{-1}$ for $g \in \pi_1(M,x_0)$.

The main topics of interest for the present survey are:

(1) To what extent does the fundamental group $\pi_1(M,x_0)$ and the equivariant intersection form s_M determine the topology of a closed, oriented 4-manifold M ?
(2) What special properties hold for the equivariant interesection form if M is a *smooth* 4-manifold ?

The material will be divided into sections according to the complexity of the fundamental group. From now on, all manifolds considered will be closed, connected and oriented.

Research partially supported by NSERC Discovery Grant A4000.

2. Simply-connected 4-manifolds

Wall [58], [57] showed in the 1960's that homotopy equivalent, simply-connected smooth 4-manifolds M_1, M_2 are smoothly h-cobordant, and hence are stably diffeomorphic

$$M_1 \# r(S^2 \times S^2) \cong M_2 \# r(S^2 \times S^2)$$

for some integer $r \geq 0$. It is still not known whether the existence of such a stable diffeomorphism actually requires more than one copy of $S^2 \times S^2$ (see [46], [37] for other aspects of smooth h-cobordisms).

Spectacular results concerning 4-manifolds were proved in the 1980's by S. Donaldson and M. Freedman, building on work of Atiyah, Casson, Hitchin, Taubes and Uhlenbeck. If M is simply-connected, then $\pi_2(M) \cong \mathbb{Z}^r$ is a free abelian group and the ordinary intersection form

$$q_M \colon H_2(M;\mathbb{Z}) \times H_2(M;\mathbb{Z}) \to \mathbb{Z}$$

is a symmetric, unimodular bilinear form. The signature of this form, denoted $\operatorname{sign}(M)$, is the difference between the number of positive and negative eigenvalues of a matrix representing q_M.

Freedman [16], [17] proved that any such form is realized by one or two *topological* 4-manifolds. Moreover, M is classified up to homeomorphism by q_M and the Kirby-Siebenmann invariant $KS(M) \in \mathbb{Z}/2$ (see [38] for the definition). Donaldson [6], [7], [8] showed using gauge theory that if q_M is a positive definite form then

$$q_M \cong \langle 1 \rangle \perp \langle 1 \rangle \perp \cdots \perp \langle 1 \rangle$$

is standard, and that the h-cobordant, smooth, simply-connected 4-manifolds are not necessarily diffeomorphic.

These results show a striking difference between smooth and topological 4-manifolds. By combining them, it follows that a smooth, non-spin, simply-connected 4-manifold M is homeomorphic to a connected sum of copies of $\pm\mathbf{CP}^2$. If M is smooth, simply-connected and spin, then M is homeomorphic to a connected sum of copies of $S^2 \times S^2$ and $\pm K3$ surfaces, provided that

$$b_2(M) \geq \frac{11}{8}|\operatorname{sign}(M)|,$$

where $b_2(M) = \operatorname{rank}(H_2(M;\mathbb{Z}))$. The well-known $\frac{11}{8}$-conjecture, still unresolved, states that this inequality always holds for smooth, spin 4-manifolds: the best partial result to date is $b_2(M) \geq \frac{5}{4}|\operatorname{sign}(M)| + 2$, if q_M is indefinite, proved by Furuta [20]. The exciting subsequent developments in the study of smooth, simply-connected 4-manifolds are outside the scope of this survey (there is a large and growing literature: for example, the work of Fintushel-Stern [13], [14], Gompf [21], Friedman-Morgan [19], Kronheimer-Mrowka [41], Ozsváth-Szabó [48], [47], Jongil Park [50], Taubes [53], and Seiberg-Witten [61]).

3. Infinite cyclic fundamental groups

If $\pi_1(M) = \mathbb{Z}$ and $\Lambda = \mathbb{Z}[\mathbb{Z}]$, then $\pi_2(M)$ is a finitely-generated, free Λ-module of rank $b_2(M)$ and s_M is a non-singular hermitian form. The classification theorem for these manifolds uses the full equivariant intersection form.

Theorem 3.1 (Freedman-Quinn [17]). *A closed, oriented topological 4-manifold M with $\pi_1(M) = \mathbb{Z}$ is classified up to homeomorphism by s_M and $KS(M)$. Any non-singular hermitian form on a finitely-generated free Λ-module can be realized by one or two manifolds.*

More precisely, such forms are *even* and realized by a unique spin manifold, or *odd* and realized by two non-spin manifolds with different Kirby-Siebenmann invariants. The equivariant intersection form of a connected sum $M = (S^1 \times S^3) \# N$, with a 1-connected manifold N, is said to be *extended from the integers*. In other words, $s_M = q_N \otimes_{\mathbb{Z}} \Lambda$. Conversely, by the classification theorem, any manifold whose equivariant intersection form is extended from the integers must be homeomorphic to a connected sum with $S^1 \times S^3$.

Fintushel and Stern [12] constructed a smooth 4-manifold M, which was homeomorphic but not diffeomorphic to a connected sum with $S^1 \times S^3$. The existence of indecomposable topological 4-manifolds with $\pi_1 = \mathbb{Z}$ and $\chi(M) > 0$ was settled later.

Theorem 3.2 ([23]). *There exists a closed, oriented topological 4-manifold M with $\pi_1(M) = \mathbb{Z}$ and $\chi(M) = 4$, and M is not homotopy equivalent to a connected sum $(S^1 \times S^3) \# N$ for any 1-connected N.*

The main step in the proof was the construction of a non-extended hermitian form L on a free Λ-module (using a certain odd, definite, rank 4 form over $\mathbb{Z}[t]$ found by Quebbemann [51, §6]). We also showed that any 4-manifold M with $\pi_1(M) = \mathbb{Z}$ and $b_2(M) - |\operatorname{sign}(M)| \geq 6$ splits off $S^1 \times S^3$ and is determined up to homeomorphism by the explicit invariants b_2, sign, w_2 and KS.

Question. If M is a smooth, closed, oriented 4-manifold with $\pi_1(M) = \mathbb{Z}$, then is s_M extended from the integers ?

This is a natural question after comparing the example $M = M_L$ in Theorem 3.2 with the Fintushel-Stern example.

Theorem 3.3 ([18]). *The manifold M_L is not smoothable.*

The idea of the proof is to consider the n-fold cyclic coverings $M_n \to M_L$. Since q_{M_L} is standard of rank 4, and both Euler characteristic and signature multiply by the index of a finite covering, the forms q_{M_n} of rank $= 4n$ are all definite, odd, unimodular forms over $\mathbb{Z}$. This seems to be an interesting series of definite forms: we showed that for $n \geq 3$ they were all non-standard, and for $n = 3, 4$ they were the unique indecomposable odd lattices in dimension 12 and 16 respectively. In any case, by Donaldson's theorem M_n is non-smoothable for $n \geq 3$ and hence M_L is non-smoothable. In [18] we found many more

examples of non-extended forms, and manifolds realizing these forms with a wide variety of other infinite fundamental groups.

4. The quadratic 2-type and surgery

In the non simply-connected case, the obvious homotopy invariants are the equivariant intersection form s_M and the first k-invariant

$$k_M \in H^3(\pi; \pi_2(M)),$$

which together with $\pi := \pi_1(M, x_0)$ and π_2 specifies the algebraic 2-type $B = B(M)$ as introduced by MacLane and Whitehead [44]. The space B is a fibration over $K(\pi, 1)$, classified by k_M, with fibre $K(\pi_2(M), 2)$ and there is a 3-connected reference map $\tilde{c}\colon M \to B(M)$ lifting the classifying map $c\colon M \to K(\pi, 1)$ for the universal covering $\widetilde{M} \to M$. In [22] we introduced the *quadratic 2-type* of M as the quadruple

$$[\pi_1(M, x_0), \pi_2(M), k_M, s_M] \ .$$

An *isometry* of two such quadruples is an isomorphism on π_1, π_2 inducing an isometry of the equivariant intersection forms, and respecting the k-invariants.

In general, not much is known about these homotopy invariants, but they are related by an exact sequence

$$0 \to H^2(\pi; \Lambda) \to H^2(M; \Lambda) \to \mathrm{Hom}_\Lambda(H_2(M; \Lambda), \Lambda) \to H^3(\pi; \Lambda) \to 0 \qquad (1)$$

arising from the universal coefficient spectral sequence. In this sequence, $H^2(M; \Lambda) \cong H_2(M; \Lambda) \cong \pi_2(M)$ by Poincaré duality, and the middle map

$$H^2(M; \Lambda) \to \mathrm{Hom}_\Lambda(H_2(M; \Lambda), \Lambda)$$

is the adjoint of s_M. The radical $R(s_M)$ of the intersection form s_M is isomorphic to the π-module $R(\pi) := H^2(\pi; \Lambda)$, and $\pi_2(M)$ is a finitely-generated Λ-module.

If $\pi := \pi_1(M, x_0)$ is a non-trivial finite group, then $\pi_2(M)$ is a finitely-generated free abelian group with a $\Lambda := \mathbb{Z}\pi$-module structure, as studied in integral representation theory. In general, there are infinitely many non-isomorphic indecomposable integral representations (e.g. for $\pi = \mathbb{Z}/p \times \mathbb{Z}/p$), and there is no known classification. If $\pi_1(M)$ is infinite, the precise structure of $\pi_2(M)$ is unknown except in very special cases, such as $\pi_1(M) = \mathbb{Z}$ mentioned in Section 3.

The study of these modules can be simplified somewhat by considering stable equivalence classes: two modules L_1, L_2 are *stably isomorphic*, denoted $L_1 \simeq_s L_2$, if there exists a free module Λ^r such that $L_1 \oplus \Lambda^r \cong L_2 \oplus \Lambda^r$. For example, the kernel

$$0 \to \Omega^{n+1}\mathbb{Z} \to F_n \to F_{n-1} \to \cdots \to F_1 \to F_0 \to \mathbb{Z} \to 0$$

after n-steps in a free resolution $\{F_*\}$ of the trivial module $\mathbb{Z}$ is stably unique by Schanuel's Lemma. For $n = 3$, such modules arise as $\pi_2(K) = H_2(K; \Lambda)$, where K is a finite 2-complex with $\pi_1(K, x_0) = \pi$, and the resolution is obtained from the chain complex $C_*(\widetilde{K})$ of the universal covering. Finite 2-complexes K provide examples of smooth

4-manifolds by taking the boundary of a thickening (i.e. a regular neigbourhood) of K in $\mathbb{R}^5$.

The stabilization operation in algebra has analogues in topology. For 2-complexes, $K \mapsto K \vee S^2$ gives the stabilization $\pi_2(K) \mapsto \pi_2(K) \oplus \Lambda$. Whitehead [60, Theorem 19] showed that any two finite 2-complexes K, K' with isomorphic fundamental groups are stably simple-homotopy equivalent, but the problem of finding the minimal Euler characteristic realized by a 2-complex with given π_1 is still unsolved. This is a cancellation problem. Note that Whitehead's Theorem implies that the stable isomorphism type of the Λ-modules $H_2(K;\Lambda)$ and $H^2(K;\Lambda)$ depend only on the fundamental group, and not on the choice of finite 2-complex K.

It turns out that the stable structure of $\pi_2(M)$ for a 4-manifold is very special.

Theorem 4.2. *Let M be a closed, oriented 4-manifold with fundamental group π. Then $\pi_2(M)$ is stably isomorphic as a Λ-module to a certain extension*

$$\mathcal{E}_M : 0 \to H_2(K;\Lambda) \to E \to H^2(K;\Lambda) \to 0$$

where K is any finite 2-complex with $\pi_1(K, x_0) = \pi$.

Remark 4.3. The boundaries of thickenings of 2-complexes yield trivial extensions. The finite fundamental group case was done in [22, 2.4], and in that case the extension class of $\mathcal{E}_M$ corresponds to the image of the fundamental class $c_*[M] \in H_4(\pi;\mathbb{Z})$ under a natural isomorphism $\theta\colon H_4(\pi;\mathbb{Z}) \cong \mathrm{Ext}^1_\Lambda(H^2(K;\Lambda), H_2(K;\Lambda))$.

Proof. We will use a chain complex argument. By stabilizing $K \mapsto K \vee rS^2$ and $M \mapsto M \# t(S^2 \times S^2)$ if necessary, we may assume that K is the sub-complex of 2-cells of M. Consider the cellular chain complex $C_* = C_*(\widetilde{M})$ of finitely-generated free Λ-modules. We have the exact sequences

$$0 \to \mathcal{Z}_2 \to C_2 \to C_1 \to C_0 \to \mathbb{Z} \to 0$$

and

$$0 \to \mathfrak{B}_3^* \to C_3^* \to C_4^* \to \mathbb{Z} \to 0$$

showing that $\mathfrak{B}_3^* = \mathrm{Hom}_\Lambda(\mathfrak{B}_3, \Lambda)$ is stably isomorphic to the 2-boundaries $\mathfrak{B}_2$. The details here depend on whether π is finite or infinite: in the latter case note that

$\mathrm{Ext}^1_\Lambda(\mathfrak{B}_3, \Lambda) \cong H^4(M;\Lambda) = \mathbb{Z}$. We now form the pull-back diagram

$$\begin{array}{ccccccccc}
 & & & & 0 & & 0 & & \\
 & & & & \uparrow & & \uparrow & & \\
0 & \longrightarrow & H^2(M;\Lambda) & \longrightarrow & H^2(K;\Lambda) & \longrightarrow & \mathfrak{B}_2 & \longrightarrow & 0 \\
 & & \| & & \uparrow & & \uparrow & & \\
0 & \longrightarrow & H^2(M;\Lambda) & \longrightarrow & E & \longrightarrow & C_2 & \longrightarrow & 0 \\
 & & & & \uparrow & & \uparrow & & \\
 & & & & H_2(K;\Lambda) & = & \mathcal{Z}_2 & & \\
 & & & & \uparrow & & \uparrow & & \\
 & & & & 0 & & 0 & &
\end{array} \tag{4}$$

where $\mathcal{Z}_2 = H_2(K;\Lambda)$ since K is the 2-skeleton of M. The middle horizontal sequence splits since C_2 is a free Λ-module. The middle vertical sequence is $\mathcal{E}_M$, and $E \cong \pi_2(M) \oplus C_2$ is a stabilization of $\pi_2(M)$. □

Surgery theory as developed by Browder, Novikov, Sullivan and Wall [59] provides a powerful framework for classifying manifolds of dimension ≥ 5 within a fixed homotopy type. However, in dimension 4 there are serious obstacles arising from the failure of the Whitney trick. One approach, developed by Cappell and Shaneson [3], is based on Wall's idea of studying smooth 4-manifolds after stabilization with copies of $S^2 \times S^2$. The drawback is that information about the original (unstabilized) homotopy type is lost in the process.

Freedman's work [16] fully established 4-dimensional surgery theory for topological manifolds whose fundamental groups do not "grow" too quickly. This class includes the poly-cyclic by finite groups, but it is not known at present if 4-dimensional topological surgery theory works for (non-cyclic) free fundamental groups. Note that Donaldson's results show that smooth surgery theory definitely does not work in dimension 4, and there are s-cobordant smooth 4-manifolds which are not diffeomorphic.

The modified surgery theory of M. Kreck [39] requires less initial information about the homotopy type: for example, one can try to classify smooth 4-manifolds which have the same algebraic 2-type (up to smooth s-cobordism). In this theory, the key step is to compute certain bordism groups $\Omega_4(B, \xi)$, where ξ is a bundle over B whose pullback $c^*(\xi) \cong \nu_M$ is the stable normal bundle of M. For such computations there are a variety of methods available, including the Atiyah-Hirzebruch and Adams spectral sequences. If two manifolds $[M_1.\tilde{c}_1]$, $[M_2, \tilde{c}_2]$ are bordant over the type (B, ξ), then the triviality of an algebraically defined invariant implies that M_1 and M_2 are smoothly s-cobordant (see [39, Theorem B]).

One possible way of analysing the final step is to note that the relation $[M_1, \tilde{c}_1] = [M_2, \tilde{c}_2] \in \Omega_4(B, \xi)$ implies that

$$M_1 \# r_1(S^2 \times S^2) \cong M_2 \# r_2(S^2 \times S^2)$$

are stably diffeomorphic [39, Cor. 3], with control on the reference maps to B (see [40], [5] for further applications of stabilization). The *cancellation* problem is to find techniques for removing $S^2 \times S^2$ factors from both sides. In algebra, cancellation theorems for modules and quadratic forms over noetherian rings were proved by Bak [1], Bass [2], Stafford [52] and Vaserstein [56]. In [23], we realized that these algebraic results could be combined with the constructions by [3, 1.5] of self-diffeomorphisms of 4-manifolds to prove cancellation theorems for certain 4-manifolds. For example, the integral group rings $\mathbb{Z}[\pi]$ of poly-cyclic by finite groups are noetherian rings, but the group ring of a free group on 2 generators is not noetherian. This theme has recently been taken up again in [4].

5. Finite fundamental group

In early joint work with M. Kreck [23] we studied the topology of 4-manifolds with finite fundamental groups, and obtained a good description of the homotopy types within a prescribed algebraic 2-type. We also showed that there are only finitely many homeomorphism types of closed, oriented 4-manifolds with given finite π_1, and given Euler characteristic (see [22, p. 87]). To obtain precise classification results up to homeomorphism we needed to restrict to special fundamental groups. We say that M has w_2-type (I) if $w_2(\widetilde{M}) \neq 0$, w_2-type (II) if $w_2(M) = 0$, and w_2-type (III) if $w_2(M) \neq 0$ but $w_2(\widetilde{M}) = 0$.

Theorem 5.1 ([22], [24])**.** *Closed, oriented topological 4-manifolds with finite cyclic fundamental groups are classified up to homeomorphism by $\pi_1(M)$, q_M, the w_2-type, and $KS(M)$.*

This is a generalization of Freedman's Theorem in the simply-connected case (where the w_2-type is determined by the intersection form). One interesting consequence is that certain automorphisms of the cohomology ring of a smooth 4-manifold are induced by self-homeomorphisms but not by a self-diffeomorphism (Donaldson's work is used to rule out a self-diffeomorphism, see the example [22, p.87]).

However, for more complicated fundamental groups we can not expect a classification in terms of the ordinary intersection form q_M on $H_2(M; \mathbb{Z})$ (see [54], [55]). Here is a sample result involving the quadratic 2-type.

Theorem 5.2. *Closed, oriented, topological 4-manifolds with $w_2(M) = 0$ and odd order finite fundamental groups are classified up to homeomorphism by the simple isometry class of the quadratic 2-type $[\pi_1(M), \pi_2(M), k_M, s_M]$.*

Proof. This is essentially an exercise in the methods of [39] and [25], and the information needed to use the odd order assumption is provided by [26, Section 4]. The definition of *simple isometry class* will be explained below.

Here are some details of the steps in the argument. We first notice that the normal 2-type for such a spin manifold M is $B \times BTOPSPIN$, where $B = B(M)$ is the algebraic 2-type. Since $\pi := \pi_1(M)$ has odd order,

$$\Omega_4^{TOPSPIN}(K(\pi,1)) = \mathbb{Z} \oplus H_4(\pi;\mathbb{Z}),$$

and the stable homeomorphism class of M is determined by its signature and the image of the fundamental class $c_*[M] \in H_4(\pi;\mathbb{Z})$. If M and M' have isometric quadratic 2-types, then there exists an isometry $\alpha\colon s_M \cong s_{M'}$ respecting the k-invariants. We use α to identify the 2-types $B(M) \cong B(M')$, and conclude that the images of their fundamental classes agree by [26, p. 168]. Hence M and M' are spin bordant over B, and there exists a stable homeomorphism

$$h\colon M \# r(S^2 \times S^2) \cong M' \# r(S^2 \times S^2)$$

respecting the reference maps to $K(\pi,1)$. If α is a simple isometry of the quadratic 2-types, then we claim that this data will allow us to construct another stable homeomorphism h' between these manifolds, with the additional property that the induced isometry h'_* of equivariant intersection forms induces the identity on the hyperbolic summands (in fact, we will obtain $h'_* = (\alpha \oplus 1)$). With that additional property, one can attach handles to both domain and range to obtain an s-cobordism between M and M' (see [39, §4]).

To modify the homeomorphism h we proceed as follows. Let $M_r := M \# r(S^2 \times S^2)$ denote the r-fold stabilization of M. By composition, we obtain an element

$$\beta := (\alpha \oplus 1)^{-1} \circ h_* \in \mathrm{Isom}[\pi_1(M_r), \pi_2(M_r), k_{M_r}, s_{M_r}].$$

The braid diagram of [26, p. 168], combined with [26, Theorem B], now shows that $\beta = \phi_*$ for some $\phi \in \mathrm{Aut}_\bullet(M_r)$, such that ϕ is induced by an inertial s-cobordism $(W; M_r, -M_r)$. We have further assumed that α is a *simple* isometry of the quadratic 2-types. By definition, this means that the Whitehead torsion $\tau(\phi) = 0 \in \mathrm{Wh}(\mathbb{Z}\pi)$, and hence $\tau(W, M_r) = u \in \mathrm{Wh}(\mathbb{Z}\pi)$ is self-dual ($\bar{u} = u$). Note that this definition of a simple isometry is independent of the choice of h-cobordism inducing ϕ. From the exact sequences in the proof of [26, 4.1], and the fact that the discriminant map $L_6^h(\mathbb{Z}\pi) \to \widehat{H}^0(\mathbb{Z}/2; \mathrm{Wh}(\mathbb{Z}\pi))$ is surjective (since π has odd order), we can realize the self-equivalence ϕ by an s-cobordism W': if necessary, we modify our first choice by the action of $L_6^h(\mathbb{Z}\pi)$ on $\mathcal{H}(M)$. It follows that the homotopy self-equivalence induced by W' is realized by a self-homeomorphism $f\colon M_r \to M_r$. We now define $h' := h \circ f^{-1}$ and notice that $h'_* = \alpha \oplus 1$, as required. □

Stabilization and cancellation techniques can also be used effectively for manifolds with arbitrary finite fundamental groups (see [25]). For 4-manifolds, the connected sum operation gives the stabilization

$$\pi_2(M \#(S^2 \times S^2)) = \pi_2(M) \oplus \Lambda \oplus \Lambda$$

where the equivariant intersection form is stabilized by adding a hyperbolic plane

$$H(\Lambda) = (\Lambda \oplus \Lambda, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix})$$

The cancellation problem for 4-manifolds with finite fundamental group has the following optimal solution:

Theorem 5.3 ([25]). *Let M, M' be closed, oriented topological 4-manifolds with finite fundamental group. If $M = M_0 \#(S^2 \times S^2)$, and*

$$M \# r(S^2 \times S^2) \cong M' \# r(S^2 \times S^2)$$

then $M \cong M'$.

Note that even in the simply-connected case, non-isomorphic forms can become isomorphic after adding a hyperbolic plane, so the statement is best possible.

6. Fundamental groups of aspherical 2-complexes

A finitely-presented group π is *geometrically 2-dimensional* (g-dim $\pi \leq 2$) if there exists a finite aspherical 2-complex with fundamental group π. Examples of geometrically 2-dimensional groups include free groups, 1-relator groups (e.g. surface groups) and small cancellation groups [43], provided they are torsion-free, as well as many word-hyperbolic groups, see also [31, 2.3], [34, §10].

Recall that the radical $R(s_M)$ of the equivariant intersection form s_M is isomorphic to the π-module $R(\pi) := H^2(\pi; \Lambda)$. A closed oriented 4-manifold M will be called *minimal* if the equivariant intersection form on $\pi_2(M)$ vanishes, or equivalently, if $\pi_2(M) = R(s_M) \cong R(\pi)$. It turns out that a thickening of an aspherical 2-complex for π gives a minimal, smooth 4-manifold M_0 with fundamental group π, whenever g-dim $\pi \leq 2$ (see [27, Lemma 3.7]). For example, the manifolds $\#_r(S^1 \times S^3)$ and $S^2 \times \Sigma$ are minimal, where Σ denotes an oriented surface of genus ≥ 1.

In a series of papers [31], [33], [32], [34], [35], J. Hillman investigated the homotopy classification of Poincaré 4-complexes under various fundamental group assumptions. In the case of g-dim ≤ 2, the problem was reduced to the minimal case, where the homotopy classification was completed for free or surface fundamental groups (see also [49] where these cases were studied from a different viewpoint).

In recent joint work with Matthias Kreck and Peter Teichner (described below), we used the modified surgery approach to obtain classification results for topological 4-manifolds with geometrically 2-dimensional fundamental groups, up to homeomorphism (in favourable cases) or s-cobordism.

A particular nice family of examples if provided by the solvable Baumslag-Solitar groups

$$BS(k) := \{a, b \,|\, aba^{-1} = b^k\}, \quad k \in \mathbb{Z}.$$

The groups $BS(k)$ have geometrical dimension ≤ 2 because the 2-complex corresponding to the above presentation is aspherical. The easiest cases are

$$BS(0) = \mathbb{Z}, \quad BS(1) = \mathbb{Z} \times \mathbb{Z}, \quad \text{and } BS(-1) = \mathbb{Z} \rtimes \mathbb{Z},$$

and these are the only Poincaré duality groups in this family. Each $BS(k)$ is solvable, so is a "good" fundamental group for topological 4-manifolds [16]. This implies that Freedman's s-cobordism theorem is available to complete the homeomorphism classification. This had been done previously only for the three special cases above, see [17] for $BS(0)$, and [33] for $BS(\pm 1)$, using a more classical surgery approach.

Theorem 6.1 ([27, Theorem A])**.** *For closed oriented 4-manifolds with solvable Baumslag-Solitar fundamental groups, and given w_2-type and Kirby-Siebenmann invariant, any isometry between equivariant intersection forms can be realized by a homeomorphism.*

In particular, we showed that a minimal 4-manifold is unique up to homeomorphism and established some relations between the invariants in general (based in part on [55]). For fundamental groups π with $H_4(\pi; \mathbb{Z}) = 0$ we showed that the signature is determined by s_M via the formula $\text{sign}(M) = \text{sign}(s_M \otimes_\Lambda \mathbb{Z})$. This formula does not hold for arbitrary 4-manifolds, as one can see from examples of surface bundles over surfaces with nontrivial signature (but vanishing π_2).

For $\pi_1(M) = BS(k)$, type (III) can only occur if k is odd. In this case, we gave a generalization of Rochlin's formula (see [27, Corollary 6.10]):

$$KS(M) \equiv \text{sign}(M)/8 + \text{Arf}(M) \pmod 2$$

where $\text{Arf}(M) \in \mathbb{Z}/2$ is a codimension 2 Arf invariant. In contrast, for spin manifolds $KS(M) \equiv \text{sign}(M)/8 \pmod 2$.

We also proved a realization theorem for hermitian forms in this setting. If M has fundamental group $BS(k)$, then the quotient module $\pi_2(M)^\dagger := \pi_2(M)/R(s_M)$ is a finitely-generated, stably-free Λ-module, and the equivariant intersection form s_M is non-singular on this quotient. It turns out that any such hermitian form can be realized by one or two 4-manifolds.

A close inspection of the arguments shows that we used a number of special facts about the Baumslag-Solitar groups. For more general fundamental groups π, we need to assume the corresponding properties for its algebraic K-theory and L-theory.

Definition 6.2. A group π satisfies properties (W-AA) whenever

(1) The Whitehead group $\text{Wh}(\pi)$ vanishes,
(2) The assembly map $A_5 \colon H_5(\pi; \mathbb{L}_0) \to L_5(\mathbb{Z}\pi)$ is surjective.
(3) The assembly map $A_4 \colon H_4(\pi; \mathbb{L}_0) \to L_4(\mathbb{Z}\pi)$ is injective.

Note that these properties (and more) do hold whenever the group π satisfies the Farrell-Jones isomorphism conjectures [11] (see [42] for a survey of results on these conjectures).

Theorem 6.3 ([27, Theorem C]). *Let π be a geometrically 2-dimensional group satisfying properties* (W-AA). *For closed oriented 4-manifolds with fundamental group π, and given Kirby-Siebenmann invariant, any isometry between equivariant intersection forms inducing an isomorphism of w_2-types can be realized by an s-cobordism.*

The w_2-type mentioned in this statement is actually a refinement of the notion defined in Section 4, in which we now keep track of the class $w \in H^2(\pi; \mathbb{Z}/2)$ determining $w_2(M)$ in type (III).

7. Some questions

Here are a few questions and problems concerning smooth and topological 4-manifolds with non-trivial fundamental group.

(1) For a smooth 4-manifold M with geometrically 2-dimensional fundamental group, is M homeomorphic to $M_0 \# N$, where M_0 is minimal and N is simply-connected ? In other words, is the equivariant intersection form s_M always extended from the integers ?

(2) Construct distinct smooth structures on indecomposable, non-simply connected 4-manifolds. Is there a minimal 4-manifold with more than one smooth structure ?

(3) For a given group π, there exist 4-manifolds $M(\alpha)$ with $\pi_1(M) = \pi$ and $c_*[M]$ a given element $\alpha \in H_4(\pi; \mathbb{Z})$. How does the minimal possible Euler characteristic and signature of $M(\alpha)$ depend on the class α ?

(4) For a given group π, does there exists a stable range constant $c(\pi)$, with the property that a stable homeomorphism or diffeomorphism $M_1 \# r(S^2 \times S^2) \cong M_2 \# r(S^2 \times S^2)$ between manifolds with fundamental group π admits cancellation of at least one copy of $S^2 \times S^2$ (up to s-cobordism) whenever $r > c(\pi)$?

(5) Compare the actions of $\mathrm{Diff}(M)$ and $\mathrm{Homeo}(M)$ on the equivariant intersection form of a smooth 4-manifold.

Remark 7.1. There are many interesting problems related to the study the existence and uniqueness of non-free finite group actions on smooth or topological 4-manifolds. One may ask, for example, which equivariant intersection forms are realized by smooth actions of finite cyclic groups on simply-connected 4-manifolds. For topological actions there is a satisfactory picture, particularly for cyclic groups of prime order (see Edmonds [9], [10], Edmonds-Ewing [9], and McCooey [45]). For smooth actions, there are restrictions detected by equivariant gauge theory [28], [29] and the answer is interesting even for the permutation representations which arise for actions on connected sums of $\mathbf{CP}^2$'s (see [30, 1.18]). A striking contrast between smooth and topological actions is shown by the recent paper of Finstushel, Stern and Sunukjian [15], where infinite families of topologically equivalent but smoothly distinct cyclic group actions are constructed on 4-manifolds with non-trivial Seiberg-Witten invariants.

Acknowledgement. I would like to thank Matthias Kreck and Peter Teichner for many interesting and fruitful discussions on this subject. This survey is largely based on our joint work.

References

[1] A. Bak, *On modules with quadratic forms*, Algebraic K-Theory and its Geometric Applications (Conf., Hull, 1969), Springer, Berlin, 1969, pp. 55–66.
[2] H. Bass, *Algebraic K-theory*, W. A. Benjamin, Inc., New York-Amsterdam, 1968.
[3] S. E. Cappell and J. L. Shaneson, *On four dimensional surgery and applications*, Comment. Math. Helv. **46** (1971), 500–528.
[4] D. Crowley and J. Sixt, *Stably diffeomorphic manifolds and* $l_{2q+1}(Z[\pi])$, (arXiv:0808.2008v2 [math.GT]), 2008.
[5] J. F. Davis, *The Borel/Novikov conjectures and stable diffeomorphisms of 4-manifolds*, Geometry and topology of manifolds, Fields Inst. Commun., vol. 47, Amer. Math. Soc., Providence, RI, 2005, pp. 63–76.
[6] S. K. Donaldson, *An application of gauge theory to four-dimensional topology*, J. Differential Geom. **18** (1983), 279–315.
[7] ———, *Irrationality and the h-cobordism conjecture*, J. Differential Geom. **26** (1987), 141–168.
[8] ———, *The orientation of Yang-Mills moduli spaces and* 4*-manifold topology*, J. Differential Geom. **26** (1987), 397–428.
[9] A. L. Edmonds, *Aspects of group actions on four-manifolds*, Topology Appl. **31** (1989), 109–124.
[10] ———, *Periodic maps of composite order on positive definite 4-manifolds*, Geom. Topol. **9** (2005), 315–339 (electronic).
[11] F. T. Farrell and L. E. Jones, *Isomorphism conjectures in algebraic K-theory*, J. Amer. Math. Soc. **6** (1993), 249–297.
[12] R. Fintushel and R. J. Stern, *A fake* 4*-manifold with* $\pi_1 = \mathbf{Z}$ *and* $b^+ = 4$, Turkish J. Math. **18** (1994), 1–6.
[13] ———, *Knots, links, and* 4*-manifolds*, Invent. Math. **134** (1998), 363–400.
[14] ———, *Double node neighborhoods and families of simply connected 4-manifolds with* $b^+ = 1$, J. Amer. Math. Soc. **19** (2006), 171–180 (electronic).
[15] R. Fintushel, R. J. Stern, and N. Sunukjian, *Exotic group actions on simply connected smooth 4-manifolds*, (arXiv: 0902.0963 [math.GT]), 2009.
[16] M. H. Freedman, *The disk theorem for four-dimensional manifolds*, Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Warsaw, 1983) (Warsaw), PWN, 1984, pp. 647–663.
[17] M. H. Freedman and F. Quinn, *Topology of 4-manifolds*, Princeton University Press, Princeton, NJ, 1990.
[18] S. Friedl, I. Hambleton, P. Melvin, and P. Teichner, *Non-smoothable four-manifolds with infinite cyclic fundamental group*, Int. Math. Res. Not. IMRN (2007), 20 pages.
[19] R. Friedman and J. W. Morgan, *Smooth four-manifolds and complex surfaces*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)], vol. 27, Springer-Verlag, Berlin, 1994.
[20] M. Furuta, *Monopole equation and the* $\frac{11}{8}$*-conjecture*, Math. Res. Lett. **8** (2001), 279–291.
[21] R. E. Gompf, *Smooth* 4*-manifolds and symplectic topology*, Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994) (Basel), Birkhäuser, 1995, pp. 548–553.
[22] I. Hambleton and M. Kreck, *On the classification of topological* 4*-manifolds with finite fundamental group*, Math. Ann. **280** (1988), 85–104.
[23] ———, *Smooth structures on algebraic surfaces with cyclic fundamental group*, Invent. Math. **91** (1988), 53–59.

[24] ———, *Cancellation, elliptic surfaces and the topology of certain four-manifolds*, J. Reine Angew. Math. **444** (1993), 79–100.
[25] ———, *Cancellation of hyperbolic forms and topological four-manifolds*, J. Reine Angew. Math. **443** (1993), 21–47.
[26] ———, *Homotopy self-equivalences of 4-manifolds*, Math. Z. **248** (2004), 147–172.
[27] I. Hambleton, M. Kreck, and P. Teichner, *Topological* 4*-manifolds with geometrically* 2*-dimensional fundamental groups*, (arXiv:0802.0995v1 [math.GT]), 2008.
[28] I. Hambleton and R. Lee, *Perturbation of equivariant moduli spaces*, Math. Ann. **293** (1992), 17–37.
[29] ———, *Smooth group actions on definite* 4*-manifolds and moduli spaces*, Duke Math. J. **78** (1995), 715–732.
[30] I. Hambleton and M. Tanase, *Permutations, isotropy and smooth cyclic group actions on definite 4-manifolds*, Geom. Topol. **8** (2004), 475–509.
[31] J. A. Hillman, *Four-manifolds, geometries and knots*, Geometry & Topology Monographs, vol. 5, Geometry & Topology Publications, Coventry, 2002.
[32] ———, PD_4*-complexes with free fundamental group*, Hiroshima Math. J. **34** (2004), 295–306.
[33] ———, PD_4*-complexes with fundamental group of* PD_2*-group*, Topology Appl. **142** (2004), 49–60.
[34] ———, PD_4*-complexes with strongly minimal models*, Topology Appl. **153** (2006), 2413–2424.
[35] ———, *Strongly minimal* PD_4*-complexes*, Topology Appl. (2009), (to appear: arXiv:math.GT 0712.1069).
[36] M. A. Kervaire, *Smooth homology spheres and their fundamental groups*, Trans. Amer. Math. Soc. **144** (1969), 67–72.
[37] R. Kirby, *Akbulut's corks and h-cobordisms of smooth, simply connected* 4*-manifolds*, Turkish J. Math. **20** (1996), 85–93.
[38] R. C. Kirby and L. C. Siebenmann, *Foundational essays on topological manifolds, smoothings, and triangulations*, Princeton University Press, Princeton, N.J., 1977, With notes by John Milnor and Michael Atiyah, Annals of Mathematics Studies, No. 88.
[39] M. Kreck, *Surgery and duality*, Ann. of Math. (2) **149** (1999), 707–754.
[40] M. Kreck, W. Lück, and P. Teichner, *Stable prime decompositions of four-manifolds*, Prospects in topology (Princeton, NJ, 1994), Ann. of Math. Stud., vol. 138, Princeton Univ. Press, Princeton, NJ, 1995, pp. 251–269.
[41] P. B. Kronheimer and T. S. Mrowka, *Embedded surfaces and the structure of Donaldson's polynomial invariants*, J. Differential Geom. **41** (1995), 573–734.
[42] W. Lück and H. Reich, *The Baum-Connes and the Farrell-Jones conjectures in* K*- and* L*-theory*, Handbook of K-theory. Vol. 1, 2, Springer, Berlin, 2005, pp. 703–842.
[43] R. C. Lyndon and P. E. Schupp, *Combinatorial group theory*, Classics in Mathematics, Springer-Verlag, Berlin, 2001, Reprint of the 1977 edition.
[44] S. MacLane and J. H. C. Whitehead, *On the* 3*-type of a complex*, Proc. Nat. Acad. Sci. U. S. A. **36** (1950), 41–48.
[45] M. P. McCooey, *Symmetry groups of four-manifolds*, Topology **41** (2002), 835–851.
[46] J. W. Morgan and Z. Szabó, *Complexity of* 4*-dimensional h-cobordisms*, Invent. Math. **136** (1999), 273–286.
[47] P. Ozsváth and Z. Szabó, *Holomorphic disks and three-manifold invariants: properties and applications*, Ann. of Math. (2) **159** (2004), 1159–1245.
[48] ———, *Holomorphic disks and topological invariants for closed three-manifolds*, Ann. of Math. (2) **159** (2004), 1027–1158.
[49] M. Pamuk, *Homotopy self-equivalences of four-manifolds*, Ph.D. thesis, McMaster University, 2008.
[50] J. Park, *Simply connected symplectic 4-manifolds with* $b_2^+ = 1$ *and* $c_1^2 = 2$, Invent. Math. **159** (2005), 657–667.
[51] H.-G. Quebbemann, *Definite lattices over real algebraic function domains*, Math. Ann. **272** (1985), 461–475.

[52] J. T. Stafford, *Absolute stable rank and quadratic forms over noncommutative rings*, K-Theory **4** (1990), 121–130.
[53] C. H. Taubes, *The Seiberg-Witten invariants and symplectic forms*, Math. Res. Lett. **1** (1994), 809–822.
[54] P. Teichner, *Topological four manifolds with finite fundamental group*, Ph.D. thesis, University of Mainz, 1992.
[55] ———, *On the signature of four-manifolds with universal covering spin*, Math. Ann. **295** (1993), 745–759.
[56] L. N. Vaserštein, *Stabilization of unitary and orthogonal groups over a ring with involution*, Mat. Sb. (N.S.) **81 (123)** (1970), 328–351.
[57] C. T. C. Wall, *Diffeomorphisms of* 4*-manifolds*, J. London Math. Soc. **39** (1964), 131–140.
[58] ———, *On simply-connected* 4*-manifolds*, J. London Math. Soc. **39** (1964), 141–149.
[59] ———, *Surgery on compact manifolds*, second ed., American Mathematical Society, Providence, RI, 1999, Edited and with a foreword by A. A. Ranicki.
[60] J. H. C. Whitehead, *Simplicial spaces, nuclei and m-groups*, Proc. Lond. Math. Soc. **45** (1939), 243–327.
[61] E. Witten, *Monopoles and four-manifolds*, Math. Res. Lett. **1** (1994), 769–796.

Department of Mathematics & Statistics
McMaster University
Hamilton, ON L8S 4K1, Canada
E-mail address: hambleton@mcmaster.ca

Proceedings of 15$^{\text{th}}$ Gökova
Geometry-Topology Conference
pp. 151–169

Exotic embeddings of $\#6\mathbb{R}\mathrm{P}^2$ in the 4-sphere

Sergey Finashin

Abstract. We construct an infinite sequence of smooth embeddings of $\#6\mathbb{R}\mathrm{P}^2$ in S^4, which are all ambient homeomorphic, but pairwise ambient non-diffeomorphic. The double covers of S^4 ramified along these surfaces form a family of the exotic $\mathbb{C}\mathrm{P}^2\#5\overline{\mathbb{C}\mathrm{P}^2}$ constructed by Park, Stipsicz and Szabó.

1. Introduction

The main goal of this paper is the following result.

Theorem 1.1. *For any $k \geq 6$ there exists an infinite family of smoothly embedded surfaces $F_i \subset S^4$, $i = 1, 2, \dots$, homeomorphic to $F = \#k\mathbb{R}\mathrm{P}^2$ (connected sum of k copies of $\mathbb{R}\mathrm{P}^2$) and with the normal Euler number $F^2 = 2k - 4$, such that*

(1) *the pairs (S^4, F_i) are all homeomorphic; the ambient homeomorphisms can be assumed to be diffeomorphisms in some neighborhoods of F_i;*
(2) *(S^4, F_i) are all pairwise non-diffeomorphic.*

Theorem 1.1 improves the result of [FKV1]-[FKV2], where a similar family of $F_i \subset S^4$ was constructed for $k = 10$. Our construction of F_i is based on similar ideas and makes use of the examples of exotic $\mathbb{C}\mathrm{P}^2\#5\overline{\mathbb{C}\mathrm{P}}^2$ in [PSS]. More precisely, our goal can be stated as follows.

Theorem 1.2. *There exists an infinite family of smoothly embedded surfaces $F_i \subset S^4$, $i = 1, 2, \dots$ which are all homeomorphic to a connected sum $\#6\mathbb{R}\mathrm{P}^2$, such that $\pi_1(S^4 \smallsetminus F_i) = \mathbb{Z}/2$, and the double covers $X_i \to S^4$ branched along F_i provide an infinite family of exotic $\mathbb{C}\mathrm{P}^2\#5\overline{\mathbb{C}\mathrm{P}}^2$ constructed in* [PSS].

Recall the well-known diffeomorphism $\mathbb{C}\mathrm{P}^2/\mathrm{conj} = \mathrm{S}^4$. The image of $\mathbb{R}\mathrm{P}^2$ in the quotient-space S^4 represents an isotopy class of *unknotted embeddings* of $\mathbb{R}\mathrm{P}^2$ with the normal Euler number -2. Another isotopy class of unknotted embeddings (with the normal Euler class 2) is obtained by reversing the orientation of S^4. It is represented by the image of $\mathbb{R}\mathrm{P}^2$ in $S^4 = \overline{\mathbb{C}\mathrm{P}}^2/\mathrm{conj}$, which will be denoted $\overline{\mathbb{R}\mathrm{P}}^2 \subset S^4$. A non-orientable surface $F \subset S^4$ will be called *unknotted* if it splits into an ambient connected sum of such unknotted embeddings, that is $F = p\mathbb{R}\mathrm{P}^2\#q\overline{\mathbb{R}\mathrm{P}}^2$.

Key words and phrases. Knotted surfaces, equivariant surgery, exotic smooth structures.

1.1. Theorem 1.2 implies Theorem 1.1

It is proven in [PSS] that X_i are pairwise non-diffeomorphic. This implies that the pairs (S^4, F_i) are non-diffeomorphic to each other. It is known moreover that for any $\ell \geq 1$, $X_i \# \ell \overline{\mathbb{C}\mathrm{P}}^2$, $i = 1, 2, \ldots$, remain pairwise non-diffeomorphic, which implies that $F_i \# \ell \overline{\mathbb{R}\mathrm{P}}^2 \subset S^4$ are also all ambient non-diffeomorphic to each other.

The values $F_i^2 = 8$ of the normal Euler numbers for F_i in Theorem 1.2 can be obtained from the the signature formula $\sigma(X_i) = -4 = 2\sigma(S^4) - \frac{1}{2}F_i^2$. For the connected sum $F_i \# \ell \overline{\mathbb{R}\mathrm{P}}^2$ the Euler number becomes $8 + 2\ell$, that is $2k - 4$ where $k = \ell + 6$.

The obstructions for pairwise ambient homeomorphism of surfaces F_i were analyzed in [FKV2], Proposition 3, where it was shown that they belong to some finite set (it is crucial for this proof that $\pi_1 = \mathbb{Z}/2$). This implies that we can choose an infinite subsequence of pairwise ambient homeomorphic surfaces F_i in the infinite sequence of pairwise non-diffeomorphic ones (which are covered by the corresponding exotic $\mathbb{C}\mathrm{P}^2 \# 5\overline{\mathbb{C}\mathrm{P}}^2$). □

Remark 1.1. In [K], it is shown that the finite ambiguity observed in [FKV2] for the exotic $\#10\mathbb{R}\mathrm{P}^2$ actually vanishes. This means that all the examples of embedded $\#10\mathbb{R}\mathrm{P}^2$ that were constructed in [FKV2] are actually homeomorphic to an unknotted $\mathbb{R}\mathrm{P}^2 \# 9\overline{\mathbb{R}\mathrm{P}}^2$. Formally speaking the arguments in [K] concern only the case of $\#10\mathbb{R}\mathrm{P}^2$, and it is not clear if they can be appropriately modified for our case of $\#6\mathbb{R}\mathrm{P}^2$. If it were possible, then all the examples of F_i in Theorem 1.2 would be actually ambient homeomorphic to an unknotted $\mathbb{R}\mathrm{P}^2 \# 5\overline{\mathbb{R}\mathrm{P}}^2$ (without a need to select a subsequence).

1.2. Scheme of the proof of Theorem 1.2

There are several alternative constructions of an exotic $\mathbb{C}\mathrm{P}^2 \# 5\overline{\mathbb{C}\mathrm{P}}^2$ in [PSS], and the one suitable for us is obtained by some surgery from a rational elliptic surface, X, with a fiber of type $\mathbb{I}_8$. The first step is a double node neighborhood knot surgery on X, which yields a 4-manifold X_K containing a nodal *pseudo-section*. Next, X_K is blown up at several points so that we obtain a suitable chain of spheres, $C = C_1 \cup \cdots \cup C_k$, which can be rationally blown down on the last step. Our aim is to perform these constructions equivariantly.

In Section 2, we construct a special example of a rational elliptic surface, X, with a fiber $\mathbb{I}_8$, which is defined over reals, and thus has an involution, c, of the complex conjugation. It is essential for the further constructions that the components of the $\mathbb{I}_8$-fiber as well as the four remaining singular fishtail fibers are all *real* (i.e., invariant under the complex conjugation). In Section 4, we observe that a certain non-singular real fiber, T, which is constructed in Section 2, can be used for an equivariant double node knot surgery. We check that the nodal pseudo-section S_K obtained after such a surgery can be chosen invariant under the involution. Following the construction in [PSS], we blowup several points, which turn out to be all real in our example of X. Finally we obtain as in [PSS] a chain of spheres $C = C_1 \cup \cdots \cup C_n$, whose components are all c-invariant.

In Section 3, we discuss equivariant blowing down such chains C. It is crucial for us that the quotient X/c turns out to be S^4 and that the quotient by the involution remains the same as we modify the 4-manifold and the involution. So, all the involved equivariant transformations of X (knot surgery, blowup at a real point and rational blowdown) just modify the fixed point set F in the quotient S^4. Another crucial fact is that the fundamental group $\pi_1(S^4 \smallsetminus F) = \mathbb{Z}/2$ is preserved under these modifications of F.

More precisely, π_1 is preserved after a rational blowdown of C if we put a certain condition on C. This condition is satisfied for two of the configurations proposed in Proposition 2.5 of [PSS]: for $C_{79,44}$ and for $C_{89,9}$, which are the chains

$$(-2,-5,-11,-2,-2,-2,-2,-2,-2,-3,-2,-2,-3), \quad \text{and}$$

$$(-10,-11,-2,-2,-2,-2,-2,-2,-3,-2,-2,-2,-2,-2,-2,-2,-2).$$

Following the construction of [PSS] in the equivariant setting, we obtain a certain 4-manifold $\widehat{X}$ homeomorphic but not diffeomorphic to $\mathbb{C}\mathrm{P}^2\#5\overline{\mathbb{C}\mathrm{P}}^2$, with an involution, $\widehat{c} : \widehat{X} \to \widehat{X}$. In the quotient $\widehat{X}/\widehat{c} = S^4$ there is a surface $F \subset S^4$ which is the fixed point set of $\widehat{c}$. Observing that F is connected, non-orientable (because $F^2 = 8 \neq 0$), and estimating its Euler characteristic $\chi(F) = 2\chi(S^4) - \chi(\widehat{X})$, we deduce that it is $\#6\mathbb{R}\mathrm{P}^2$.

A sequence of the twist-knots K_i that can be used for the knot surgery on the first step of the construction (see Figure 6a)) yields a sequence $\widehat{X}_i$ of exotic $\mathbb{C}\mathrm{P}^2\#5\overline{\mathbb{C}\mathrm{P}}^2$ with involutions $\widehat{c}_i$, and a sequence of surfaces $F_i \subset S^4$ required for Theorem 1.2.

In the last section we pass once more through all the steps and verify some conditions which are required to be satisfied (for instance, to apply Lemma 3.2) in the particular case of $C_{79,44}$-chain. For convenience of the reader we included in the last section a review of the construction of this chain in $\mathbb{C}\mathrm{P}^2\#18\overline{\mathbb{C}\mathrm{P}}^2$ which was given in [PSS].

2. Real rational elliptic surfaces with special singular fibers

2.1. Double planes ramified along quartics

It is well known that the double covering over $\mathbb{C}\mathrm{P}^2$ branched along a non-singular quartic, $A \subset \mathbb{C}\mathrm{P}^2$, yields a del Pezzo surface $X_A = \mathbb{C}\mathrm{P}^2\#7\overline{\mathbb{C}\mathrm{P}}^2$. A pencil of lines, $L_s \subset \mathbb{C}\mathrm{P}^2$, centered at $x \in \mathbb{C}\mathrm{P}^2$, is covered by an elliptic pencil, $T_s \subset \mathbb{C}\mathrm{P}^2\#7\overline{\mathbb{C}\mathrm{P}}^2$, whose singular fibers cover the lines tangent to A. Blowing up the pull-back of x in $\mathbb{C}\mathrm{P}^2\#7\overline{\mathbb{C}\mathrm{P}}^2$, we obtain an elliptic fibration, $p : X \to \mathbb{C}\mathrm{P}^1$.

Assume that the quartic A has a singular point, $y \in A$, of the type $\mathbb{A}_n$ (by definition at such a singularity A is locally defined as $z_1^{n+1} + z_2^2 = 0$), and the basepoint x is generic. Then X_A also has a singularity of type $\mathbb{A}_n$ (i.e., locally defined as $z_1^{n+1} + z_2^2 + z_3^2 = 0$) and after its resolution we obtain an elliptic fibration $p : X \to \mathbb{C}\mathrm{P}^1$ with a singular fiber T_0 of the type $\mathbb{I}_{n+1}$ (in Kodaira's classification), i.e., a cyclic chain of (-2)-spheres. The fiber T_0 is the pull-back of the line $L_0 \subset \mathbb{C}\mathrm{P}^2$ passing through x and y. If in addition L_0 is tangent to A at the basepoint $x \in A$, then the fiber T_0 is of the type $\mathbb{I}_{n+3}$. We explain it in detail for the example of $n = 5$ that we will need.

Namely, in the case of our interest we obtain $\mathbb{I}_8$-fiber T_0 if we choose a quartic $A = A_1 \cup A_3$ that splits into a cubic, A_3, and a real line, A_1, tangent to A_3 at its inflection point, y (see Figure 1a)). The corresponding line L_0 should pass through y and be tangent to A_3 at some other point, x, which will be the center of the pencil of the lines L_s. The fiber $\mathbb{I}_8$ can be seen as the double cover of the chain of spheres $(-1,-2,-2,-2,-1)$ with four points of ramification: two points on each (-1)-curve. This chain appears in $\mathbb{C}\mathrm{P}^2\#5\overline{\mathbb{C}\mathrm{P}}^2$ as we resolve $\mathbb{A}_5$-singularity at y by a triple blowup, which gives a $(-1,-2,-2)$-chain, and then blowup the intersection of L_0 with A at x twice. The line L_0 (blown up three times) gives the third (-2)-curve in the chain, and the exceptional curve of the second blowup at x is the last (-1)-curve of the chain. The exceptional curve of the first blowup at x is (-2)-curve $E_x^* \subset \mathbb{C}\mathrm{P}^2\#5\overline{\mathbb{C}\mathrm{P}}^2$ (since the second blowup is in its point); it is a section, which belongs to the ramification locus of the double cover of $\mathbb{C}\mathrm{P}^2\#5\overline{\mathbb{C}\mathrm{P}}^2$ by the elliptic surface $X = \mathbb{C}\mathrm{P}^2\#9\overline{\mathbb{C}\mathrm{P}}^2$. So, E_x^* lifts to a (-1)-curve, $E_x \subset X$, which is a section of our elliptic fibration. The two ramification points on the first (-1)-component of the $(-1,-2,-2,-2,-1)$-chain are intersections with the two branches of A at y. On the last (-1)-component, one of the ramification points appear from the branch of A at x, and the other is the intersection point with E_x^*.

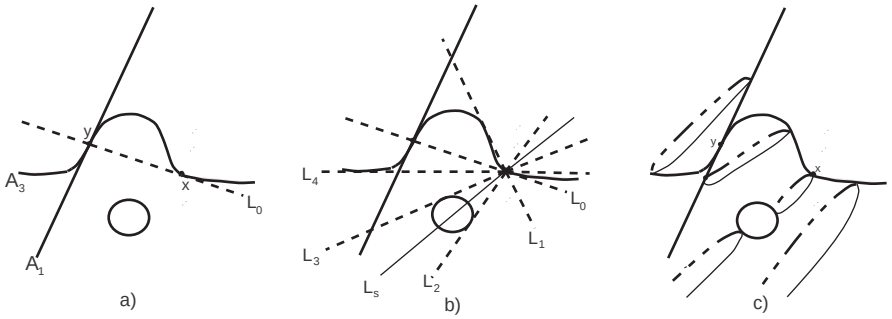


FIGURE 1. a) Quartic $A = A_1 \cup A_3$ and the basepoint x. b) The tangent lines L_i, $i = 1,2,3,4$ of the pencil and line L_s corresponding to the fiber $T = T_s$. c) The real locus of the double plane X_A.

2.2. Construction of a special real elliptic fibration

Consider the double covering $q : X_A \to \mathbb{C}\mathrm{P}^2$ ramified along a degree $2n$ curve $A \subset \mathbb{C}\mathrm{P}^2$ having an equation $f = 0$ with real coefficients. It is well-known (and trivial) that the complex conjugation in $\mathbb{C}\mathrm{P}^2$ can be lifted to X_A in two ways. These two liftings correspond to two real algebraic models of X_A, namely, the ones defined by a weighted homogeneous equation $y^2 = \pm f(x_0, x_1, x_2)$ in the quasi-projective space $P_{1,1,1,n}$ with the coordinates x_0, x_1, x_2, y of weights $1, 1, 1, n$. The corresponding involutions $c_\pm : X_A \to X_A$, induced

from the complex conjugation in $P_{1,1,1,n}$, have fixed point sets $\mathrm{Fix}(c_\pm|_{X_A})$ which are projected by q to the regions $\mathbb{R}\mathrm{P}^2_{\pm\mathrm{f}} = \{\mathrm{x} \in \mathbb{R}\mathrm{P}^2 \mid \pm\mathrm{f(x)} \geq 0\}$ bounded by the curve $A_\mathbb{R} = A \cap \mathbb{R}\mathrm{P}^2$. It is convenient to specify the choice of one of these two involution by referring to the corresponding region $\mathbb{R}\mathrm{P}^2_\mathrm{f}$ (the projection of the real locus of X_A).

In our example of a real quartic $A = A_1 \cup A_3$, we choose the region $\mathbb{R}\mathrm{P}^2_\mathrm{f}$ as is shown on Figure 1c) and the corresponding involution $c_A : X_A \to X_A$ whose fixed point set is $q^{-1}(\mathbb{R}\mathrm{P}^2_\mathrm{f})$. As it was explained, blowing up the $\mathbb{A}_5$-singularity and then the two infinitely near base-points of X_A, we obtain a *real elliptic fibration*, $p : X \to \mathbb{C}\mathrm{P}^1$, endowed with an involution $c : X \to X$ commuting with p. Let $F = \mathrm{Fix}(c)$ denote its fixed point set.

Lemma 2.1. *The real elliptic fibration $p : X \to \mathbb{C}\mathrm{P}^1$ constructed above has the following properties.*

(1) *X contains a real singular fiber $T_0 = C_1 \cup \cdots \cup C_8$ of the type $\mathbb{I}_8$, whose components C_i are c-invariant.*
(2) *X contains 4 other c-invariant singular fibers, T_i, $i = 1, 2, 3, 4$, which are ordinary fishtails.*
(3) *The elliptic fibration $p : X \to \mathbb{C}\mathrm{P}^1$ admits a c-invariant section.*
(4) *X can be blown down to $\mathbb{C}\mathrm{P}^2$, so that each of the nine consecutively contracted exceptional curves is real. This transforms the non-singular real fibers T_s to non-singular real cubics.*
(5) *$X/c = S^4$.*
(6) *If the singular fibers T_i, $i = 0, 1, 2, 3, 4$ are the pre-images of the tangent lines L_i on Figure 1b), then a non-singular real fiber $T = T_s$ chosen between T_2 and T_3 has real locus, $T \cap F$, formed by two connected components, as is shown on Figure 2c). The two vanishing curves in T, which are contracted as T is degenerated into the singular fibers T_2 and T_3, are isotopic. A vanishing curve from this isotopy class can be chosen c-invariant, and so that c reverses its orientation.*
(7) *The complement $F \smallsetminus (F \cap T)$ is connected.*

Proof. As is explained above, our fiber T_0 is of type $\mathbb{I}_8$ and is a double cover of the chain $(-1, -2, -2, -2, -1)$. All the components of the chain have real structures equivalent to that of $\mathbb{C}\mathrm{P}^1$, since we blowup only at real points. The branching points on the (-1)-curves are also real, which implies that all the components of $\mathbb{I}_8$ are c-invariant and (1) is proven.

Fibers T_i in (2) cover the tangent lines L_i shown on Figure 1b). Since L_i are real lines with ordinary tangency, the fibers T_i are c-invariant fishtails.

The section $E_x \subset X$, which was already described justifies (3). There are also other real (i.e., c-invariant) sections that we need for proving (4). For instance there is the proper transform, $A_1^* \subset X$, of the line A_1 (it has self-intersection -2 in $\mathbb{C}\mathrm{P}^2\#5\overline{\mathbb{C}\mathrm{P}}^2$ after 3 blowups and becomes (-1)-curve after lifting to X). The proper transform of a real line $L' \subset \mathbb{C}\mathrm{P}^2$ passing through y and tangent to A_3 (see Figure 2a)) splits into two components, L'_1, L'_2, which are both real sections of p. Another tangent line L'' shown on Figure 2a) gives similarly components L''_1 and L''_2, which are also real sections.

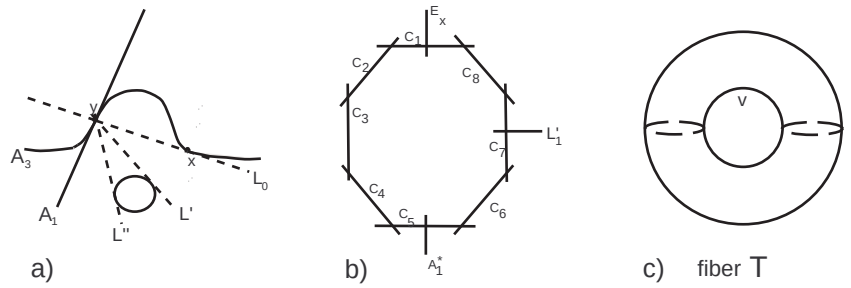


FIGURE 2. a) Tangent lines L' and L''. b) Fiber T_0 and 3 disjoint real sections E_x, A_1^*, and L_1'. c) The fixed point set of c dividing the fiber T, and the vanishing cycle v.

Let us numerate cyclically the components $C_1, \dots, C_8$ of the fiber T_0, so that C_1 is the double cover of the last (-1)-component in the chain $(-1,-2,-2,-2,-1)$ considered above, and thus section E_x intersects T_0 at a point of C_1. Note that the proper image of A_1 intersects this chain at the first (-1)-component, and the proper images of L' and L'' at the middle (-2)-component. This implies that A_1^* intersects C_5, and the curves L_j', L_j'', $j = 1, 2$, intersect C_7 and C_3. Without loss of generality, we may suppose that L_1' intersects C_7, like it is shown on Figure 2b). We have $C_i^2 = -2$, $E_x^2 = (L_1')^2 = (A_1^*)^2 = -1$, and so we can blow down consecutively nine real exceptional curves, for example in the following order: E_x, C_1, C_2, C_3, C_4, A_1^*, L_1', C_7, and C_6. Such a blowdown must give $\mathbb{C}\mathrm{P}^2$, which proves (4). After blowing down the remaining components, C_5 and C_8, will represent a line and a conic in $\mathbb{C}\mathrm{P}^2$ (since their self-intersection indices are 1 and 4 respectively), and non-singular real fibers become non-singular real cubics.

We can deduce (5) from (4) using that $\mathbb{C}\mathrm{P}^2/\mathrm{conj} = \mathrm{S}^4$, which implies that a blowup at a real point effects to the quotient as a connected sum with $S^4 = \overline{\mathbb{C}\mathrm{P}}^2/\mathrm{conj}$, and thus, differential-topologically does not change the quotient.

Inspecting Figure 1c) we observe that the fixed point set $T \cap F$ of the complex conjugation acting on T has two connected components, as it is shown on Figure 2c). The types of the vanishing cycles on T are determined by the index of the real critical points of the projection $F = \mathrm{Fix}(c) \to \mathbb{R}\mathrm{P}^1$ (restriction of p). From Figure 1c) we can see saddle points at the fibers T_2 and T_3, and so each of the vanishing cycles have two real points (the real vanishing cycle S^0). This implies that the complex vanishing cycles at the non-singular fiber T_s for the both saddle points are isotopic to the curve v on Figure 2c). This is because up to isotopy, there is a unique c-invariant curve on a torus T_s which intersects once each component of its real locus. Namely, c should act on curve v as a reflection, and thus the homology class $[v]$ belongs to the one-dimensional (-1)-eigenspace of the c-action in $H_1(T_s)$. This proves (6).

Property (7) is clear from Figure 1c) if we take into account connectedness of the real locus of our real form of singularity $\mathbb{A}_5$ (local model $x^6 - y^2 + z^2 = 0$) after its resolution. Namely, this connectedness implies that the two branches of the surface at the singularity shown on Figure 1c) become connected after resolution. □

3. Equivariant rational blowdown

3.1. Rational blowdown surface surgery

Let X denote a smooth 4-manifold with *a chain of spheres* $C = C_1 \cup \cdots \cup C_k \subset X$, which intersect each other consecutively and transversely, so that their dual weighted graph is a chain-tree shown on the bottom of Figure 3. We can (and always will) orient the spheres so that their intersection indices are positive. A regular neighborhood, $N(C)$, of C is a plumbing 4-manifold, P_C, corresponding to this weighted graph.

Certain chains C can be *rationally blown down*, i.e., we can remove $N(C) = P_C$ from X and replace it by some rational homology ball, Q_C, with the same boundary $\partial Q_C = \partial N(C)$. This gives a new 4-manifold $\widehat{X} = X' \cup Q_C$, where $X' = X \smallsetminus N(C)$.

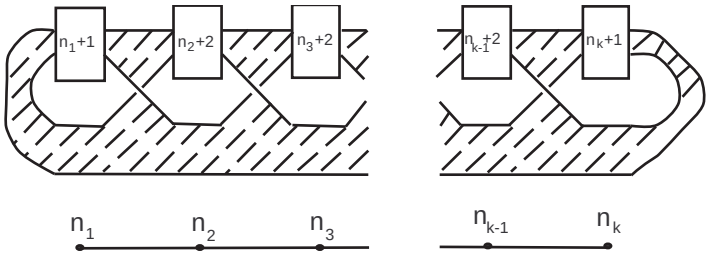


FIGURE 3. The plumbing surface F_C described by a chain-tree can be presented as the span-surface of a two-bridge link diagram (the interior of F_C is pushed inside D^4). The numbers in the boxes count the half-twists.

It is well known and easy to see that P_C can be described as the double cover over D^4 branched along a surface, F_C, obtained by plumbing of the unknotted bands, $F_{n_i} \subset D^4$, $i = 1, \ldots, k$, where n_i stands for the framing of the band (number of its half-twists which is taken with sign "$-$" in the case of left-hand half-twists). Such a plumbing surface can be sketched as is shown on Figure 3.

As it is observed in [FS1], Q_C is the double cover of D^4 branched along another surface, $R_C \subset D^4$, bounded by the same link as F_C, $L_C = \partial R_C = \partial F_C$. More details about the surface R_C can be extracted from [CH], and we will only mention the idea and illustrate it by an example. We start with a plumbing surface $F_{C'}$ representing some chain $(n_1, \ldots, n_k)$, which can be blown down to the chain (0), like for instance the chain $(-1, -1)$. The boundary link $L_{C'} = \partial F_{C'}$ is a 2-component unlink, as the boundary

of an untwisted band F_0. For instance Figure 4c) illustrates an example of $L_{C'}$ for the chain $(-1,-1)$. There is a $(+1)$-twisted band in the middle of $F_{C'}$ on Figure 4c), and cutting it yields another plumbing surface $F_C = F_{-4}$ on Figure 4a). In general, cutting a similar band replaces a pair of consecutive numbers n_i, n_{i+1} in the chain C' by one number $n_i + n_{i+1} - 2$ in the chain C. On the other hand, $L_{C'}$ as an unlink bounds two disjoint disks $D_1 \sqcup D_2$, and the operation of smoothing the $(+1)$-crossing on Figure 4c) can be interpreted as "a ribbon move", that is gluing a ribbon to the boundary of a pair of disjoint unknotted discs, $D_1 \sqcup D_2$. Such a ribbon may intersect the discs, but as we push in the interior of the disks from S^3 inside D^4, we obtain an embedded surface $R_C \subset D^4$ bounding $L_C = \partial F_C$. If the ribbon connects ∂D_1 with ∂D_2, then L_C is a ribbon knot and R_C is a disk embedded in D^4. The ribbon may also connect the boundary of one of the discs, say D_1, to itself, then R_C is a disjoint union of the disc D_2 with a Möbius band (one can show that the band cannot be orientable, because ∂F_C cannot have 3 boundary components as a 2-bridge link). On Figure 4b) the disk D_2 bounding one component is pushed inside D^4 and so is disjoint from the Möbius band bounding another component (the band shaded on Figure 4b)).

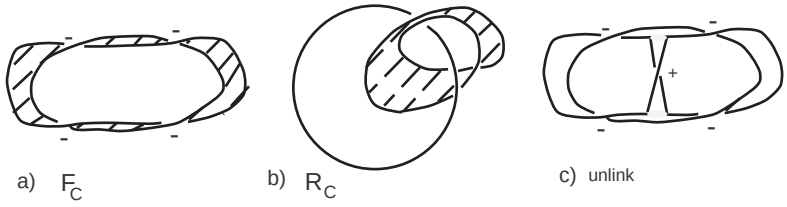


FIGURE 4. a) A band $F_C = F_{-4}$ with the four negative half-twists (marked with the signs "−"). b) Link $L_C = \partial F_C$ bounds also R_C (a disc and a band). c) An unlink $L_{C'}$ which yields L_C after a ribbon move.

To obtain similarly the surface R_C for $C = C_{79,44}$ we should start with a chain $C' = (-2,-5,-8,-1,-2,-2,-2,-2,-2,-2,-3,-2,-2,-3)$.

Lemma 3.1. *Consider a rational blowdown of a chain of spheres $C \subset X$, which yields $\widehat{X} = X' \cup Q_C$. Assume that X is endowed with an orientation preserving involution $c : X \to X$, which keeps each of the components, $C_i \subset C$ invariant, and reverses its orientation, so that* $\mathrm{Fix}(c) \cap C_i \neq \emptyset$. *Then, if $N(C)$ is chosen c-invariant, the rational blowdown can be made equivariantly, which yields an involution $\widehat{c} : \widehat{X} \to \widehat{X}$.*

Such a blowdown gives the same quotient $Y = X/c = \widehat{X}/\widehat{c}$. $F = \mathrm{Fix}\, c$ and $\widehat{F} = \mathrm{Fix}(\widehat{c})$ descend to Y and give the same locus $F' = F \cap Y' = \widehat{F} \cap Y'$, inside $Y' = X'/c$. The modification of the fixed point set is supported in a ball $N(C)/c = D^4$, where $F \cap D^4$

is isotopic to the plumbing surface F_C. The piece of surface $\widehat{F} \cap D^4$ is isotopic to the surface R_C.

Proof. Under these assumptions, $c|_{N(C)}$ is equivalent to the deck transformation of the double branched covering $N(C) \to D^4$. The involution $\widehat{c}$ just extends the involution $c|_{X \setminus N(C)}$ to Q_C as the deck transformation of the branched covering $Q_C \to D^4$. □

We say that $(\widehat{X}, \widehat{c})$ is obtained by an *equivariant rational blowdown* from (X, c).

3.2. Characteristic sub-configurations

It is trivial to see that the number of components of a compact surface with non-empty boundary is greater by one than the nullity of its intersection form with $\mathbb{Z}/2$-coefficients. On the other hand, Figure 3 presents a 2-bridge link, and such links may have at most two connected components. Thus, the link $L_C = \partial F_C$ bounded by the plumbing of bands, F_C, has one component if the intersection matrix $M_C = (C_i \circ C_j)$ is non-singular modulo 2 (has odd determinant), and has two components otherwise (if the nullity over $\mathbb{Z}/2$ is 1).

We say that the union of some components C_i forms a *characteristic sub-configuration*, $W \subset C$, and call the corresponding components C_i *characteristic components* if the fundamental class $[W] \in H_2(C; \mathbb{Z}/2)$ is a Wu element of the intersection form $(C_i \circ C_j)$, that is $C_i^2 = C_i \circ W \pmod 2$ for all $i \in \{1, \dots, k\}$. If the matrix M_C is non-singular modulo 2, then a characteristic sub-configuration W does exist and is unique. Otherwise, a characteristic sub-configuration W still exists, but is not unique: it is trivial to see that another sub-configuration $W' \subset C$ is also characteristic if and only if the sum $[W] + [W'] \in H_2(C; \mathbb{Z}/2)$ belongs to the null-space of the intersection form modulo 2.

Remark 3.1. To find a characteristic sub-configuration of C (and to prove in turn its existence) one can first reduce modulo 2 the corresponding chain of integers, then blow it down (mod 2) to obtain a chain of zeros, whose characteristic sub-configuration is trivial. It is an easy exercise to prove that each blowup (mod 2) preserves all the previous characteristic components, and that the new exceptional component is also characteristic if and only if it is not a neighbor of another characteristic component. (One hint: observe first that a pair of characteristic components cannot be adjacent.)

To illustrate this method let us consider a chain $(-3, -2, -3, -4)$, which has modulo 2 reduction $(1, 0, 1, 0)$. Blowing down (mod 2) the component C_1 gives $(1, 1, 0)$, and the subsequent blowdown of C_2 gives $(0, 0)$, which implies that the remaining components C_3 and C_4 are not characteristic. As we blow down C_2 its neighbor C_3 is not characteristic, and so C_2 is characteristic. On the contrary, C_1 is not characteristic, because its neighbor C_2 is characteristic. Thus, the characteristic sub-configuration of the given chain includes only C_2. Another example illustrating this method will appear in 5.3

Remark 3.2. It is also simple to determine W using an arbitrarily chosen orientation of the link diagram of L_C shown on Figure 5. Namely, we should include component C_i in W if and only if the opposite sides of the band F_{n_i} are co-directed, as is shown on Figure 5. This description follows from the fact that $W \cap F_C$ realizes the first Stiefel-Whitney class

$w_1(F_C)$ (because w_1 is dual to the Wu element of the intersection form in $H_1(F_C;\mathbb{Z}/2)$, which is described by the matrix M_C modulo 2).

3.3. Commutativity of π_1 after rational blowdowns

Suppose that a sphere $C_0 \subset X$ extends the chain C to a longer chain $\widetilde{C} = C_0 \cup C_1 \cup \cdots \cup C_n$. This means that C_0 is a c-invariant sphere (like the others C_i) which intersects C_1 transversely at a single point and does not intersect C_i, if $i > 1$.

Lemma 3.2. *Assume that*

(1) *the link L_C is a knot,*
(2) *a characteristic sub-configuration $W \subset C$ is not characteristic for $\widetilde{C}$.*

Then $\pi_1(Y \smallsetminus \widehat{F}) = \pi_1(Y \smallsetminus F)$.

Remark 3.3. According to the definitions, the second assumption of the Lemma means that C_1 is not included into W if $n_0 = C_0^2$ is odd, and $C_1 \subset W$ if n_0 is even.

Proof. Applying the Van Kampen theorem, we see that $\pi_1(Y \smallsetminus F) = G' *_{G_L} G_C$, where $G' = \pi_1(Y' \smallsetminus F')$, $G_C = \pi_1(D^4 \smallsetminus F_C)$, and $G_L = \pi_1(S^3 \smallsetminus L_C)$. Similarly, $\pi_1(Y \smallsetminus \widehat{F}) = G' *_{G_L} \widehat{G}$, where $\widehat{G} = \pi_1(D^4 \smallsetminus R_C)$.

The plan of the proof is to observe that the inclusion homomorphisms $G_L \to G_C$, $G_L \to \widehat{G}$ are epimorphisms, and that their kernels, K and $\widehat{K}$, vanish under the inclusion homomorphism $G_L \to G'$. This implies that the homomorphisms $G' \to G' *_{G_L} G_C$ and $G' \to G' *_{G_L} \widehat{G}$ are isomorphisms.

First of all, note that the upper Wirtinger presentation for the link L_C whose diagram is shown on Figure 3 implies that its group G_L is generated by two elements a, b, presented by the loops around the overpasses ℓ_a and ℓ_b shown in Figure 5.

The homomorphism $G_L \to G_C$ is epimorphic because G_C is a cyclic group (since F_C is a connected span-surface for L_C, whose interior is pushed out from S^3 inside D^4). Inspecting the homology, we see that $G_C = \mathbb{Z}$ is obtained from G_L by adding the relation $a = b$ in the case of orientable surface F_C. If F_C is non-orientable, then $G_C = \mathbb{Z}/2$ is obtained from G_L by adding two relations: $a = b$ and $a = b^{-1}$, which generate K.

The homomorphism $G_L \to \widehat{G}$ is also epimorphic, because R_C is a ribbon-surface. The kernel $\widehat{K}$ is contained in the commutator subgroup $[G_L, G_L]$, which is the kernel of the product map $G_L \to \widehat{G} \to H_1(D^4 \smallsetminus R_C) = H_1(S^3 \smallsetminus L_C) = \mathbb{Z}$, where the latter two equalities are due to our assumption that L_C is connected, and so R_C is a disc. Thus, $[G_L, G_L]$ is generated by the relation $a = b$ if overpasses ℓ_a and ℓ_b on Figure 5 inherit co-directed orientation from L_C, and by the relation $a = b^{-1}$ if these overpasses inherit opposite orientations.

Showing that the images of a, b under the inclusion homomorphisms $G_L \to G'$ (for which we keep the same notation a, b) satisfy both relations $a = b$ and $a = b^{-1}$ will complete the proof.

One of these relations comes from a regular neighborhood $N(D_0)$ of the disc $D_0 = C_0/c$ in Y. Note that $H = N(D_0) \cap Y'$ is a 4-ball containing an unknotted disc $F_H = F' \cap H$, so

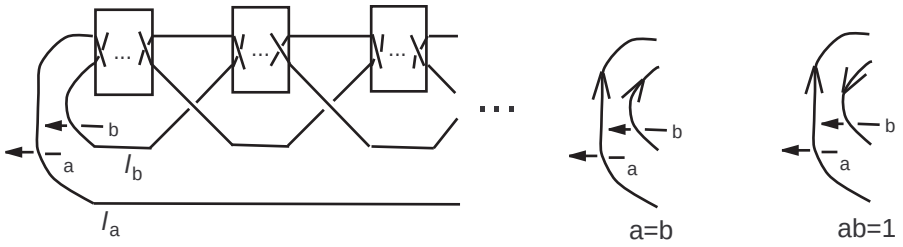


FIGURE 5. Overpasses ℓ_a, ℓ_b and the corresponding generators a and b of $\pi_1(S^3 \smallsetminus L_C)$. The case of co-directed and oppositely directed overpasses ℓ_a, ℓ_b, with the corresponding relations between a and b (after adding the commutativity relation $ab = ba$)

that $\pi_1(H \smallsetminus F_H) = \mathbb{Z}$. The common piece of the boundary of H and D^4 is a 3-ball, which intersects F along a pair of arcs, $\ell_a \cup \ell_b$. It is not difficult to see that in $\pi_1(H \smallsetminus F_H)$ we obtain the relation $a = b$ if n_0 is even, and $a = b^{-1}$ if odd. With this relation, the group G_L becomes abelian, and we obtain another relation (which depends on the orientation of ℓ_a, ℓ_b induced from L_C, as was explained). Under the second assumption of our Lemma, these two relations are different, that is both relations $a = b$ and $a = b^{-1}$ are satisfied in $\pi_1(Y \smallsetminus \widehat{F})$. □

4. Equivariant version of the Fintushel-Stern double node knot surgery

4.1. Equivariant knot surgery

The 4-dimensional knot surgery consists of removing from a 4-manifold X a trivialized tubular neighborhood $N(T) = T \times D^2$ of a torus $T \subset X$, and replacing it by $S^1 \times C(K)$, where $C(K)$ is a knot complement (see [FS1]). It is supposed that the gluing map $S^1 \times \partial C(K) \to T \times \partial D^2$ identifies a longitude $\mathrm{pt} \times \ell \subset S^1 \times C(K)$ with a meridian of T, $m_T = \mathrm{pt} \times \partial D^2 \subset \partial N(T)$, which yields a 4-manifold X_K, homologically equivalent to X.

In the equivariant version of this construction, we suppose that X is endowed with an orientation preserving involution, c, which keeps invariant the torus T as well as its neighborhood $N(T)$. We say that a trivialization $N(T) = T \times D^2$ of a tubular neighborhood, $N(T)$, of T is *equivariant* if the action of c on $N(T)$ can be presented as the direct product of $c|_T$ and the complex conjugation in $D^2 \subset \mathbb{C}$. Note that equivariant trivializability is equivalent to the existence of a projection $N(T) \to D^2$ which commutes with $c|_{N(T)}$ and the complex conjugation in D^2. In the case of our interest, T is a real non-singular fiber in a real elliptic fibration and thus admits such an equivariantly trivializable neighborhood.

Let us assume in addition that $c|_T$ reverses the orientation of T and has two-component fixed point set, $F \cap T$ (see Figure 2c)). In this case the quotient $\mathcal{A} = T/c$ is an annulus

and in the coordinates defined by some diffeomorphism $T = S^1 \times S^1$ the action of c looks like $(z_0, z_1) \mapsto (z_0, \bar{z}_1)$. Thus for an appropriate diffeomorphism $N(T) = S^1 \times S^1 \times D^2$, this action looks like $(z_0, z_1, z_2) \mapsto (z_0, \bar{z}_1, \bar{z}_2)$.

From a knot $K \subset S^3$ we require that it has an axis of symmetry, and intersects this axis at a pair of points. It will be convenient to choose the complex conjugation, conj, in $S^3 \subset \mathbb{C}^2$ as such a symmetry, so that the axis is $S^1_{\mathbb{R}} = S^3 \cap \mathbb{R}^2$. It is not difficult to see that such a knot K admits an equivariant tubular neighborhood, $N(K)$, in which c acts as $(z_1, z_2) \mapsto (\bar{z}_1, \bar{z}_2)$, with respect to a trivialization $S^1 \times D^2 = N(K)$. We should choose such a trivialization to be *null-framed*, which means that a longitude $\ell_K = S^1 \times \mathrm{pt}$ is null-homologous in the knot complement $C(K) = \mathrm{Cl}(S^3 \setminus N(K))$. Note that one can actually choose an equivariant trivialization with any framing: if we split $N(K) = S^1 \times D^2$ into a pair of cylinders $I \times D^2$ permuted by c and change a trivialization of one cylinder by n half-twists, then this trivialization can be c-symmetrically extended to another cylinder and will change the trivialization of $S^1 \times D^2$ by n full twists.

We can glue $S^1 \times C(K)$ to $X \setminus N(T)$ via an equivariant gluing map $g : S^1 \times \partial C(K) \to \partial N(T)$. Using the coordinates (z_0, z_1, z_2) in $\partial C(K) = \partial N(K)$ and $\partial N(T)$ from the above trivializations of $N(K)$ and $N(T)$, we define the map g as $(z_0, z_1, z_2) \mapsto (z_0, z_2, z_1)$. Such an equivariant knot surgery yields a 4-manifold X_K endowed with an involution, $c_K : X_K \to X_K$.

4.2. The tangle surgery in quotient-spaces

The quotient $N(K)/\mathrm{conj}$ is a 3-ball, which can be viewed as a regular neighborhood, $N(\mathfrak{s}_K)$, of the arc $\mathfrak{s}_K = K/\mathrm{conj}$ in $S^3 = S^3/\mathrm{conj}$. Thus, $B_K = C(K)/\mathrm{conj}$ is also a 3-ball, complementary to $N(\mathfrak{s}_K)$. The unknot $S^1_{\mathbb{R}} \subset S^3/\mathrm{conj}$ splits into a trivial tangle $\mathfrak{t} = S^1_{\mathbb{R}} \cap N(\mathfrak{s}_K)$ and a non-trivial one, $\mathfrak{t}_K = S^1_{\mathbb{R}} \cap B_K$ (see Figure 6d)).

Example 4.1. The twist-knot $K = K_n$, which will be used in our construction, admits a conj-invariant presentation, as it is sketched in Figure 6a). Figures 6d)–6e) present the corresponding tangle splitting $S^1_{\mathbb{R}} = \mathfrak{t} \cup \mathfrak{t}_K \subset S^3$.

The quotient-space X_K/c_K is obtained from X/c by removing a regular neighborhood, $N = N(\mathcal{A})$, of the annulus $\mathcal{A} = T/c$, which can be viewed as $N = S^1 \times N(\mathfrak{s}_K) = S^1 \times B^3$, and replacing it by $S^1 \times B_K = S^1 \times B^3$. Such a surgery does not change the differential-topological type of a 4-manifold, so we can identify both quotients, $Y = X/c = X_K/c_K$.

The branching locus F_K of the double covering $X_K \to Y$ is obtained from F after replacing $F_N = F \cap N = S^1 \times \mathfrak{t}$ by $S^1 \times \mathfrak{t}_K$ inside $N = S^1 \times B^3$. Note that the components of $\mathfrak{t}_K$ and of $\mathfrak{t}$ connect the same pairs of their common endpoints. We denote these four endpoints $p_1^{\pm}, p_2^{\pm}$, and assume that p_i^+ is connected with p_i^-. Moreover, both tangles must have *the same framing*. This means by definition that the kernel of the inclusion homomorphism $H_1(S^2 \setminus \partial\mathfrak{t}) \to H_1(D^3 \setminus \mathfrak{t})$ is the same as for $H_1(S^2 \setminus \partial\mathfrak{t}_K) \to H_1(D^3 \setminus \mathfrak{t}_K)$.

This kind of surgery will be called *tangle surgery of $F \subset Y$ along an annulus membrane $\mathcal{A}$*. It can be applied to any surface F in a 4-manifold Y under the assumption that the annulus membrane $\mathcal{A}$ with the boundary on this surface is *null-framed*. This means

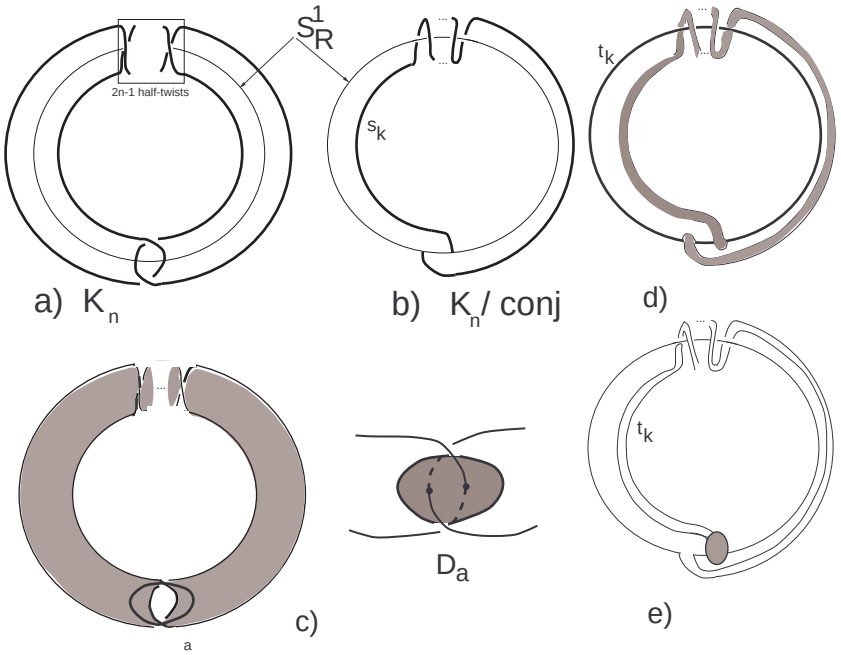


FIGURE 6. a) The twist-knot $K = K_n$ with the axis of symmetry $S^1_{\mathbb{R}}$. b) The arc $\mathfrak{s}_K = K/\mathrm{conj}$. c) Seifert surface S°_K of genus 1 bounded by K. It contains conj-invariant curve a, which bounds a conj-invariant disc D_a. d) The ball $N(\mathfrak{s}_K)$ and tangle $\mathfrak{t}_K$ in its complement. e) $\mathfrak{t}_K$ after an isotopy of $N(\mathfrak{s}_K)$ (the ball shaded on the Figure).

by definition that for some trivialization $N = S^1 \times D^3$ of its regular neighborhood, $N = N(\mathcal{A})$, the part of surface $F \cap N$ is identified with $S^1 \times \mathfrak{t}$, and $\mathcal{A}$ is identified with $S^1 \times \mathfrak{s}$, where $\mathfrak{s}$ is a line segment connecting the midpoints of the components of $\mathfrak{t}$ (see Figure 7a)). The following Lemma summarizes our observations.

Lemma 4.1. *An equivariant knot surgery on (X, c) along a c-invariant torus $T \subset X$ gives (X_K, c_K) with the same quotient-space $Y = X_K/c_K = X/c$. The fixed point set $F_K \subset Y$ of c_K is obtained from the fixed point set F by the tangle surgery along the annulus membrane $\mathcal{A} = T/c$.* □

4.3. Commutativity of π_1 throughout the knot surgery

Lemma 4.2. *Assume that $F \subset Y$ is a surface in a 4-manifold and $\mathcal{A}$ is a null-framed annulus membrane on F such that $F \smallsetminus \partial\mathcal{A}$ is connected. Assume that F_K is obtained from F by applying the tangle surgery along $\mathcal{A}$ with respect to $\mathfrak{t}_K$, where $K = K_n$ is the*

twist-knot from Example 4.1. Assume furthermore that the group $\pi_1(Y \smallsetminus (F \cup \mathcal{A}))$ *is abelian. Then* $\pi_1(Y \smallsetminus F_K)$ *is also abelian and isomorphic to* $\pi_1(Y \smallsetminus F)$.

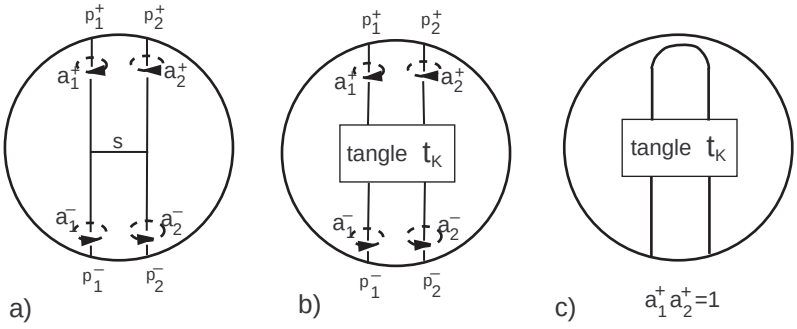


FIGURE 7. a) Trivial tangle $\mathfrak{t}$ with the connecting line segment $\mathfrak{s}$. The generators $a_i^{\pm}$ of $\pi_1(S^2 \smallsetminus \partial\mathfrak{t})$. b) The result of a tangle surgery. c) Adding the relation $a_1^+ a_2^+ = 1$ to the group of tangle $\mathfrak{t}_K$ effects like connecting together the points p_1^+ and p_2^+.

Proof. Let $Y' = \mathrm{Cl}(Y \smallsetminus N)$ and $F' = F \cap Y'$, $F_{N,K} = F_K \cap N$. By the Van Kampen theorem, $\pi_1(Y \smallsetminus F_K) = G' *_H G_N$, where $G' = \pi_1(Y' \smallsetminus F')$, $H = \pi_1(\partial N \smallsetminus \partial F_N)$, and $G_N = \pi_1(N \smallsetminus F_{N,K})$. Note that $N \smallsetminus (\mathcal{A} \cup F_N)$ can be deformation retracted to its boundary $\partial N \smallsetminus \partial F_N$, which implies that group G' is abelian, by the assumption on $\pi_1(Y \smallsetminus (F \cup \mathcal{A}))$. Note that $G' *_H G_N = G' *_H (G_N/K)$, where K is the image in G_N of the kernel of the homomorphism $H \to G'$. We will show that G_N/K is an abelian group and the product homomorphism $H \to G_N \to G_N/K$ is epimorphic. This implies that $G' *_H G_N$ is a quotient of G' and thus is also abelian.

Note that $H = \mathbb{Z} \times \pi_1(S^2 \smallsetminus \partial\mathfrak{t})$, where the second factor is a free group or rank 3. It is convenient to present this free group by 4 generators, $a_1^{\pm}$, $a_2^{\pm}$, satisfying the relation $a_1^+ a_1^- a_2^- a_2^+ = 1$. These generators correspond to the loops around the tangle endpoints, $p_1^{\pm}, p_2^{\pm} \in S^2$, in the positive direction, see Figure 7a).

Let us fix some element $a \in G'$ presented by a loop around $F \smallsetminus \partial\mathcal{A}$. Commutativity of G' and connectedness of $F \smallsetminus \partial\mathcal{A}$ imply that the inclusion homomorphism $H \to G'$ sends each of $a_i^{\pm}$ either to a, or to a^{-1} (depending on the topology of the boundary $\partial\mathcal{A}$ as an oriented curve in F). Such a relation, $a_i^{\pm} = a$ or $a_i^{\pm} = a^{-1}$, is inherited by the quotient-group G_N/K. To complete our proof of the Lemma it is enough to show that by adding this relation we transform group $\pi_1(D^3 \smallsetminus \mathfrak{t}_K)$ into a cyclic group with a generator a (since the factor $\mathbb{Z}$ in $G_N = \mathbb{Z} \times \pi_1(D^3 \smallsetminus \mathfrak{t}_K)$ lies in the center and comes from the corresponding factor in $H = \mathbb{Z} \times \pi_1(S^2 \smallsetminus \{p_1^+, p_1^-, p_2^+, p_2^-\})$).

We will present two arguments. The first one can be applied to any knot K admissible for an equivariant knot surgery, but it works only if we have a relation $a_1^+a_2^+ = 1$, or $a_1^-a_2^- = 1$. Note that if we connect the endpoints p_1^+ and p_2^+ as shown on Figure 7c), we modify the group $\pi_1(B^3 \smallsetminus \mathfrak{t}_K)$ by adding a relation $a_1^+a_2^+ = 1$. In the case of tangles $\mathfrak{t}_K$ constructed from conj-invariant knots K, this modification transforms the tangle into an unknotted arc in D^3. Thus, the group $\pi_1(D^3 \smallsetminus \mathfrak{t}_K)$ becomes cyclic and generated by any of the elements $a_i^{\pm}$. The case of adding relation $a_1^-a_2^- = 1$ is analogous.

Our second argument is specific for the twist-knot $K = K_n$, but can be applied in the case of relation $a_1^+ = a_2^+$ or $a_1^- = a_2^-$ (as well as in the case of relation $a_1^{\pm}a_2^{\pm} = 1$ considered before). First, we observe that the upper Wirtinger presentation gives 5 generators for $\pi_1(D^3 \smallsetminus \mathfrak{t}_K)$, namely $a_i^{\pm}$, $i = 1, 2$, and one more generator b shown on Figure 8.

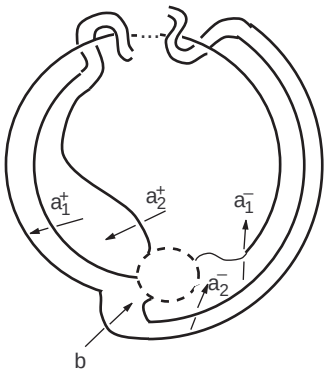


FIGURE 8. The tangle group becomes abelian after adding the relations $a_1^+ = a_2^+$, $a_1^- = (a_1^+)^{\pm 1}$, and $a_2^- = (a_1^+)^{\pm 1}$.

The two strands of the tangle $\mathfrak{t}_K$ with the origins at the points p_1^+ and p_2^+ pass together several times under $S^1_{\mathbb{R}}$. These underpasses separate the consecutive overpasses on the first strand which yield elements $a_1^+, b^{-1}a_1^+b, \ldots$, which are all conjugate to a_1^+. The similar overpasses on the second strand give elements $a_2^+, b^{-1}a_2^+b, \ldots$, which are conjugate to a_2^+ by the same sequence of elements. In the end of the sequence, we obtain elements $b^{-1} = x^{-1}a_1^+x$ and $(a_2^-)^{-1} = x^{-1}a_2^+x$, which are conjugate to a_1^+ and a_2^+ via the same element $x \in \pi_1(D^3 \smallsetminus \mathfrak{t}_K)$. So, a relation $a_1^+ = a_2^+$ which we add implies that $b = a_2^-$, whereas $a_1^+a_2^+ = 1$ implies $ba_2^- = 1$. In any case, generator b can be eliminated, and after adding two more relations to $\pi_1(D^3 \smallsetminus \mathfrak{t}_K)$, namely $a_1^- = a_1^+$ (or $a_1^- = (a_1^+)^{-1}$) and $a_2^- = a_1^+$ (or $a_2^- = (a_1^+)^{-1}$), we obtain a cyclic group.

Finally, we can observe that the surface F_K is homologically equivalent to F, and thus $H_1(Y \smallsetminus F_K) = H_1(Y \smallsetminus F)$. The group $\pi_1(Y \smallsetminus F)$ is abelian due to the assumption of the Lemma, because $Y \smallsetminus F$ is obtained from $Y \smallsetminus (F \cup \mathcal{A})$ after puncturing an annulus (which

may only add a relation to π_1) and then puncturing an arc (which adds a 3-cell and thus does not change π_1). Thus, we obtain an isomorphism $\pi_1(Y \setminus F_K) = \pi_1(Y \setminus F)$. □

Lemma 4.3. *Let (X, c) be the real elliptic surface constructed in Section 2, and T be the real fiber from Lemma 2.1(6). Then the membrane $\mathcal{A} = T/c$ satisfies the assumptions of Lemma 4.2, and thus $\pi_1(S^4 \setminus F_K) = \mathbb{Z}/2$ (here $S^4 = X/c$ by Lemma 2.1(6) and $F_K \cong F \cong \#10\mathbb{R}\mathrm{P}^2$).*

Proof. Connectedness of $F \setminus \mathcal{A}$ is observed in Lemma 2.1(7). The group $\pi_1(S^4 \setminus (F \cup \mathcal{A}))$ was shown to be cyclic in [FKV2], §4, under the assumption that $T \subset X$ is obtained from a real non-singular cubic in $\mathbb{C}\mathrm{P}^2$ by blowing up the base-points of a real pencil of cubics. This is so in our case, as follows from property (4) of Lemma 2.1. □

4.4. The equivariant double node surgery

To justify that a pseudo-section $S_K \subset X_K$ can be chosen c-invariant we recall first its construction in [FS2]. Consider a disc $\Delta_1 \subset \mathbb{C}\mathrm{P}^1$ which contains inside precisely two critical values $s_+, s_- \subset \Delta_1$ of an elliptic Lefschetz fibration $p : X \to \mathbb{C}\mathrm{P}^1$. Assume moreover that the corresponding two vanishing cycles in a non-singular fiber, T_s, $s \in \Delta_1$, are isotopic. Let $\Delta \subset \Delta_1$ denote a smaller disc not containing points $s_\pm$, and $U = p^{-1}(\Delta)$. Consider a section $S \subset X$ of p. Its restriction over Δ is the disc $\widetilde{\Delta} = S \cap U$. Using that the gluing map in the definition of the knot surgery which yields X_K may be adjusted by an isotopy, we can make the boundary $\partial\widetilde{\Delta}$ match with the boundary of a Seifert surface $S^\circ_K \subset \mathrm{pt} \times C(K) \subset X_K$ and obtain a closed surface $S^*_K = (S \setminus \widetilde{\Delta}) \cup S^\circ_K$ in X_K. If $K = K_n$ (the twist-knot on Figure 6a)), then S^*_K is a torus which has a certain disc membrane $D^*_a \subset X_K$ bounded by curve $a = \partial D^*_a = D^*_a \cap S^*_K$ and having self-intersection $(D^*_a)^2 = -1$ (relative to the boundary on the surface S^*_K). The torus S^*_K can be deformed and degenerated into a fishtail $S_K \subset X_K$, as we pinch curve $a \subset S^*_K$ along disc D^*_a. The local topology of S_K near its singular point is like near an algebraic double point, and the embedded surface $S_K \subset X_K$ is differential-topologically equivalent to a rational curve with a single node and self-intersection $(S_K)^2 = -1$.

To construct the disc D^*_a, we first take a disc $D_a \subset S^3$ bounded by a, so that D_a intersects K at a pair of points (see Figure 6c)). Disc D_a punctured at these points is embedded in $C(K) = \mathrm{pt} \times C(K) \subset S^1 \times C(K) \subset X_K$. The boundary of the punctures are two meridians, $m_\pm \subset \partial C(K)$, around K, which represent certain parallel curves in two different fibers of p after a knot surgery. We have to choose the knot surgery so that $m_\pm$ are the vanishing cycles corresponding to the singular values $s_\pm$ (this is possible because of our assumption that the vanishing curves corresponding to $s_\pm$ are isotopic). Then the two holes in the punctured disc D_a can be filled with the discs $D_{m_\pm}$ centered at the nodes of the singular fibers over $s_\pm$ and bounded by the vanishing cycles $m_\pm$. This gives D^*_a.

Lemma 4.4. *Consider the real elliptic surface (X, c) constructed in Section 2. Let $K = K_n$ be the twist-knot embedded* conj*-invariantly in S^3, as is shown on Figure 6a). Assume that (X_K, c_K) is obtained from (X, c) by an equivariant knot surgery along the*

non-singular fiber $T = T_s$ specified in Lemma 2.1(6). Then the pseudo-section S_K can be chosen c_K-invariant.

Proof. By Lemma 2.1(3), we can suppose that section S is c-invariant. We consider a disc $\Delta \ni s$ which is invariant under the complex conjugation in $\mathbb{C}\mathrm{P}^1$ (the base of the elliptic fibration), and denote by $r_\pm$ the endpoints of the interval $\Delta \cap \mathbb{R}\mathrm{P}^1$. In our example of real elliptic surface described in Section 2, we have a pair of real critical values, $s_\pm$, whose fibers $T_2 = p^{-1}(s_-)$, $T_3 = p^{-1}(s_+)$ can be used for a double node knot surgery, as it follows from Lemma 2.1(6). The curve $S \cap \partial U$ is a conj-invariant longitude of the knot K in the boundary of $C(K) \cong \mathrm{pt} \times C(K)$. This longitude spans a conj-invariant Seifert surface $S^\circ_K \subset C(K)$, as it is shown on Figure 6c). This implies that the torus S^*_K can be chosen c_K-invariant. Furthermore, we can choose the disc D^*_a to be c_K-invariant as follows. First, one can obviously choose a conj-invariant disc D_a (see Figure 6c)), so that the two intersection points $D_a \cap K$ are both real. This intersection can be made orthogonal by an equivariant isotopy of K near the intersection points, so that the meridians $m_\pm$ become also conj-invariant. As we perform an equivariant knot surgery, meridians $m_\pm$ become c-invariant vanishing cycles at the real fibers $p^{-1}(r_\pm) \subset \partial U$ (cycle $m_\pm$ corresponds to the critical value $s_\pm$). Note that the discs $D_{m_\pm}$ can be also chosen c-invariant. Namely, such a disc can be seen as the trace of the vanishing cycle $m_\pm$ as a non-singular fiber $p^{-1}(r_\pm)$ moves towards $p^{-1}(s_\pm)$ so that $m_\pm$ collapses. More precisely, the vanishing cycle is moved by the flow of a gradient-like vector field which covers the tangent vector field along a path in $\mathbb{C}\mathrm{P}^1$ connecting $r_\pm$ with $s_\pm$. If this path goes along $\mathbb{R}\mathrm{P}^1$ and the gradient-like vector field is c-invariant (for instance, a c-symmetrization of any given gradient-like vector field), then the disc $D_{m_\pm}$ becomes c-invariant, and hence, the disc D^*_a is also c-invariant.

Finally, note that there is a c_K-equivariant deformation of S^*_K, which contracts disc D^*_a and degenerates the torus S^*_K into S_K. Its construction goes like in the non-equivariant case: we deform S_K using a flow of a vector field tangent to D^*_a. Note that the deformation is equivariant if we choose a c-invariant vector field (it can be done like before, by symmetrization of any sample vector field whose flow contracts S^*_K to S_K). □

5. Proof of Theorem 1.2

Now we can combine together all the ingredients of the proof of Theorem 1.2 and see in details how it works for the configuration $C_{79,44}$ (the case of $C_{89,9}$ is analogous).

5.1. The construction of $C_{79,44}$

We recall that the chain $C_{79,44} = (-2,-5,-11,-2,-2,-2,-2,-2,-2,-3,-2,-2,-3)$ was constructed in [PSS] as follows. We start with an elliptic surface $\mathbb{C}\mathrm{P}^2\#9\overline{\mathbb{C}\mathrm{P}}^2$ which contains a fiber $T_0 = \mathbb{I}_8$ and four fishtails T_i, $i = 1,2,3,4$, two of which, T_2 and T_3, can be used for the double node knot surgery of [FS2] producing a pseudo-section S_K. Blowing up the node of T_1 we obtain a (-4)-sphere, and blowing up at the node of T_4 twice (the second time at the intersection of T_4 with the exceptional curve) we obtain a $(-2,-5)$-chain.

Blowing up T_0 4 times gives a chain-cycle $(-6,-2,-2,-2,-2,-2,-2,-3,-2,-2,-2,-1)$ from which we drop its (-1)-component to obtain a usual chain. The points to blowup are chosen so that S_K intersects this chain at a point of (-6)-sphere. We blowup also the node of S_K to obtain a (-5)-sphere and then smooth its intersection points with (-4) and (-6)-spheres produced by the fibers T_1 and T_0. This gives a (-11)-sphere, and together with $(-2,-5)$ and the remaining part of the long chain we obtain

$$(-2,-5,-11,-2,-2,-2,-2,-2,-2,-3,-2,-2,-2)$$

in $\mathbb{C}\mathrm{P}^2\#17\overline{\mathbb{C}\mathrm{P}}^2$. After blowing up the very last (-2)-component we obtain the configuration $C_{79,44}$ in $\mathbb{C}\mathrm{P}^2\#18\overline{\mathbb{C}\mathrm{P}}^2$. Its rational blowdown yields an exotic $\mathbb{C}\mathrm{P}^2\#5\overline{\mathbb{C}\mathrm{P}}^2$. As a knot K for the knot surgery one can use any one from a sequence of the twist knots K_i, which gives infinitely many exotic $\mathbb{C}\mathrm{P}^2\#5\overline{\mathbb{C}\mathrm{P}}^2$ (pairwise homeomorphic but non-diffeomorphic).

5.2. Construction of $F_i \subset S^4$ by equivariant surgery

In Section 2, Lemma 2.1, we provided a real elliptic surface suitable for the equivariant subsequent constructions, such that $\mathbb{C}\mathrm{P}^2\#9\overline{\mathbb{C}\mathrm{P}}^2/\mathrm{conj} = \mathrm{S}^4$. In Lemma 4.4 we show how to perform the double node knot surgery equivariantly, so that the pseudo-section is c-invariant. By Lemma 4.1 the quotient X/c is preserved. In what follows, all the blowups and smoothings are made at some real points and so are equivariant; real blowups do not change X/c as well. In Lemma 3.1 we described equivariant rational blowdowns and proved that they also do not change the quotient. So, in the quotient $S^4 = X/c$ of an exotic $\mathbb{C}\mathrm{P}^2\#5\overline{\mathbb{C}\mathrm{P}}^2$ we obtain a surface $F = \mathrm{Fix}(c)$. Pairwise non-diffeomorphism of an infinite family of such exotic 4-manifolds implies pairwise non-diffeomorphism of the corresponding embedded surfaces $F_i \subset S^4$.

5.3. The fundamental group $\pi_1(S^4 \smallsetminus F_i)$

In addition we justified that the fundamental group of the complement $S^4 \smallsetminus F = X/c \smallsetminus \mathrm{Fix}(c)$, where (X,c) is the corresponding 4-manifold X with an involution, is preserved abelian under all the equivariant surgery operations involved. At the first step, for $X = \mathbb{C}\mathrm{P}^2\#9\overline{\mathbb{C}\mathrm{P}}^2$, $F = \mathbb{R}\mathrm{P}^2\#9\overline{\mathbb{R}\mathrm{P}}^2$, and c being the complex conjugation, we need a bit stronger fact that the complement $S^4 \smallsetminus (F \cup \mathcal{A})$ has an abelian (and therefore cyclic) group π_1, where A is the annulus membrane on F represented by the quotient of a non-singular real elliptic fiber by the complex conjugation. Van Kampen's theorem easily implies that blowing up $\mathbb{C}\mathrm{P}^2$ at real points outside $\mathcal{A}$ preserves group $S^4 \smallsetminus (F \cup \mathcal{A})$ abelian (note that after blowing up at n points, F becomes $\mathbb{R}\mathrm{P}^2\#(\mathrm{n}+9)\overline{\mathbb{R}\mathrm{P}}^2$, while $S^4 = X/c$ and $\mathcal{A}$ remain the same). The details can be found in [FKV2], see the proof of Proposition 6 in Section 4.7. Note that these facts were also used in [F], and a brief proof was reproduced in the appendix there. In Lemmas 4.2–4.3 we prove that after an equivariant double node knot surgery applied to the real elliptic surface from Lemma 2.1, we obtain a 4-manifold with involution (X,c) such that $S^4 \smallsetminus F = X/c \smallsetminus \mathrm{Fix}(c)$ has an abelian π_1 (and thus, $\pi_1 = \mathbb{Z}/2$).

As we mentioned, the subsequent blowing up does not change π_1, and so to justify that $\pi_1(S^4 \smallsetminus F_i) = \mathbb{Z}/2$ in Theorem 1.2, it is left to prove that π_1 remains abelian after an equivariant rational blowdown of the configuration $C_{79,44}$ that we described in 5.1. Lemma 3.2 proves it under some assumptions on the chain C that we rationally blow down. Let us check that these assumptions are satisfied for $C_{79,44}$.

First, note that the last blowdown in the above construction of $C_{79,44}$ gives a (-1)-sphere which extends the chain $C = C_{79,44}$ to a longer chain $\widetilde{C}$ as required for Lemma 3.2. As we explained in Section 3.2, the condition (1) of Lemma 3.2 is equivalent to non-singularity modulo 2 of the intersection matrix $M_C = (C_i \circ C_j)$. Non-singularity can be justified for instance via modulo 2 blowdown of $C_{79,44}$ (or a reader can directly calculate the determinant). Namely, (mod 2) reduction of $C_{79,44}$ gives $(0,1,1,0,0,0,0,0,0,1,0,0,1)$, and after blowing down 9 times the first occurrence of "1" in the sequence, we obtain $(0,0,0,1)$, which can be then completely blown down. The characteristic sub-configuration of $C_{79,44}$ contains the components C_i with $i = 1,4,6,8,10,12$ (it can be either directly verified, or found using the algorithm explained in Remark 3.1). To verify the assumption (2) of Lemma 3.2 we use Remark 3.3 after this Lemma. The auxiliary (-1)-component extending our configuration C to $\widetilde{C}$ is adjacent to the last component C_{13}. Since (-1) is odd and C_{13} is not characteristic, this assumption is satisfied. □

References

[CH] A. Casson, J. Harrer, *Some homology lens spaces which bound rational homology balls*, Pacific J. Math. **96** (1981) 23–36.

[F] S. Finashin, *Knotting of Algebraic Curves in* $\mathbb{C}\mathrm{P}^2$, Topology **41** (2002) 47–55.

[FKV1] S. Finashin, M. Kreck, V. Viro, *Exotic knottings of surfaces in the 4-sphere*, Bull. AMS **17** (1987) 287–290.

[FKV2] S. Finashin, M. Kreck, V. Viro, *Non-diffeomorphic but homeomorphic knottings of surfaces in the 4-sphere*, Lecture Notes in Math., Springer, Berlin **1346** (1988) 157–198.

[FS1] R. Fintushel and R. Stern, *Rational blowdowns of smooth 4-manifolds*, J. Diff. Geom. **46** (1997) 181–235.

[FS2] R. Fintushel and R. Stern, *Double node neighborhoods and families of simply connected 4-manifolds with* $b^+ = 1$, J. Amer. Math. Soc. **19** (2006) 171–180.

[K] M. Kreck, *On the homeomorphism classification of smooth knotted surfaces in the 4-sphere*, Geometry of low-dimensional manifolds, 1 (Durham, 1989) London Math. Soc. Lecture Notes Ser. **150** (1990) 63-72.

[PSS] J. Park, A. Stipsicz, Z. Szabo, *Exotic smooth structures on* $\mathbb{C}\mathrm{P}^2\#5\overline{\mathbb{C}\mathrm{P}}^2$, Math. Res. Lett. **12** (2005) 701–712.

Middle East Technical University, Department of Mathematics
Ankara 06531 Turkey
E-mail address: serge@metu.edu.tr