ASPECTS OF MASS TRANSPORTATION
IN DISCRETE CONCENTRATION INEQUALITIES

A Thesis
Presented to
The Academic Faculty

by

Marcus D. Sammer

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

School of Mathematics
Georgia Institute of Technology
April 2005

ASPECTS OF MASS TRANSPORTATION
IN DISCRETE CONCENTRATION INEQUALITIES

Approved by:

Dr. Prasad Tetali, Advisor
School of Mathematics
Georgia Institute of Technology

Dr. Wilfrid Gangbo
School of Mathematics
Georgia Institute of Technology

Dr. Ellis Johnson
School of Industrial and Systems
Engineering
Georgia Institute of Technology

Dr. Thomas Morley
School of Mathematics
Georgia Institute of Technology

Dr. Michael Loss
School of Mathematics
Georgia Institute of Technology

Date Approved: 11 April, 2005

To Marla,

who never let me give up.

ACKNOWLEDGEMENTS

I first want to thank my advisor, Prasad Tetali, who worked harder for me than I deserved. He inspired me with his excitement about our work, and he saw me through to the end.

I would like to thank my other thesis defense committee members: Wilfrid Gangbo, for acting as my co-advisor while Prasad was out of town; Ellis Johnson, for his input and serving on my oral exam committee; Tom Morley, whose ideas are always interesting; and Michael Loss, for filling in when I needed someone. And I would like to thank Eric Carlen for serving on my oral exam committee and for the helpful discussions throughout my research.

I am indebted to many other wonderful professors at Georgia Tech, but I'm especially grateful for Alfred Andrew and Robert Ghrist who spent hours of their own time preparing me for the comprehensive exam.

And I would like to thank my undergraduate professors: Robert Moore, Arthur Richert, and Larry Hanson, who inspired me to study math in the first place.

I will always appreciate the friendship of my office mate for three years, Sean Curran, without whom school would have been much less fun. Also, I thank Max von Renesse for his friendship and positive outlook on my work.

I would like to thank my parents, Christine Sammer and Edwin Sammer, for their love and support throughout my education. Doug Sammer and Julie Sammer deserve thanks for being my favorite brother and sister. I will always remember my grandfather, Cecil Davis, who taught me about math at a young age. In addition, Marla's family has been great sources of support (and entertainment).

More than anything I want to thank my wife, Marla Sammer, who has encouraged me, pushed me, and put up with me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

During the last half century there has been a resurgence of interest in Monge's 18th century mass transportation problem, with most of the activity limited to continuous spaces. This thesis, consequently, looks at the role of mass transportation in the context of the measure concentration phenomenon in a discrete setting. Inequalities capturing such concentration on n-fold products of graphs, equipped with product measures, have been well investigated using combinatorial and probabilistic techniques, the most notable being martingale techniques. The emphasis here, is instead on the analytic viewpoint. Of particular relevance and focus is the so-called subgaussian constant, which is an optimal constant in a transportation inequality on a graph, equipped with a probability measure on the vertices of the graph. The relationship between the transportation inequality and the Poincaré and modified log-Sobolev inequalities is also examined. In such comparisons, different versions of the transportation inequality are considered, and the role of the particular distance function employed on the underlying graph is studied. The duality shown by Bobkov and Götze of the transportation inequality and a generating function inequality is utilized in finding the asymptotically correct value of the subgaussian constant of a cycle. This result tensorizes to give a concentration inequality on the discrete torus. Finally, a candidate notion of a *discrete* Ricci curvature for finite Markov chains based on coupling of Markov chains and given in terms of mass transportation is considered. This analog of curvature is then compared to another put forward by Schmuckenschläger, with the conclusion that this notion merits further investigation and development. Overall, the thesis demonstrates the utility of using the mass transportation problem in the study of discrete concentration inequalities.

CHAPTER I

INTRODUCTION AND BACKGROUND

The mass transportation problem was developed by Gaspard Monge [27] in 1781, but beginning in the mid 20th century, interest in this concept has surged. As Vershik [37] describes, this era was heralded by the 1942 publication of L.V. Kantorovich's note [24] in which he details the problem, its dual, and the optimality condition. The problem slowly gained recognition through the second half of the century, but only in 1987, after the publication of a note by Yann Brenier [13] (the event Villani [38] cites as the beginning of the subject's "extreme popularity"), did mathematicians from many disparate fields realize their connection with the problem.

Throughout this thesis we look at one such connection—the role of the mass transportation problem in the study of discrete concentration inequalities. In continuous settings, like on a Riemannian manifold, the transportation problem has had notable success as a tool for proving concentration results. Here, we place particular emphasis on carrying these techniques over to the discrete setting when possible.

To set the stage for exploring the mass transportation problem in the discrete concentration of measure phenomenon, the remainder of this chapter explains the definitions, notation, and background material used throughout the thesis. The following chapter consolidates many of the technical lemmas we use throughout the thesis—lemmas that are general in nature to the mass transportation problem or the transportation inequalities. The next three chapters are largely independent, and each attempts to capture a useful aspect of the mass transportation problem in the study of discrete concentration inequalities.

Chapter 3 looks at the relationship between four different concentration inequalities, two of which are written in terms of the mass transportation problem. A new result proven in this chapter is that the modified log-Sobolev inequality implies the transportation inequality. We discuss how these relationships can be exploited to find bounds on the concentration

constants defined by the inequalities. We also see how the underlying distance function can be modified to tighten some of these bounds. Chapter 4 focuses on the subgaussian constant. As shown by Bobkov and Götze [7], this constant is equivalently defined by both a generating function inequality and a transportation inequality. We use this dual formulation to find the asymptotically correct value of the subgaussian constant of a cycle. The physical intuition provided by the mass transportation problem is the key to finding the upper bound on the subgaussian constant when the cycle contains an odd number of vertices. Chapter 5 begins the development of a notion of discrete Ricci curvature for finite Markov chains. Bounds on Ricci curvature play a key role in many concentration results in the Riemannian setting, and the lack of a good analog in discrete spaces is a serious roadblock for realizing many continuous space techniques in the discrete setting. This work is still in a developmental stage, but has several promising aspects. First, the definition we put forward is closely related to the idea of path coupling in Markov chains, used for years to prove fast convergence to stationarity. Next, the discrete Ricci curvature is relatively easy to compute on example graphs, aiding both the intuitive process of further developing the theory, and the eventual practicality of the concept. Finally, simply the fact that the notion is defined in terms of the mass transportation problem, which has shown itself to be useful in so many other areas, makes the definition appealing. Concluding thoughts and several intriguing questions generated by this work are summarized in Chapter 6.

1.1 The Setting

Here we formalize what we mean by a discrete setting. Our work is done on finite graphs and Markov chains on the graphs.

1.1.1 Graphs and Product Graphs

By a graph $G = (V, E)$, we mean a finite undirected graph without self-loops or multiple edges with vertex set V and edge set E. Unless otherwise specified, we also assume that the graph G is connected. Edges in E are written as $\{x, y\}$ where $x, y \in V$ are adjacent vertices in G. At times we denote the fact that x and y are adjacent by $x \sim y$. With each graph G we associate a probability measure π on V and a finite distance function d between the

vertices of G. We commonly choose π to be the uniform probability measure and d to be the graph distance. Recall here that the graph distance between vertices x and y is defined as the length of a shortest path between x and y. We also denote the set of probability measures on V by $P(G)$.

If $\{G_i = (V_i, E_i)\}_{i=1}^n$ is a family of graphs with associated measures π_i and associated distances d_i, then we may define the (Cartesian) product graph $G = \prod_{i=1}^n G_i = (V, E)$ as follows. The vertex set $V = \prod_{i=1}^n V_i$. We can write $x \in V$ in component form as $x = (x_i, \ldots, x_n)$ where $x_i \in V_i$ for each i. If $x, y \in V$, then $\{x, y\} \in E$ if and only if for some j, $\{x_j, y_j\} \in E_j$ and $x_i = y_i$ for all $i \neq j$. We write G^n for $\prod_{i=1}^n G$. The measure π we associate with G is the product measure defined by:

$$\pi(x_1, \ldots, x_n) = \prod_{i=1}^n \pi_i(x_i).$$

The distance d we associate with G is the l_1 distance defined by:

$$d(x, y) = \sum_{i=1}^n d_i(x_i, y_i).$$

And we note that the graph product is associative:

$$(G_1 \square G_2) \square G_3 = G_1 \square (G_2 \square G_3),$$

which is to be contrasted with the product Markov chains described in the next section.

1.1.2 Continuous Time Markov Chains and Product Chains

In the following, we use some definitions from [21], [28], and [10].

We begin with a graph $G = (V, E)$ with associated measure π and associated distance d as described in the previous section. We define a continuous time Markov chain on G that respects the graph structure and the associated measure of G. We often refer to this as simply a Markov chain on G. For the Markov chain to respect the graph structure and the associated measure we require two things. First, if $\{x, y\} \notin E$, then the Markov chain transition rates between x and y must be zero. Next, the stationary distribution of the chain must be the measure associated with the graph. However, the relationship between the distance d and both the graph structure and the Markov chain on the graph is

3

flexible and problem dependent. In the literature, the term "Markov process" is sometimes used to distinguish a continuous time Markov chain from a discrete time Markov chain. In this thesis, instead, we assume that a Markov chain is run in continuous time unless we specifically specify that we are considering a discrete time Markov chain.

Here we describe the generator L of a Markov chain on G that respects the graph structure of G and the associated measure π. For $x, y \in V$ with $x \neq y$, $L(x, y)$ is the transition rate from x to y. Because the Markov chain respects the graph structure, we have $L(x, y) = L(y, x) = 0$ if $\{x, y\} \notin E$. We define $L(x, x) = -\sum_{\substack{y \in V \\ y \neq x}} L(x, y)$ so that $\sum_{y \in V} L(x, y) = 0$ for each $x \in V$. Because the Markov chain respects the associated measure π, π must be the unique stationary measure of the chain, which in particular implies that $\pi L = 0$. The continuous time Markov chain generated by L has the transition semigroup $\{P_t = e^{tL} : t \geq 0\}$.

Throughout this thesis, we work exclusively with reversible Markov chains. Reversibility of a chain with generator L is equivalent to the requirement that $L(x, y)\pi(x) = L(y, x)\pi(y)$ for each $x, y \in V$ (also known as the detailed balanced condition).

As a normalization so that comparisons may be made between constants, we will sometimes require that

$$\sum_{x \in V} \sum_{\substack{y \in V \\ y \neq x}} L(x, y)\pi(x) \leq 1. \tag{1}$$

The potential need for this normalization factor is explained in Section 1.4.2, and it is used in Sections 3.2 and 3.3.

At times we will be interested in discrete time chains on G. Suppose P is the transition probability matrix of a reversible discrete time chain that respects the graph structure of G and the associated probability measure π. As in the continuous time setting, by this we mean that if $x \neq y$ we have $P(x, y) = P(y, x) = 0$ for $\{x, y\} \notin E$, and that π is the unique stationary distribution of the chain. To avoid unnecessary technicalities we also assume that the chain is aperiodic. Recall that the conditions of irreducibility and aperiodicity are sometime collectively referred to using the common term *ergodicity*.

Let $L = P - I$. Then the continuous time chain generated by L also respects the graph

structure and the associated measure, and is called the "continuization" of the discrete time chain (see [1] for standard terminology and facts).

Next assume we start with a continuous time chain on G whose generator L satisfies:

$$\sum_{\substack{y \in V \\ y \neq x}} L(x,y) \leq 1 \tag{2}$$

for each $x \in V$. Define a probability transition matrix P by:

$$P(x,y) = \begin{cases} L(x,y), & x \neq y \\ 1 + L(x,x), & x = y \end{cases}$$

Then the discrete time chain with probability transition matrix P also respects the graph structure and associated measure of G. Furthermore the continuization of this discrete time chain is the original continuous time chain.

Now consider a family of graphs $\{G_i = (V_i, E_i)\}_{i=1}^n$ with associated measures π_i. For each i let L_i be the generator of a continuous time Markov chain on G_i. Let $G = (V, E)$ be the product graph $\prod_i G_i$ with associated measure $\pi = \prod_{i=1}^n \pi_i$ as defined in Section 1.1.1. Then we define the generator L of a product Markov chain on G by:

$$L((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \frac{1}{n} \sum_{i=1}^n L_i(x_i, y_i) \prod_{\substack{j=1 \\ j \neq i}}^n \delta^{x_j}(y_j),$$

where $\delta^{x_j}(y_j)$ is one if $x_j = y_j$ and zero if $x_j \neq y_j$. The verification that the Markov chain generated by L respects the graph structure of G and the measure π associated with it is straight forward, and we omit it.

Suppose again that we have a Markov chain with generator L on the graph $G = (V, E)$ with associated probability measure π. If the chain starts with distribution ν, then the distribution at time $t \geq 0$ is νP_t. Let $\nu_t = \nu P_t$, and note that $\nu = \nu_0$. If f_t is the density of ν_t with respect to the stationary distribution π, then we also have $f_t = P_t f_0$. For each $x \in V$ we can calculate the derivative of $\nu_t(x)$ and $f_t(x)$ with respect to t by:

$$\frac{d}{dt}\nu_t(x) = \nu_t L(x) \quad \text{and} \quad \frac{d}{dt}f_t(x) = Lf_t(x).$$

5

We will be interested in the derivative of functions of ν_t. Some of these derivatives will be defined in Section 1.4.2, while the derivative of the Wasserstein distance between two Markov chains will be described in Section 2.2.2.

1.2 Concentration of Measure

Suppose $G = (V, E)$ is a graph with associated probability measure π and distance function d. For $A \subset V$, we define the h enlargement of A by:

$$A_h = \{x \in V : d(x, A) < h\}.$$

Then the concentration function α of the graph G with measure π and distance d is defined by:

$$\alpha(h) = \max\{\pi((A_h)^c) : \pi(A) \geq 1/2\},$$

where B^c is the complement of B for any $B \subset V$. We will use subscripts on α when necessary to clarify which graph, measure, or distance function is being considered. The measure on the graph is said to have exponential concentration if there exist constants k and K for which

$$\alpha(h) \leq K e^{-kh},$$

while it is said to have normal concentration if there exist constants $\hat{k}$ and $\hat{K}$ for which

$$\alpha(h) \leq \hat{K} e^{-\hat{k}h^2}$$

for each $h > 0$. Under this definition, every measure on every graph has normal concentration, since we are considering finite graphs with finite distance functions. But we are more interested in concentration in product graphs, which motivates us to look at sequences of graphs $\langle G_n \rangle_{i=n}^{\infty}$. As introduced by Gromov and Milman [22], such a sequence is said to be a *normal Lévy family* if there exist constants $\tilde{k}$ and $\tilde{K}$ for which

$$\alpha_{G_n, \pi_n}(h) \leq \tilde{K} e^{-\tilde{k}nh^2}$$

so that graph G_n is normally concentrated with constant $\hat{k} = n\tilde{k}$. Since we are interested in product graphs, we will consider sequences in which $G_n = \prod_{i=1}^{n} H_i$ where $\langle H_i \rangle_{i=1}^{\infty}$ is some

other family of graphs. The simplest case of this form is when $G_n = H^n$ for a given graph H.

1.3   The Mass Transportation Problem

Let $G = (V, E)$ be a graph with associated measure π and distance function d. The mass transportation problem requires two probability measures ν_1 and ν_2 on V. In particular applications these measures may be related to the measure π associated with the graph, but in general there is no connection. The problem does, however, depend very specifically on the distance function d.

The problem consists of finding an optimal way of reconfiguring a mass distributed on the vertices of G according to the measure ν_1 into a mass distributed according to the measure ν_2. We consider measures μ on $V \times V$ which specify how the mass is to be transferred. For each pair of distinct vertices $x, y \in V$, $\mu(x, y)$ gives the amount of mass that moves from vertex x to vertex y, and $\mu(x, x)$ indicates the amount of mass that remains fixed at x. Since $\sum_{x \in V} \mu(v, x)$ must be the amount of mass that starts at vertex v and $\sum_{x \in V} \mu(x, v)$ is the amount of mass that ends up at vertex v, we get that μ has first and second marginals ν_1 and ν_2 respectively. Next we say that the cost of moving mass from vertex x to vertex y is proportional to $d(x, y)$, where d is the distance function associated with the graph. The total cost of reconfiguring the mass is then $\sum_{x,y} d(x, y)\mu(x, y)$. The problem of minimizing this cost becomes the following linear program:

$$
\begin{aligned}
\text{minimize:} \quad & M(\mu) = \sum_{x,y \in V} d(x, y)\mu(x, y) \\
\text{subject to:} \quad & \sum_{y \in V} \mu(x, y) = \nu_1(x) && \text{for each } x \in V \\
& \sum_{x \in V} \mu(x, y) = \nu_2(y) && \text{for each } y \in V \\
& 0 \leq \mu(x, y) && \text{for each } x, y \in V.
\end{aligned}
\tag{3}
$$

In this problem $\mu(x, y)$ is a variable for each $x, y \in V$, while d, ν_1, and ν_2 are fixed data. The mass transportation problem was originally formulated by Monge [27] in 1781, so any feasible μ that minimizes $M(\mu)$ will be referred to as a solution to Monge's problem with respect to ν_1 and ν_2.

The linear programming dual to this problem is:

$$\text{maximize:} \quad K(h,g) = \sum_{x \in V} h(x)\nu_1(x) + \sum_{x \in V} g(x)\nu_2(x)$$

$$\text{subject to:} \quad h(x) + g(y) \leq d(x,y) \qquad \text{for every } x,y \in V. \tag{4}$$

Here the variables are $h(x)$ and $g(x)$ for each $x \in V$, and the given data are d, ν_1, and ν_2. In 1942, Kantorovich [24] formulated the dual to Monge's problem in a general setting, where it is not simply a linear programming problem. So any feasible functions g and h which maximize $K(g,h)$ are called a solution to Kantorovich's problem with respect to ν_1 and ν_2.

Except in Section 1.4.3, we will assume that the distance function d is an actual metric on V satisfying the usual metric properties:

1. $d(x,y) = 0$ for $x,y \in V$ if and only if $x = y$.

2. $d(x,y) = d(y,x)$ for each $x,y \in V$.

3. $d(x,y) + d(y,z) \geq d(x,z)$ for each $x,y,z \in V$.

Under these assumptions we may always find solutions to Kantorovich's problem in which $h(x) = -g(x)$ for each $x \in V$. So we may simplify Kantorovich's problem to:

$$\text{maximize:} \quad K(g) = \sum_{x \in V} g(x)(\nu_1(x) - \nu_2(x))$$

$$\text{subject to:} \quad |g(x) - g(y)| \leq d(x,y) \qquad \text{for every } x,y \in V. \tag{5}$$

In this case, then a single feasible function f that minimizes $K(f)$ is called a solution to Kantorovich's problem with respect to ν_1 and ν_2. For completeness we give a proof of the equivalence of the two forms of Kantorovich's problem when d is a metric in Proposition 2.1.1.

While the joint optimal value of the problems of Monge and Kantorovich has several different names (including KROV distance), in the following it is referred to as the Wasserstein distance between ν_1 and ν_2, and is denoted by $W(\nu_1, \nu_2)$. We will use a subscript on W if necessary to clarify which distance function is being used on the graph.

1.4   The Inequalities

We begin with a graph $G = (V, E)$ with associated measure π and distance function d, and a reversible continuous time Markov chain with generator L that respects the graph structure and the associated measure of G. We have already defined the Wasserstein distance between two measures, but we need several more quantities before we can describe the inequalities. Expectations for any functions on V are taken with respect to the associated measure π unless indicated otherwise by a subscript. Since we are in a discrete setting, for a probability measure ν to be absolutely continuous with respect to π, we mean that $\nu(x) = 0$ whenever $\pi(x) = 0$. And for $x \in V$ with $\nu(x) = \pi(x) = 0$, by convention we define $\frac{\nu(x)}{\pi(x)} = 1$. When ν is absolutely continuous with respect to π, we sometimes denote the density of ν with respect to π as $\frac{d\nu}{d\pi}$.

First, the expected value of a function f on V is denoted by $\mathrm{E}(f)$, $\mathrm{E}[f]$, or $\mathrm{E}f$ and defined by:

$$\mathrm{E}(f) = \sum_{x \in V} f(x)\pi(x).$$

Next, the variance of f is denoted $\mathrm{Var}(f)$ or $\mathrm{Var}[f]$ and is defined by:

$$\mathrm{Var}(f) = \mathrm{E}(f^2) - (\mathrm{E}f)^2.$$

The entropy of f is denoted by $\mathrm{Ent}(f)$ or $\mathrm{Ent}[f]$ and is defined by:

$$\mathrm{Ent}(f) = \mathrm{E}(f \log f) - \mathrm{E}(f) \log(\mathrm{E}f).$$

For a measure ν absolutely continuous with respect to π, the relative entropy of ν with respect to π is denoted by $D(\nu||\pi)$ and is defined by:

$$D(\nu||\pi) = \sum_{x \in V} \nu(x) \log\left(\frac{\nu(x)}{\pi(x)}\right) = \mathrm{E}_\nu\left[\frac{d\nu}{d\pi}\right].$$

We note that for a probability measure ν absolutely continuous with respect to π, if f is the density of ν with respect to π, then $\mathrm{Ent}(f) = D(\nu||\pi)$. Also, by convention we define $0 \log(0) = 0$, so that $D(\nu||\pi)$ is a continuous function of ν on $P(G)$.

For two functions f and g on V, the Dirichlet form of f and g is denoted by $\mathcal{E}(f, g)$ and is defined by:

$$\mathcal{E}(f, g) = -\mathrm{E}[fLg].$$

9

Because the Markov chain generated by L is reversible, we also get:

$$\mathcal{E}(f, g) = \sum_{x,y \in V} (f(y) - f(x))(g(y) - g(x)) L(x, y) \pi(x).$$

Subscripts will be used on $\mathcal{E}$ when necessary to clarify which Markov generator is being considered. For example, and since we will need this equality later, we note that for a positive constant c we have:

$$\mathcal{E}_{cL}(f, g) = \mathrm{E}[f(cL)g] = c\mathrm{E}[fLg] = c\mathcal{E}_L(f, g).$$

And now we can describe the inequalities.

1.4.1 Transportation and Variance Transportation Inequalities

In this section we at last see the convergence of the mass transportation problem and the concentration of measure phenomenon embodied in these two transportation inequalities.

A graph $G = (V, E)$ with associated measure π and distance function d satisfies the transportation inequality with constant σ^2 if:

$$W^2(\nu, \pi) \leq 2\sigma^2 D(\nu || \pi) \tag{6}$$

for each probability measure ν that is absolutely continuous with respect to π. The smallest constant for which this inequality holds for each ν is known as the *subgaussian constant* $\sigma^2(G)$. The variance transportation inequality is satisfied with constant c^2 if:

$$W^2(\nu, \pi) \leq c^2 \operatorname{Var}\left(\frac{d\nu}{d\pi}\right) \tag{7}$$

for all ν absolutely continuous to π. The smallest constant c^2 for which the inequality holds for all ν is known as the *spread constant* $c^2(G)$. Subscripts may be used on $\sigma^2(G)$ or $c^2(G)$ to clarify which probability measure and distance function is being associated with the graph.

Both of these inequalities have what may loosely be called dual formulations. As shown By Bobkov and Götze [7], the subgaussian constant can be equivalently defined as the smallest constant σ^2 for which

$$\mathrm{E}\left[e^{t(f - \mathrm{E}[f])}\right] \leq e^{\sigma^2 t^2 / 2}$$

10

for every Lipschitz function f and real number t (see Proposition 2.3.2). Throughout this thesis, a function f on V is said to be Lipschitz if $|f(x) - f(y)| \leq d(x, y)$ for each $x, y \in V$ (i.e. we mean Lipschitz with constant one). The dual formulation of the spread constant (which was actually its original definition) is given by Alon, Boppana, and Spencer [3] as:

$$c^2(G) = \max_{Lipschitz f} \mathrm{Var}(f).$$

We prove the equivalence of these definitions of the spread constant in Proposition 2.3.1.

Both the transportation and variance transportation inequalities give upper bounds on the concentration function α defined in Section 1.2. Using the transportation definition of the subgaussian constant, the following bound can be derived (see [26] for example):

$$\alpha(h) \leq e^{-\frac{h^2}{8\sigma^2(G)}} \tag{8}$$

for $h \geq 2\sqrt{2\sigma^2(G)\log 2}$. While using the generating function definition, Bobkov, Houdré, and Tetali [8] derive the bound:

$$\alpha(h) \leq e^{-\frac{(h-\sigma)^2}{2\sigma^2(G)}} \tag{9}$$

for $h \geq \sqrt{\sigma^2(G)}$. As $\sqrt{\sigma^2(G)}$ can be on the order of the diameter of G, these bounds are of limited interest on their own. But implicit in [3] is the fact that

$$\sigma^2\left(\prod_{i=1}^n G_i\right) \leq \sum_{i=1}^n \sigma^2(G_i)$$

and in particular $\sigma^2(G^n) = n\sigma^2(G)$. So using the transportation definition, for example, we get:

$$\alpha_{G^n}(h) \leq e^{-\frac{h^2}{8n\sigma^2(G)}}$$

for $h \geq 2\sqrt{2n\sigma^2(G)\log 2}$. Note that the diameter of $\prod_{i=1}^n G_i$ is at least n, when using the graph distance, and this bound becomes useful when $h \gg \sqrt{n}$. In fact, if d is the distance on G and we use the normalized distance $\hat{d}(x, y) = \frac{1}{n}\sum_{i=1}^n d(x_i, y_i)$ on G^n, then $\langle G^n\rangle_{n=1}^\infty$ becomes a normal Lévy family since:

$$\alpha_{G^n, \hat{d}}(h) \leq e^{-\frac{nh^2}{8\sigma^2(G)}}.$$

The spread constant was introduced in [3] specifically because it gives the asymptotically correct value of the concentration function:

$$\alpha_{G^n}(h) = e^{-\frac{h^2}{2c^2(G)n}(1+o(1))}$$

for $\sqrt{n} \ll h \ll n$. From (9) and the tensoring property, the subgaussian constant gives us:

$$\alpha_{G^n}(h) \le e^{-\frac{h^2}{2\sigma^2(G)n}(1+o(1))}$$

for all $h \gg \sqrt{n}$. This shows that $\sigma^2(G) \ge c^2(G)$, which we prove directly in Proposition 3.1.1.

At this point we will mention one upper bound on the subgaussian constant (and hence on the spread constant). In [3], it is shown that

$$\mathrm{E}\left[e^{t(f-\mathrm{E}f)}\right] \le e^{t^2 D^2/8}$$

where D is the diameter of the graph. This gives the general upper bound $\sigma^2(G) \le \frac{D^2}{4}$.

1.4.2 Poincaré and Modified Log-Sobolev Inequalities

In Chapter 3 we compare the transportation inequalities of the previous section with the Poincaré and modified log-Sobolev inequalities introduced here. The Poincaré and modified log-Sobolev inequalities are of interest because of their connection with the concentration of measure phenomenon on graphs, and their role in bounding mixing times of Markov chains on graphs. While the relevance of the Poincaré inequality to the mixing time is classical, the connection between the modified log-Sobolev inequality and mixing (in the relative entropy sense) of finite Markov chains is more recent (see e.g., [10]).

We start with the Poincaré inequality. The graph G with associated measure π satisfies the Poincaré inequality with constant λ_1 if

$$\lambda_1 \mathrm{Var}(f) \le \mathcal{E}(f, f) \tag{10}$$

for all functions f on V. The largest constant for which this inequality holds for all f is known as the *spectral gap* $\lambda_1(L)$ of the Markov chain. It is also the smallest positive eigenvalue of $-L$.

The graph G with associated measure π satisfies the modified log-Sobolev inequality with constant ρ_0 if

$$\rho_0 \operatorname{Ent}(f) \leq \frac{1}{2}\mathcal{E}(f, \log f) \tag{11}$$

for all functions f on V. The largest constant for which this inequality holds for all f is known as the *modified log-Sobolev constant* $\rho_0(L)$ of the Markov chain. The inequality and the constant have also been introduced under the names entropy inequality and *entropy constant* in [20]. Although we will not be concerned with it, for comparison we also mention the (usual) log-Sobolev inequality, which is satisfied if

$$\rho \operatorname{Ent}(f^2) \leq 2\mathcal{E}(f, f) \tag{12}$$

for all functions f on V. Here the largest constant for which the inequality holds is the *log-Sobolev constant* $\rho(L)$. In a continuous setting where the chain rule for differentiation holds, the log-Sobolev and modified log-Sobolev inequalities are equivalent, while in the discrete setting they may be quite different.

We note that the left hand side of the Poincaré and the modified log-Sobolev inequalities does not depend on the Markov generator, while the Dirichlet form on the right hand side of both inequalities does. Using the fact observed earlier that $\mathcal{E}_{cL}(f, g) = c\mathcal{E}_L(f, g)$ for positive constants c, we get that $\rho_0(cL) = c\rho_0(L)$ and $\lambda_1(cL) = c\lambda_1(L)$. This is why a normalization factor such as (1) is needed when comparing these quantities to constants that do not depend specifically on a Markov generator on the graph.

Let ν_t be a Markov chain on G with generator L, and let f_t be the density of ν_t with respect to π. Some motivation for the definitions of the Poincaré and the modified log-Sobolev inequalities comes from the derivatives of $\operatorname{Var}(f_t)$ and $D(\nu_t\|\pi)$. As shown in [10], for example, we have:

$$\frac{d}{dt}\operatorname{Var}(f_t) = \frac{d}{dt}\sum_{x\in V} f_t(x)^2\pi(x) - 1 = \sum_{x\in V} 2f_t(x)Lf_t(x)\pi(x) = -2\mathcal{E}(f_t, f_t)$$

and

$$\frac{d}{dt}D(\nu_t||\pi) = \frac{d}{dt}\operatorname{Ent}[f_t]$$
$$= \frac{d}{dt}\sum_{x \in V} f_t(x)\log(f_t(x))\pi(x)$$
$$= \sum_{x \in V}(Lf_t(x) + Lf_t(x)\log f_t(x))\pi(x)$$
$$= \sum_{x \in V}\log(f_t(x))Lf_t(x)\pi(x)$$
$$= -\mathcal{E}(f_t, \log f_t).$$

These derivatives together with the Poincaré and the modified log-Sobolev inequalities give us:

$$\operatorname{Var}(f_t) \leq \operatorname{Var}(f_0)e^{-2\lambda_1 t} \quad \text{and} \quad D(\nu_t||\pi) \leq D(\nu_0||\pi)e^{-2\rho_0 t}$$

which provide bounds on the convergence of the Markov chain to stationarity.

1.4.3 Quadratic Transportation Inequality

In the continuous setting, the quadratic cost transportation inequality plays a greater role than the transportation inequality we are studying. Here we examine a couple of the reasons for using a quadratic cost transportation inequality and some of the hurdles faced when trying to use such an inequality in a discrete setting. The quadratic cost transportation inequality is not considered anywhere else in this thesis (with the exception of Proposition 2.3.7).

First we describe the quadratic cost transportation inequality. Let (X, d) be a metric space, with an associated probability measure π on X. Let $c : X \times X \to \mathbb{R}$ be the cost function $c(x, y) = d^2(x, y)$. Then (X, d) and π satisfy a quadratic cost transportation inequality with constant t if:

$$W_c(\nu, \pi) \leq tD(\nu||\pi) \tag{13}$$

for each ν absolutely continuous with respect to π. If we let $\tilde{c}(x, y) = \frac{1}{t}c(x, y)$, then this is equivalent to the inequality:

$$W_{\tilde{c}}(\nu, \pi) \leq D(\nu||\pi)$$

holding for each ν absolutely continuous with respect to π. In the literature the constant is often taken to be one by absorbing it into the cost function, but we will keep the constant explicit. As with the transportation inequality, the quadratic cost transportation inequality has a dual representation. It is defined in terms of an infimum convolution. The infimum convolution of the function $f : V \to \mathbb{R}$ with respect to the cost function $\frac{1}{t}c$ is defined to be:

$$Q_{\frac{1}{t}c}f(x) = \inf_{y \in V}\{f(y) + \frac{1}{t}c(x,y)\}.$$

Then [26] notes that the transportation inequality of (13) holds with constant t, for each function f on X, if and only if the infimum convolution inequality:

$$\sum_{x \in V} e^{Q_{\frac{1}{t}c}f(x)}\pi(x) \leq e^{\sum_{x \in V} f(x)\pi(x)} \tag{14}$$

holds with constant t for every f. For completeness we include a proof of this in Proposition 2.3.7.

An important reason that quadratic cost transportation inequalities are of interest is that they give dimension free concentration results. Consider for $i = 1, 2, \ldots, n$ metric spaces (X_i, d_i) with measures π_i and quadratic cost functions $c_i(x,y) = d_i^2(x,y)$. Assume that they satisfy quadratic cost transportation inequalities (13) with constants t_i. Let (X, d) be the metric space with $X = \prod_{i=1}^{n} X_i$ and $d(x,y) = \left(\sum_{i=1}^{n} d_i^2(x_i, y_i)\right)^{\frac{1}{2}}$. Let π be the product measure on X and $c(x,y) = d^2(x,y)$ be the quadratic cost function on X. Then (X, d) and π satisfy the quadratic cost transportation inequality with constant $\max_i t_i$. A proof of this using the infimum convolution inequality is found in [26], while Talagrand [36] proves this in $\mathbb{R}^n$ using the quadratic cost transportation inequality.

In the Euclidean setting, the quadratic cost transportation inequality is also favored because the solution to a quadratic cost transportation problem is unique and characterized by an optimal map which is the gradient of a convex function [13] (see also Gangbo and McCann [19] for results on general classes of cost functions and other references). This optimal map allows an interpolation between points in the space of square integrable probability measures on $\mathbb{R}^n$, along which the entropy functional is convex. Otto and Villani [30] use this to prove an inequality that gives a partial converse to the fact that the log-Sobolev inequality implies the quadratic cost transportation inequality. Sturm and Renesse

[34] use these ideas to prove equivalent lower bounds on the Ricci curvature of Riemannian manifolds. Analogs of either of these results would be very welcome in the discrete setting.

Now we move on to the hurdles, first showing that under mild conditions no quadratic cost transportation inequality can hold on a graph. Let $G = (V, E)$ be a graph with associated cost function c and probability measure π. Suppose there exists $A \subset V$ for which $c(x, y) > 0$ for each $x \in A$ and $y \in V \setminus A$. If c is the square of the graph distance, for example, then any subset A works. We will show that for any fixed $t > 0$, the infimum convolution inequality:

$$\sum_{x \in V} e^{Q_t f(x)} \pi(x) \leq e^{\sum_{x \in V} f(x) \pi(x)}$$

will not hold for some function f on V. So let $t > 0$ be fixed, and let

$$m = \min\{c(x, y) : x, y \in V \text{ and } c(x, y) > 0\}.$$

Define $f : V \to \mathbb{R}$ by

$$f(x) = \begin{cases} \frac{1}{2t} m & x \in A \\ 0 & x \in V \setminus A. \end{cases}$$

For $x \in V \setminus A$, we have $Q_t f(x) = 0$. For $x \in A$, we have $f(x) \geq Q_t f(x) = \min_{y \in V}\{f(y) + \frac{1}{t} c(x, y)\} \geq \min\{f(x), \frac{m}{t}\} = f(x)$. So $Q_t f(x) = f(x)$ for each $x \in V$. Since e^t is a strictly convex function and $f(x)$ is not a constant function, Jensen's inequality gives us:

$$\sum_{x \in V} e^{Q_t f(x)} \pi(x) = \sum_{x \in V} e^{f(x)} \pi(x) > e^{\sum_{x \in V} f(x) \pi(x)}$$

Hence the infimum convolution inequality does not hold for f.

Addressing the other benefits of using the quadratic cost transportation problem in the continuous setting, we note that the quadratic cost transportation problem does not yield the same benefits in the discrete setting. Squaring the graph distance in the transportation problem does not guarantee unique solutions to Monge's problem. And the solutions to the transportation problem (whether using a squared distance or not), do not directly give a useful interpolation between points in the space of probability measures on the graph. We do not claim that these hurdles are insurmountable, and we believe there is hope that a better analog of the quadratic cost transportation problem and quadratic cost

16

transportation inequality will be developed. But for this thesis, we focus on what we *can* do with the mass transportation problem and the transportation inequality.

1.5 Summary of New Results

In Chapter 3 we provide the first systematic study of the constants $\rho_0(L)$, $\sigma^2(G)$, $\lambda_1(L)$, and $c^2(G)$ in the discrete setting.

- The main new result in this chapter is the inequality $\rho_0(L) \leq \frac{1}{2\sigma^2(G)}$, showing that the *modified* log-Sobolev inequality implies the transportation inequality.

In the continuous setting, the usual log-Sobolev inequality implies the quadratic cost transportation inequality (see [30, 9]). Our result is very natural in the discrete setting since the quadratic cost transportation inequality does not hold there, and since the modified log-Sobolev inequality is based on the derivative of the relative entropy. In the continuous setting, it does not seem to be settled whether or not the quadratic cost transportation inequality implies the usual log-Sobolev inequality. In the discrete setting we are able to say that the transportation inequality does not imply the modified log-Sobolev inequality, as it can be too weak to imply even the Poincaré inequality.

In light of our result above, it is natural to wonder if the weaker Poincaré inequality might also imply the (entropy) transportation inequality. We show that the class of bounded degree expander graphs provides an answer in the negative to this question. With this class of graphs, we also answer a question of Svante Janson, whether there is an infinite family of graphs for which $c^2(G) \ll \sigma^2(G)$ when π is the uniform probability measure. More precisely, we prove the following:

- Let $\{G_i\}_{i=1}^{\infty}$ be a family of bounded degree expander graphs (i.e. there exist positive constants k and ϵ so that the maximum degree of a vertex in G_i is bounded from above by k for each i, and the spectral gap of the Markov chain on G_i is bounded below by ϵ for each i). If G_i has n_i vertices, then $\sigma^2(G_i) \geq K \log n_i$ for some constant K, so that $\lambda_1(L_i) \gg \frac{1}{\sigma^2(G_i)}$ as $i \to \infty$.

Since $\frac{1}{2c^2(G_i)} \geq \lambda_1(L_i)$ for each i, we get $c^2(G_i) \ll \sigma^2(G_i)$ as $i \to \infty$. Since a bounded degree random graph is an expander with probability tending to 1 as n tends to infinity,

this shows that among bounded degree graphs, it is typically the case that $c^2(G) \ll \sigma^2(G)$. For an explicit reference (and additional references) to the fact that random graphs provide existence of expanders, see Section 4 of [2].

In Chapter 4 we calculate the asymptotically correct value of the subgaussian constant for cycles. In fact we show:

- $\sigma^2(C_{2k}) = c^2(C_{2k})$ and $c^2(C_{2k+1}) < \sigma^2(C_{2k+1}) = c^2(C_{2k+1})$, for positive integers k,

where C_n is the cycle on n vertices.

Through tensoring, this provides a concentration result on the discrete torus consisting of a product of both even and odd length cycles. The exact value of $\sigma^2(G)$ is notoriously hard. For $n \geq 5$, the exact value of $\sigma^2(C_n)$ was previously open and remains open for odd values of n. Furthermore, extremal sets in the isoperimetric problem on the discrete torus are not known unless the torus is a product of only even cycles. Our proof uses a new general fact that $\sigma^2(G) = c^2(C)$ unless there exists a probability measure $\nu \neq \pi$ for which $W^2(\nu, \pi) = 2\sigma^2(G)D(\nu||\pi)$, and it uses new facts concerning the solutions of Monge's and Kantorovich's problems with respect to ν and π under this condition.

Finally we make a couple of new observations that provide significant direction for future research. The first, at the end of Chapter 3, concerns the existence of fast mixing chains on a graph. In fact, for graphs with the uniform probability measure and diameter bounded by a polynomial in the log of the number of vertices, there exists a Markov chain with mixing time polynomial in the log of the number of vertices. This is interesting since Glauber dynamics takes time polynomial in size of the state space as opposed to polynomial in the log of the size of the state space (see for example [11]). This leads to the practical question of actually finding a fast mixing chain, as is guaranteed to exist. The second observation consists of Chapter 5, where we see that one characterization of a lower bound on the Ricci curvature of a Riemannian manifold has a well defined analog in the discrete setting closely related to the path coupling technique of proving fast mixing of Markov chains. This leads to the wide open research area of finding analogs in the discrete setting for propositions in the Riemannian setting involving Ricci curvature.

CHAPTER II

TECHNICAL LEMMAS AND PROPOSITIONS

In this chapter we put together some facts about solutions to the mass transportation problem, the Wasserstein distance, and the dual formulations of the transportation inequalities. Although interesting in their own right, they may be skipped and referred to as needed. As described in Chapter 1, our setting is a graph $G = (V, E)$ with associated measure π and distance function d, which throughout the chapter we assume is a metric on the vertices of the graph. A function g is said to be Lipschitz if $|g(x) - g(y)| \leq d(x, y)$ for $x, y \in V$. At times we will specialize to the graph distance on G. Then it suffices that $|g(x) - g(y)| \leq 1$ for $\{x, y\} \in E$. We also recall that $P(G)$ denotes the set of probability measures on V. And finally, the norm $\|\cdot\|$ denotes the l_1 on $\mathbb{R}^n$.

$$\|x\| = \sum_{i=1}^{n} |x_i| \quad \text{for } x \in \mathbb{R}^n.$$

We apply the norm to probability measures on V by embedding them in $\mathbb{R}^{|V|}$.

2.1 Solutions to the Mass Transportation Problem

Recall the definitions of Monge's and Kantorovich's problems and the Wasserstein distance from Section 1.3.

We start with the proposition, promised in Section 1.3, that Kantorovich's problem can be simplified in the case where the cost function d is a metric. Recall that the metric conditions are given in Section 1.3. The proof is derived from one by Feldman and McCann [18] done in the Riemannian setting. After this proposition, Kantorovich's problem will always refer to (5) since we do assume that d is a metric.

Proposition 2.1.1. *Suppose the distance function d is an actual metric. If g and h are a solution to Kantorovich's problem* (4) *with respect to ν_1 and ν_2 and $\tilde{g}$ is a solution to Kantorovich's problem* (5) *with respect to ν_1 and ν_2, then $K(g, h) = K(\tilde{g})$.*

Proof. Let K_1 be the value of Kantorovich's problem (4) and let K_2 be the value of Kantorovich's problem (5). First we show the easier direction that $K_1 \geq K_2$. Suppose $\tilde{g}$ is feasible in Kantorovich's problem (5) with respect to ν_1 and ν_2. Let $g = \tilde{g}$ and $h = -\tilde{g}$. Then for $x, y \in V$ we have

$$g(x) + h(y) = \tilde{g}(x) - \tilde{g}(y) \leq d(x, y). \tag{15}$$

So g and h are feasible in Kantorovich's problem (4). And

$$K(g, h) = \sum_{x \in V} g(x)\nu_1(x) + \sum_{x \in V} h(x)\nu_2(x) = \sum_{x \in V} \tilde{g}(x)(\nu_1(x) - \nu_2(x)) = K(\tilde{g}).$$

So $K_1 \geq K_2$.

Next we show that $K_1 \leq K_2$. Let g and h be feasible in Kantorovich's problem (4). We will define functions $\tilde{g}$ and $\tilde{h}$ and eventually show that $\tilde{g}$ is Lipschitz and $K(\tilde{g}) \geq K(g, h)$. Define $\tilde{g} : V \to \mathbb{R}$ by:

$$\tilde{g}(x) = \min_{y \in V}(d(x, y) - h(y)) \quad \text{for each } x \in V.$$

By the feasibility of g and h we have for every $x \in V$:

$$g(x) \leq d(x, y) - h(y) \quad \text{for every } y \in V.$$

Hence $\tilde{g}(x) \geq g(x)$. Next we define $\tilde{h} : V \to \mathbb{R}$ by

$$\tilde{h}(y) = \min_{x \in V}(d(x, y) - \tilde{g}(x)) \quad \text{for each } y \in V.$$

By the definition of $\tilde{g}(x)$ we have

$$h(y) \leq d(x, y) - \tilde{g}(x) \quad \text{for every } y \in V.$$

Hence $\tilde{h}(y) \geq h(y)$ for each $y \in V$. Using this and the definition of $\tilde{g}(x)$ we have:

$$\tilde{g}(x) \geq \min_{y \in V}(d(x, y) - \tilde{h}(y)).$$

But

$$\tilde{g}(x) \leq d(x, y) - \tilde{h}(y)$$

for every $x, y \in V$ by the definition of $\tilde{h}$. So in fact

$$\tilde{g}(x) = \min_{y \in V}(d(x, y) - \tilde{h}(y)) \quad \text{for each } x \in V.$$

In particular this shows that $\tilde{g}$ and $\tilde{h}$ are also feasible in Kantorovich's problem (4). Now assume to the contrary that there exists $z \in V$ such that $\tilde{g}(z) + \tilde{h}(z) < 0$. By the definitions of $\tilde{g}$ and $\tilde{h}$, there exist $x, y \in V$ such that $\tilde{g}(z) = d(z, y) - \tilde{h}(y)$ and $\tilde{h}(z) = d(z, x) - \tilde{g}(x)$. Then

$$\begin{aligned}
d(y, z) + d(z, x) &= d(z, y) + d(z, x) \\
&= \tilde{g}(z) + \tilde{h}(z) + \tilde{g}(x) + \tilde{h}(y) \\
&< \tilde{g}(x) + \tilde{h}(y) \\
&\leq d(x, y)
\end{aligned}$$

which contradicts the triangle inequality. So $\tilde{g}(z) + \tilde{h}(z) \geq 0$ for every $z \in V$. Together with $\tilde{g}(z) + \tilde{h}(z) \leq d(z, z) = 0$ we get that $\tilde{g}(z) + \tilde{h}(z) = 0$ for every $z \in V$. Then

$$\begin{aligned}
K(\tilde{g}) &= \sum_{x \in V} \tilde{g}(x)(\nu_1(x) - \nu_2(x)) \\
&= \sum_{x \in V} \tilde{g}(x)\nu_1(x) + \sum_{x \in V} \tilde{h}(x)\nu_2(x) \\
&\geq \sum_{x \in V} g(x)\nu_1(x) + \sum_{x \in V} th(x)\nu_2(x) \\
&= K(g, h).
\end{aligned}$$

The only metric property we have not used is the fact that $d(x, y) = 0$ only if $x = y$. So the proof would also go through if d where a pseudo-metric. $\square$

The next lemma is a well known fact.

Lemma 2.1.2. *Kantorovich's problem is translation invariant.*

Proof. Suppose g is a solution to Kantorovich's problem with respect to ν_1 and ν_2. Let c be a real number. For $x, y \in V$, $|(g(x) + c) - (g(y) + c)| = |g(x) - g(y)| \leq d(x, y)$, so $g + c$ is a Lipschitz function. Also, $\sum_{x \in V}(g(x) + c)(\nu_1(x) - \nu_2(x)) = \sum_{x \in V} g(x)(\nu_1(x) - \nu_2(x)) +$

$c \sum_{x \in V} (\nu_1(x) - \nu_2(x)) = W(\nu_1, \nu_2)$. Hence $g + c$ is a solution to Kantorovich's problem with respect to ν_1 and ν_2. $\qquad \square$

The following lemma is a consequence of well known properties of linear programs, but we give a full proof here for completeness.

Lemma 2.1.3. *Suppose the distance d associated with the graph G is the graph distance (so in particular it is integer valued). Suppose $\langle g_i \rangle_{i=1}^{k}$ is a sequence of solutions to Kantorovich's problem with respect to ν_1 and ν_2. Then for any sequence of non-negative constants $\langle s_i \rangle_{i=1}^{k}$ with $\sum_{i=1}^{k} s_i = 1$ we have $\sum_{i=1}^{k} s_i g_i$ is a solution to Kantorovich's problem with respect to ν_1 and ν_2. Furthermore, for any solution g to Kantorovich's problem with respect to ν_1 and ν_2, there exists a sequence of integer valued solutions $\langle g_i \rangle_{i=1}^{k}$ and a sequence of non-negative constants $\langle s_i \rangle_{i=1}^{k}$ with the properties that $\sum_{i=1}^{k} s_i = 1$ and $g = \sum_{i=1}^{k} s_i g_i$.*

Proof. The first statement is easier and we begin with it. Let $\langle g_i \rangle_{i=1}^{k}$ be a family of solutions to Kantorovich's problem with respect to ν_1 and ν_2. Let $\langle s_i \rangle_{i=1}^{k}$ be non-negative real numbers with $\sum_{i=1}^{k} s_i = 1$. Let $g = \sum_{i=1}^{k} s_i g_i$. We now show that g is Lipschitz. For $x, y \in V$ we have

$$
\begin{aligned}
|g(x) - g(y)| &= |\sum_{i=1}^{k} s_i(g_i(x) - g_i(y))| \\
&\leq \sum_{i=1}^{k} s_i |g_i(x) - g_i(y)| \\
&\leq \sum_{i=1}^{k} s_i d(x, y) \\
&= d(x, y).
\end{aligned}
$$

So g is feasible in Kantorovich's problem with respect to ν_1 and ν_2. Then

$$
\begin{aligned}
\sum_{x \in V} g(x)(\nu_1(x) - \nu_2(x)) &= \sum_{x \in V} \left(\sum_{i=1}^{k} s_i g_i(x) \right) (\nu_1(x) - \nu_2(x)) \\
&= \sum_{i=1}^{k} s_i \sum_{x \in V} g_i(x)(\nu_1(x) - \nu_2(x)) \\
&= \sum_{i=1}^{k} s_i W(\nu_1, \nu_2) \\
&= W(\nu_1, \nu_2).
\end{aligned}
$$

Hence g is a solution to Kantorovich's problem with respect to ν_1 and ν_2.

Now we prove the second statement. Recall that we make a general assumption that G is a connected graph. Let g be a solution to Kantorovich's problem with respect to ν_1 and ν_2. Define the graph $G_g = (V_g, E_g)$ by $V_g = V$ and for $x, y \in V_g$, $\{x, y\} \in E_g$ if and only if $\{x, y\} \in E$ and $|g(x) - g(y)| = 1$. Let $\{G_i = (V_i, E_i)\}_{i=1}^{n}$ be the set of connected components of G_g. The proof will be by induction on n.

For the base case assume $n = 1$. Then g is a translation of an integer valued function. So there exists a real number $c \in [0, 1)$ so that the function $g + c$ is integer valued. Recall that Kantorovich's problem is translation invariant (see Lemma 2.1.2). Let $g_1 = g + c$ and $g_2 = g + c - 1$. Let $s_1 = 1 - c$ and $s_2 = c$. Then g_1 and g_2 are integer valued solutions to Kantorovich's problem with respect to ν_1 and ν_2 and $s_1 g_1 + s_2 g_2 = (1 - c)(g + c) + c(g + c - 1) = g$.

Next assume that any solution $\tilde{g}$ to Kantorovich's problem for which $G_{\tilde{g}}$ has no more than n connected components $(n \geq 1)$ can be written as $\tilde{g} = \sum_{i=1}^{k} s_i g_i$, for some positive integer k and assume g is a solution to Kantorovich's problem for which G_g has $n + 1$ connected components. Let $G_1 = (V_1, E_1)$ be one of the connected components of G_g. Let

$$
m^+ = \max\{g(x) - g(y) : \{x, y\} \in E, \ x \in V_1, \ \text{and} \ y \notin V_1\}
$$

$$
m^- = \min\{g(x) - g(y) : \{x, y\} \in E, \ x \in V_1, \ \text{and} \ y \notin V_1\}
$$

23

and define $h^+ : V \to \mathbb{R}$ and $h^- : V \to \mathbb{R}$ by

$$h^+(x) = \begin{cases} g(x) + 1 - m^+, & x \in V_1 \\ g(x), & x \notin V_1 \end{cases} \qquad h^-(x) = \begin{cases} g(x) - 1 - m^-, & x \in V_1 \\ g(x), & x \notin V_1. \end{cases}$$

Let $x \in V_1$ and $y \notin V_1$ with $\{x, y\} \in E$. By the definitions of m^+ and m^- and because g is Lipschitz, we have $-2 < g(x) - g(y) - m^+ \leq 0$ and $0 \leq g(x) - g(y) - m^- < 2$. Suppose $h^+(x) \geq h^+(y)$. Then

$$|h^+(x) - h^+(y)| = g(x) - g(y) - m + 1 \leq 1.$$

If $h(x) \leq h(y)$, then

$$|h^+(x) - h^+(y)| = -(g(x) - g(y) - m) - 1 < 2 - 1 = 1.$$

Hence h^+ is Lipschitz. Suppose $h^-(x) \geq h^-(y)$. Then

$$|h^-(x) - h^-(y)| = g(x) - g(y) - m^- - 1 < 2 - 1 = 1.$$

If $h(x) \leq h(y)$, then

$$|h^-(x) - h^-(y)| = -(g(x) - g(y) - m^-) + 1 \leq 1.$$

Hence h^+ is Lipschitz.

Let $x \in V_1$ and $y \notin V_1$ with $\{x, y\} \in E$ and $g(x) - g(y) = m^+$. Then $h^+(x) - h^+(y) = g(x) - g(y) + 1 - m = 1$. So G_{h^+} has at least one less connected component than G_g. If $x \in V_1$ and $y \notin V_1$ with $\{x, y\} \in E$ and $g(x) - g(y) = m^-$, then $h^-(x) - h^-(y) = g(x) - g(y) - 1 - m^- = -1$ and G_{h^-} has at least one less connected component than G_g. Now both h^+ and h^- are feasible in Kantorovich's problem with respect to ν_1 and ν_2, so

$$W(\nu_1, \nu_2)$$
$$\geq \sum_{x \in V} h^+(x)(\nu_1(x) - \nu_2(x))$$
$$= \sum_{x \in V} g(x)(\nu_1(x) = \nu_2(x)) + \sum_{x \in V_1} (1 - m)(\nu_1(x) - \nu_2(x))$$
$$= W(\nu_1, \nu_2) + (1 - m^+) \sum_{x \in V_1} (\nu_1(x) - \nu_2(x))$$

24

and

$$W(\nu_1, \nu_2)$$

$$\geq \sum_{x \in V} h^-(x)(\nu_1(x) - \nu_2(x))$$

$$= \sum_{x \in V} g(x)(\nu_1(x) = \nu_2(x)) + \sum_{x \in V_1} (-1 - m)(\nu_1(x) - \nu_2(x))$$

$$= W(\nu_1, \nu_2) + (-1 - m^-) \sum_{x \in V_1} (\nu_1(x) - \nu_2(x))$$

Now $1 - m > 0$ and $-1 - m < 0$, so $\sum_{x \in V_1} (\nu_1(x) - \nu_2(x)) = 0$ which means the inequalities above are actually equalities and h^+ and h^- are both solutions to Kantorovich's problem with respect to ν_1 and ν_2.

By the induction hypothesis, h^+ can be written as $\sum_{i=1}^{k^+} s_i^+ g_i^+$ and h^- can be written as $\sum_{i=1}^{k^+} s_i^+ g_i^+$, where k^+ and k^- are positive integers, $\langle s_i^+ \rangle_{i=1}^{k^+}$ and $\langle s_i^- \rangle_{i=1}^{k^-}$ are sequences of non-negative integers with $\sum_{i=1}^{k^+} s_i^+ = \sum_{i=1}^{k^-} s_i^- = 1$. Let $k = k^+ + k^-$. Let $t = \frac{1 + m^-}{(1 + m^-) + (1 - m^+)}$. Note that $0 \leq t \leq 1$. We define a sequence of non-negative integers $\langle s_i \rangle_{i=1}^{k}$ by $s_i = t s_i^+$ for $i \in \{1, 2, \ldots, k^+\}$ and $s_i = (1 - t) s_{i-k^+}$ for $i \in \{k^+ + 1, k^+ + 2, \ldots, k^+ + k^-\}$. We define a sequence of integer valued solutions to Kantorovich's problem with respect to ν_1 and ν_2 by $g_i = h_i^+$ for $i \in \{1, 2 \ldots, k^+\}$ and $g_i = h_{i-k^+}^-$ for $i \in \{k^+ + 1, k^+ + 2, \ldots, k^+ + k^-\}$. Then $\sum_{i=1}^{k} s_i = 1$ and $\sum_{i=1}^{k} s_i g_i = t h^+ + (1 - t) h^-$. For $x \notin V_1$, $t h^+(x) + (1 - t) h^-(x) = t g(x) + (1 - t) g(x) = g(x)$. For $x \in V_1$, we have

$$t h^+(x) + (1 - t) h^-(x)$$

$$= \frac{1 + m^-}{(1 + m^-) + (1 - m^+)} (g(x) + 1 - m^+)$$

$$+ \left(1 - \frac{1 + m^-}{(1 + m^-) + (1 - m^+)} \right) (g(x) - 1 - m^-)$$

$$= g(x)$$

Hence $g = \sum_{i=1}^{k} s_i = 1$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Now we state a "complementary slackness" result for solutions to Monge's and Kantorovich's problems. As with Proposition 2.1.1, the proof we give is derived from a proof by Feldman and McCann [18] in a Riemannian setting. The result on the odd cycle in Chapter

4 is based on intuition provided by the definition of "transport rays" defined in [18] based on this lemma.

Lemma 2.1.4. *Let g be a Lipschitz function on V and let μ be a probability measure on $V \times V$ with marginals ν_1 and ν_2. Then*

$$g(x) - g(y) = d(x, y) \quad \text{for every } x, y \in V \text{ with } \mu(x, y) > 0 \tag{16}$$

if and only if g is a solution to Kantorovich's problem and μ is a solution to Monge's problem both with respect to ν_1 and ν_2.

Proof. Recall the definitions of the functions M and K defined in Section 1.3. Let g be a Lipschitz function on V and μ a probability measure on $V \times V$ with marginals ν_1 and ν_2. Then

$$
\begin{aligned}
M(\mu) &= \sum_{x,y \in V} d(x, y)\mu(x, y) \\
&\geq \sum_{x,y \in V} [g(x) - g(y)]\mu(x, y) \\
&= \sum_{x,y \in V} g(x)\mu(x, y) - \sum_{x,y \in V} g(y)\mu(x, y) \\
&= \sum_{x \in V} g(x) \sum_{y \in V} \mu(x, y) - \sum_{y \in V} g(y) \sum_{x \in V} \mu(x, y) \\
&= \sum_{x \in V} g(x)\nu_1(x) - \sum_{y \in V} g(y)\nu_2(y) \\
&= \sum_{x \in V} g(x)[\nu_1(x) - \nu_2(x)] \\
&= K(f)
\end{aligned}
\tag{17}
$$

Note that we have equality in 17 exactly when property 16 is satisfied. And by linear programming duality theory, $M(\mu) = K(f)$ if and only if μ is a solution to Kantorovich's problem and μ is a solution to Monge's problem, both with respect to ν_1 and ν_2. $\square$

Solutions to Monge's problem are in general not unique. In the following lemma we show that there always exists a solution to Monge's problem in which mass is never moved both into a vertex and out of the same vertex.

Lemma 2.1.5. *Suppose $\nu_1, \nu_2 \in P(G)$. Then there exists a solution μ to Monge's problem with respect to ν_1 and ν_2 with the following properties for every $y \in V$:*

1. If $\nu_1(y) \geq \nu_2(y)$ then $\mu(x, y) > 0$ implies that $x = y$.

2. If $\nu_1(y) \leq \nu_2(y)$ then $\mu(y, z) > 0$ implies that $z = y$.

Proof. We first show that there exists an optimal solution μ to Monge's problem with the property that there are no triples of distinct vertices (x, y, z) with $\mu(x, y) > 0$ and $\mu(y, z) > 0$. We will call such triples "bad." We will call a vertex "bad" if it is in the middle of a bad triple. Given any optimal solution μ to Monge's problem, if there exist one or more bad vertices we create a new optimal solution with one less bad vertex. This can be repeated until we have an optimal solution with no bad vertices and hence no bad triples.

Let μ be an optimal solution to Monge's problem with one or more bad vertices. Let f be an optimal solution to Kantorovich's problem. Let y be a bad vertex. It suffices to create an optimal solution $\tilde{\mu}$ to Monge's problem with one less bad triple centered on y and no bad vertices under $\tilde{\mu}$ that are not bad under μ. This can then be repeated until y is no longer a bad vertex and hence we have one less bad vertex.

Let (x, y, z) be a bad triple. Figure 1 shows how to create a new solution $\tilde{\mu}$ eliminating the bad triple. If $\mu(x, y) \geq \mu(y, z)$, then define $\tilde{\mu}$ by

$$\tilde{\mu}(x, y) = \mu(x, y) - \mu(y, z)$$
$$\tilde{\mu}(x, z) = \mu(x, z) + \mu(y, z)$$
$$\tilde{\mu}(y, z) = 0$$
$$\tilde{\mu}(y, y) = \mu(y, y) + \mu(y, z)$$
$$\tilde{\mu}(r, s) = \mu(r, s) \qquad \text{for all other pairs } (r, s) \in V \times V$$

If $\mu(x, y) < \mu(y, z)$, then define $\tilde{\mu}$ by

$$\tilde{\mu}(x, y) = 0$$
$$\tilde{\mu}(x, z) = \mu(x, z) + \mu(x, y)$$
$$\tilde{\mu}(y, z) = \mu(y, z) - \mu(x, y)$$
$$\tilde{\mu}(y, y) = \mu(y, y) + \mu(x, y)$$
$$\tilde{\mu}(r, s) = \mu(r, s) \qquad \text{for all other pairs } (r, s) \in V \times V$$

By direct calculation we have $\tilde{\mu} \in P(\nu_1, \nu_2)$. We will use Lemma 2.1.4 to verify the optimality of $\tilde{\mu}$. Suppose that for a pair of distinct vertices (r, s) we have $\tilde{\mu}(r, s) > 0$ and

27
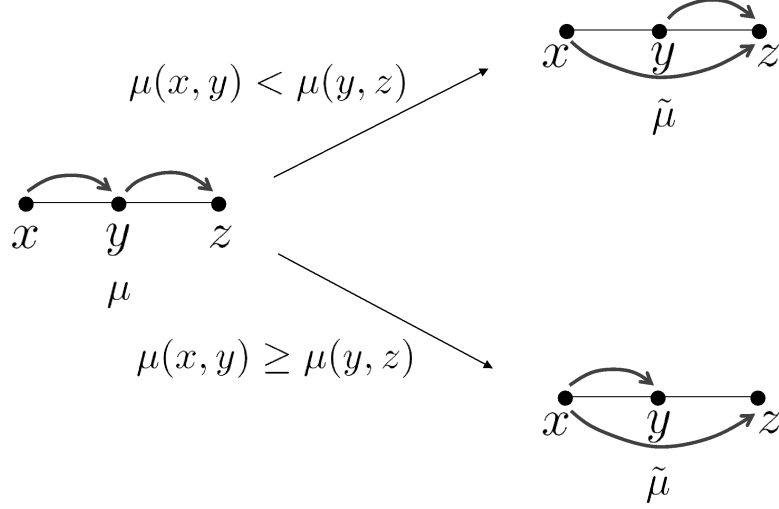
Figure 1: Eliminating a Bad Triple

$\mu(r,s) = 0$. By our definition of $\tilde{\mu}$ the only pair for which this could happen is (x,z). Hence we only need to check that $f(x) - f(z) = d(x,z)$. Since $\mu(x,y) > 0$ and $\mu(y,z) > 0$ we have $f(x) - f(y) = d(x,y)$ and $f(y) - f(z) = d(y,z)$. So $d(x,z) \geq f(x) - f(z) = f(x) - f(y) + f(y) - f(z) = d(x,y) + d(y,z) \geq d(x,z)$. So in fact $f(x) - f(z) = d(x,z)$. Hence $\tilde{\mu}$ is optimal. Now either $\tilde{\mu}(y,z) = 0$ or $\tilde{\mu}(x,y) = 0$, so x,y,z is not a bad triple under $\tilde{\mu}$. Again, since (x,z) is the only pair of vertices for which we could have $\tilde{\mu}(x,z) > 0$ but $\mu(x,z) = 0$, any triples centered on y other than (x,y,z) that are bad under $\tilde{\mu}$ are also bad under μ. So $\tilde{\mu}$ has one less bad triple centered on y. If there are any bad triples under $\tilde{\mu}$ that are not bad under μ, then they must have the form (r,x,z) or (x,z,s) for some r or s. If (r,x,z) is bad under $\tilde{\mu}$, then (r,x,y) is bad under μ and so x is already a bad vertex under μ. If (x,z,s) is bad under $\tilde{\mu}$, then (y,z,s) is bad under μ and so z is already a bad vertex under μ. Hence there are no vertices that are bad under $\tilde{\mu}$ that are not bad under μ.

Assume now that μ is an optimal solution to Monge's problem with no bad triples. Let $y \in V$. Suppose $\nu_1(y) \geq \nu_2(y)$ and $\mu(x,y) > 0$. Assume to the contrary that $x \neq y$. If $\mu(y,z) > 0$, then $z = y$ or else (x,y,z) would be a bad triple. So $\nu_2(y) \leq \nu_1(y) = \sum_{z \in V} \mu(y,z) = \mu(y,y) < \sum_{v \in V} \mu(v,y) = \nu_2(y)$, which is a contradiction. Now suppose that $\nu_1(y) \leq \nu_2(y)$ and $\mu(y,z) > 0$. Assume to the contrary that $z \neq y$. If $\mu(x,y) > 0$, then $x = y$ or else (x,y,z) would be a bad triple. So $\nu_2(y) \geq \nu_1(y) = \sum_{v \in V} \mu(y,v) > \mu(y,y) =$

28

$\sum_{x \in V} \mu(x, y) = \nu_2(y)$, which is a contradiction. $\qquad\square$

Under more specialized circumstances we show that there exists a solution to Monge's problem in which mass is only transferred to neighbors. In the proof we use Lemma 2.2.2 which is in the next section (although out of order, it fits better there).

Lemma 2.1.6. *Suppose d is the graph distance on G. Let $\nu_1 \in P(G)$. If $\nu_1(x) > 0$ for each $x \in V$ then there exists $\epsilon > 0$ so that for $\nu_2 \in P(G)$ with $\|\nu_1 - \nu_2\| \leq \epsilon$, there exists a solution μ to Monge's problem with respect to ν_1 and ν_2 with the property that for $x, y \in V$ with $x \neq y$, $\mu(x, y) = 0$ if $\{x, y\} \notin E$.*

Proof. Let D be the diameter of G. Let $\nu_1 \in P(G)$. Let $\epsilon = \frac{2}{D} \min_{x \in V} \nu_1(x)$. Let $\nu_2 \in P(G)$ with $\|\nu_1 - \nu_2\| \leq \epsilon$. For any solution μ to Monge's problem with respect to ν_1 and ν_2 let

$$B_\mu = \{(x, y) : x \neq y, \ \{x, y\} \notin E, \text{ and } \mu(x, y) > 0\}.$$

It suffices to show that for any solution μ with $|B_\mu| > 0$ there exists a solution $\tilde{\mu}$ with $|B_{\tilde{\mu}}| = |B_\mu| - 1$. So we assume μ is a solution to Monge's problem with respect to ν_1 and ν_2 with $|B_\mu| > 0$. Let $z, w \in V$ with $z \neq w$, $\{z, w\} \notin E$ and $\mu(z, w) > 0$. Let $z = x_0, x_1, x_2, \ldots, x_n = w$ be the vertices in a shortest path from z to w, where $n = d(z, w)$. Now define $\tilde{\mu}$ by

$$\tilde{\mu}(x, y) = \begin{cases} 0, & x = z \text{ and } y = w \\ \mu(x_i, x_{i+1}) + \mu(z, w), & x = x_i \text{ and } y = x_{i+1} \text{ for } i \in \{0, 1, \ldots, n-1\} \\ \mu(x_i, x_i) - \mu(z, w), & x = y = x_i \text{ for } i \in \{1, 2, \ldots, n-1\} \\ \mu(x, y), & otherwise \end{cases}$$

as shown in Figure 2. Then by definition the definition of $\tilde{\mu}$ we get that $|B_{\tilde{\mu}}| = |B_\mu| - 1$, as long $\tilde{\mu}$ is a solution to Monge's problem with respect to ν_1 and ν_2.

First we show that $\tilde{\mu}$ is feasible in Monge's problem with respect to ν_1 and ν_2. The marginals of $\tilde{\mu}$ and μ are the same, since $\tilde{\mu}$ is defined to have the same total amount of mass leave each vertex and the same total amount of mass enter each vertex as under μ (where mass remaining at a vertex is counted as both leaving and entering). But we also
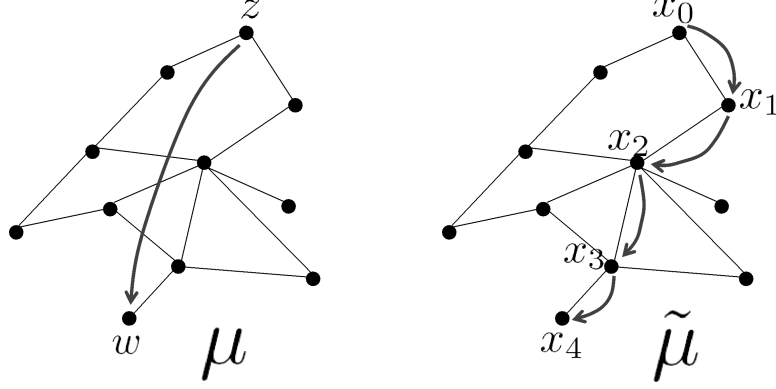
Figure 2: Nearest Neighbor Solution to Monge's Problem

need to ensure the $\tilde{\mu}(x, y) \geq 0$ for each $x, y \in V$. It suffices to check that $\tilde{\mu}(x_i, x_i) \geq 0$ for $i \in \{1, 2, \ldots, n-1\}$. So we use our assumption that $\nu_1(x) \geq \frac{D}{2}\|\nu_1 - \nu_2\|$ for each $x \in V$ and Lemma 2.2.2 to get:

$$
\begin{aligned}
\tilde{\mu}(x_i, x_i) &= \mu(x_i, x_i) - \mu(z, w) \\[2mm]
&= \nu_1(x_i) - \sum_{\substack{y \in V \\ y \neq x_i}} \mu(x_i, y) - \mu(z, w) \\[2mm]
&\geq \frac{D}{2}\|\nu_1 - \nu_2\| - \sum_{\substack{y \in V \\ y \neq x_i}} \mu(x_i, y) - \mu(z, w) \\[2mm]
&\geq W(\nu_1, \nu_2) - \sum_{\substack{y \in V \\ y \neq x_i}} \mu(x_i, y) - \mu(z, w) \\[2mm]
&= \sum_{x, y \in V} \mu(x, y) d(x, y) - \sum_{\substack{y \in V \\ y \neq x_i}} \mu(x_i, y) - \mu(z, w) \\[2mm]
&\geq \sum_{x \in V} \sum_{\substack{y \in V \\ y \neq x}} \mu(x, y) - \sum_{\substack{y \in V \\ y \neq x_i}} \mu(x_i, y) - \mu(z, w) \\[2mm]
&\geq \sum_{\substack{x \in V \\ x \notin \{x_i, z\}}} \sum_{\substack{y \in V \\ y \neq x}} \mu(x, y) \\[2mm]
&\geq 0.
\end{aligned}
$$

So $\tilde{\mu}$ is feasible in Monge's problem with respect to ν_1 and ν_2.

To finish the proof we show that $M(\tilde{\mu}) = M(\mu)$ and hence $\tilde{\mu}$ is a solution to Monge's problem with respect to ν_1 and ν_2. Recall that the function M is defined in Section 1.3. We have $M(\mu) - M(\tilde{\mu}) = \mu(z, w)d(z, w) - \sum_{i=0}^{n-1} \mu(z, w)d(x_i, x_{i+1}) = 0$ since $n = d(z, w)$. $\qquad\square$

2.2 Properties of the Wasserstein Distance

The first lemma of this section is a well known fact about the Wasserstein distance.

Lemma 2.2.1. *The Wasserstein distance is a metric on $P(G)$.*

Proof. Recall the definitions of the functions K and M from Section 1.3. First we show that $W(\nu_1, \nu_2) = 0$ if and only if $\nu_1 = \nu_2$. If $\nu_1 = \nu_2$, let g be a solution to Kantorovich's problem with respect to ν_1 and ν_2. Then

$$W(\nu_1, \nu_2) = K(g) = \sum_{x \in V} g(x)(\nu_1(x) - \nu_2(x)) = \sum_{x \in V} g(x)(\nu_1(x) - \nu_1(x)) = 0.$$

If $\nu_1 \neq \nu_2$, let μ be a solution to Monge's problem with respect to ν_1 and ν_2. If we had $\mu(x, y) = 0$ for each $x \neq y$, then μ would have the same first and second marginals implying that $\nu_1 = \nu_2$. Hence there exist $x^*, y^* \in V$ with $x^* \neq y^*$ and $\mu(x^*, y^*) > 0$. So

$$W(\nu_1, \nu_2) = M(\mu) = \sum_{x,y \in V} \mu(x, y) d(x, y) \geq \mu(x^*, y^*) d(x^*, y^*) > 0$$

since d is a metric on V.

Next we show that $W(\nu_1, \nu_2) = W(\nu_2, \nu_1)$ for $\nu_1, \nu_2 \in P(G)$. It suffices to show that for any $\nu_1, \nu_2 \in P(G)$, $W(\nu_1, \nu_2) \leq W(\nu_2, \nu_1)$. So let $\nu_1, \nu_2 \in P(G)$ and let g be a solution to Kantorovich's problem with respect to ν_1 and ν_2. Then $-g$ is a Lipschitz function, so it is feasible in Kantorovich's problem with respect to ν_2 and ν_1. Hence

$$
\begin{aligned}
W(\nu_1, \nu_2) &= \sum_{x \in V} g(x)(\nu_1(x) - \nu_2(x)) \\
&= \sum_{x \in V} (-g(x))(\nu_2(x) - \nu_1(x)) \\
&\leq W(\nu_2, \nu_1).
\end{aligned}
$$

Finally we prove the triangle inequality. Let $\nu_1, \nu_2, \nu_3 \in P(G)$. Let g be a solution to Kantorovich's problem with respect to ν_1 and ν_3. Then g is Lipschitz, so it is feasible in Kantorovich's problem with respect to ν_1 and ν_2 and in Kantorovich's problem with respect

to ν_2 and ν_3. So

$$W(\nu_1, \nu_3) = \sum_{x \in V} g(x)(\nu_1(x) - \nu_3(x))$$

$$= \sum_{x \in V} g(x)(\nu_1(x) - \nu_2(x)) + \sum_{x \in V} g(x)(\nu_2(x) - \nu_3(x))$$

$$\leq W(\nu_1(x), \nu_2(x)) + W(\nu_2(x), \nu_3(x)) \qquad \square$$

Now we compare the Wasserstein distance between ν_1 and ν_2 to the l_1 distance between ν_1 and ν_2.

Lemma 2.2.2. *Let D be the diameter of G and $\nu_1, \nu_2 \in P(G)$. Then*

$$\frac{1}{2}\|\nu_1 - \nu_2\| \leq W(\nu_1, \nu_2) \leq \frac{1}{2}D\|\nu_1 - \nu_2\|.$$

Proof. We can describe this inequality as follows. The quantity $\frac{1}{2}\|\nu - \pi\|$ is the total amount of mass that needs to be transported from one location to another. This mass must be moved a distance of at least one, but no more than a distance of D. To be precise, we let μ be a solution to Monge's problem given to us by Lemma 2.1.5. Then we may write

$$\frac{1}{2}\|\nu - \pi\| = \frac{1}{2}\sum_{x \in V} |\nu(x) - \pi(x)|$$

$$= \frac{1}{2}\sum_{x \in V} \left| \sum_{y \in V} \mu(x,y) - \sum_{y \in V} \mu(y,x) \right| \qquad (18)$$

$$= \frac{1}{2}\sum_{x \in V}\sum_{y \in V \setminus \{x\}} \mu(x,y) + \mu(y,x)$$

$$= \sum_{x \in V}\sum_{y \in V \setminus \{x\}} \mu(x,y).$$

In (18), either $\mu(x,y) = 0$ for each $y \neq x$ or $\mu(y,x) = 0$ for each $y \neq x$. So for each x, either $\sum_{y \neq x} \mu(x,y) = 0$ or $\sum_{y \neq x} \mu(y,x) = 0$. The next line then follows.

The desired inequality can then be translated to the following transparently true inequality:

$$\sum_{x \in V}\sum_{y \in V \setminus \{x\}} \mu(x,y) \leq \sum_{x \in V}\sum_{y \in V} d(x,y)\mu(x,y) \leq \sum_{x \in V}\sum_{y \in V \setminus \{x\}} D\mu(x,y)$$

which completes the proof. $\square$

Lemma 2.2.3. *The Wasserstein distance is continuous as a function of both probability measures.*

Proof. Let $\langle(\nu_i, \tilde{\nu}_i)\rangle_{i=1}^{\infty}$ be a sequence in $P(G) \times P(G)$ with $(\nu_i, \tilde{\nu}_i) \to (\nu, \tilde{\nu})$ as $i \to \infty$. Then

$$
\begin{aligned}
|W(\nu_i, \tilde{\nu}_i) - W(\nu, \tilde{\nu})| &= |W(\nu_i, \tilde{\nu}_i) - W(\nu_i, \tilde{\nu}) + W(\nu_i, \tilde{\nu}) - W(\nu, \tilde{\nu})| \\
&\leq |W(\nu_i, \tilde{\nu}_i) - W(\nu_i, \tilde{\nu})| + |W(\nu_i, \tilde{\nu}) - W(\nu, \tilde{\nu})| \\
&\leq W(\tilde{\nu}_i, \tilde{\nu}) + W(\nu_i, \nu) \\
&\leq \frac{D}{2}(\|\tilde{\nu}_i - \tilde{\nu}\| + \|\nu_i - \nu\|) \to 0
\end{aligned}
$$

as $i \to \infty$. The second inequality follows from the triangle inequality since W is a metric (see Lemma 2.2.1). The third inequality is from Lemma 2.2.2. $\qquad\square$

In the next lemma we show that solutions to Kantorovich's problem with respect to ν_1 and ν_2 are in a sense stable under small perturbations of ν_1 and ν_2. Suppose $\nu_1, \nu_2 \in P(G)$. Let $\epsilon_1 : V \to \mathbb{R}$ and $\epsilon_2 : V \to \mathbb{R}$ with the property that $\nu_1 + \epsilon_1$ and $\nu_2 + \epsilon_2$ are also probability measures on V. This simply means that $\sum_{x \in V} \epsilon_1(x) = \sum_{x \in V} \epsilon_2(x) = 0$ and $\nu_1(x) + \epsilon_1(x) \geq 0$ and $\nu_2(x) + \epsilon_2(x) \geq 0$ for each $x \in V$.

Lemma 2.2.4. *Suppose the distance function d on G is the graph distance. Then there exists $\delta > 0$ so that if $\|\epsilon_1\|, \|\epsilon_2\| < \delta$ then any solution to Kantorovich's problem with respect to $\nu_1 + \epsilon_1$ and $\nu_2 + \epsilon_2$ is also a solution to Kantorovich's problem with respect to ν_1 and ν_2.*

Proof. We prove the statement for integer valued solutions to Kantorovich's problem. The general case follows from the fact that convex combinations of solutions to Kantorovich's problem are also solutions and every solution to Kantorovich's problem can be written as a convex combination of integer valued solutions (see Lemma 2.1.3). Since Kantorovich's problem is translation invariant, without loss of generality we restrict our solutions to those whose images are contained in $[D] = \{1, 2, \ldots, D\}$, where D is the diameter of G. Let Z be the set of solutions to Kantorovich's problem with respect to ν_1 and ν_2 whose images are subsets of $[D]$. Let Z' be the set of integer valued Lipschitz functions on V whose images are subsets of $[D]$. By the feasibility condition for solutions to Kantorovich's problem, we

33

have that Z is a subset of Z'. Suppose $Z = Z'$. Then for any ϵ_1 and ϵ_2, any solution to Kantorovich's problem with respect to $\nu_1 + \epsilon_1$ and $\nu_2 + \epsilon_2$ whose image is a subset of $[D]$ is a member of $Z' = Z$, and hence would be a solution to Kantorovich's problem with respect to ν_1 and ν_2. So we would be done. Hence we assume that Z is a proper subset of Z'.

Since Z' is a finite set and Z is a proper subset of Z', we may define:

$$m = \min\left\{\left|\sum_{x\in V}(g(x) - g'(x))(\nu(x) - \mu(x))\right| : g \in Z \text{ and } g' \in Z' \setminus Z\right\}$$

Let $\delta = \frac{m}{4D}$. Note that $\delta > 0$. Now assume that $\|\epsilon_1\|, \|\epsilon_2\| < \delta$. Let Z_ϵ be the set of solutions to Kantorovich's problem with respect to $\nu_1 + \epsilon_1$ and $\nu_2 + \epsilon_2$ whose images are subsets of $[D]$. We will show that $Z_\epsilon \subset Z$.

First we find an upper bound on $|W(\nu_1, \nu_2) - W(\nu_1+\epsilon_1, \nu_2+\epsilon_2)|$. Let $g \in Z$ and $g_\epsilon \in Z_\epsilon$. Let h_1 be a solution to Kantorovich problem with respect to ν_1 and $\nu_1+\epsilon_1$ and h_2 a solution to Kantorovich's problem with respect to ν_2 and $\nu_2 + \epsilon_2$, both of whose images are subsets of $[D]$. Then

$$
\begin{aligned}
&|W(\nu_1, \nu_2) - W(\nu_1 + \epsilon_1, \nu_2 + \epsilon_2)| \\
={}& |W(\nu_1, \nu_2) - W(\nu_1, \nu_2 + \epsilon_2) + W(\nu_1, \nu_2 + \epsilon_2) - W(\nu_1 + \epsilon_1, \nu_2 + \epsilon_2)| \\
\leq{}& |W(\nu_1, \nu_2) - W(\nu_1, \nu_2 + \epsilon_2)| + |W(\nu_1, \nu_2 + \epsilon_2) - W(\nu_1 + \epsilon_1, \nu_2 + \epsilon_2)| \\
\leq{}& W(\nu_2, \nu_2 + \epsilon_2) + W(\nu_1, \nu_1 + \epsilon_1) \\
={}& \sum_{x\in V} h_2(x)(\nu_2(x) - (\nu_2(x) + \epsilon_2(x))) + h_1(x)(\nu_1(x) - (\nu_1(x) + \epsilon_1(x))) \\
={}& -\sum_{x\in V} h_2(x)\epsilon_2(x) + h_1(x)\epsilon_1(x) \\
\leq{}& \sum_{x\in V} h_2(x)|\epsilon_2(x)| + h_2(x)|\epsilon_1(x)| \\
\leq{}& D\sum_{x\in V} |\epsilon_1(x)| + |\epsilon_2(x)| \\
={}& D(\|\epsilon_1\| + \|\epsilon_2\|)
\end{aligned}
$$

34

Next we find a lower bound on $|W(\nu_1, \nu_2) - W(\nu_1 + \epsilon_1, \nu_2 + \epsilon_2)|$.

$$|W(\nu_1, \nu_2) - W(\nu_1 + \epsilon_1, \nu_2 + \epsilon_2)|$$

$$= \left| \sum_{x \in V} g(x)(\nu_1(x) - \nu_2(x)) - g_\epsilon(x)((\nu_1(x) + \epsilon_1(x)) - (\nu_2(x) + \epsilon_2(x))) \right|$$

$$= \left| \sum_{x \in V} (g(x) - g_\epsilon(x))(\nu_1(x) - \nu_2(x)) - \sum_{x \in V} g_\epsilon(x)(\epsilon_1(x) - \epsilon_2(x)) \right|$$

$$\geq \left| \sum_{x \in V} (g(x) - g_\epsilon(x))(\nu_1(x) - \nu_2(x)) \right| - \left| \sum_{x \in V} g_\epsilon(x)(\epsilon_1(x) - \epsilon_2(x)) \right|$$

$$\geq \left| \sum_{x \in V} (g(x) - g_\epsilon(x))(\nu_1(x) - \nu_2(x)) \right| - D \sum_{x \in V} |\epsilon_1(x)| + |\epsilon_2(x)|$$

$$= \left| \sum_{x \in V} (g(x) - g_\epsilon(x))(\nu_1(x) - \nu_2(x)) \right| - D(\|\epsilon_1(x)\| + \|\epsilon_2(x)\|)$$

Together, the upper and lower bounds give us:

$$\left| \sum_{x \in V} (g(x) - g_\epsilon(x))(\nu_1(x) - \nu_2(x)) \right| \leq 2D(\|\epsilon_1\| + \|\epsilon_2\|).$$

Suppose to the contrary that Z_ϵ is not a subset of Z. Let $\tilde{g}_\epsilon \in Z_\epsilon \setminus Z$. Then

$$
\begin{aligned}
m &\leq \left| \sum_{x \in V} (g(x) - \tilde{g}_\epsilon(x))(\nu(x) - \mu(x)) \right| \\
&\leq 2D(\|\epsilon_1\| + \|\epsilon_2\|) \\
&< 4D\delta \\
&= m.
\end{aligned}
$$

which is a contradiction. Hence $Z_\epsilon \subset Z$. $\qquad\square$

Now we give a condition under which the solution to Kantorovich's problem is unique up to translation.

Lemma 2.2.5. *The solution to Kantorovich's problem with respect to ν and π is unique up to translation if there does not exist $C \subsetneq V$ with $\sum_{x \in C} \nu(x) = \sum_{x \in C} \pi(x)$.*

Proof. Suppose g and $\tilde{g}$ are solutions to Kantorovich's problem with respect to ν and π. Assume there does not exist a constant c such that $g + c = \tilde{g}$. Let $x_0 \in V$. Let $C = \{x \in V : g(x) - \tilde{g}(x) = g(x_0) - \tilde{g}(x_0)\}$. Note that $x_0 \in C$. By assumption $C \neq V$. Let $z \in V \setminus C$. Let

μ be a solution to Monge's problem with respect to ν and π. Suppose to the contrary that there exists $x \in C$ with $\mu(x,z) > 0$. Then $\tilde{g}(x) - \tilde{g}(z) = g(x) - g(z) = d(x,z)$ by Lemma 2.1.4. But then we have $\tilde{g}(z) - g(z) = \tilde{g}(x) - g(x) = \tilde{g}(x_0) - g(x_0)$ which contradicts the fact that $z \notin C$. Similarly there does not exist $x \in C$ with $\mu(z,x) > 0$. Then

$$
\begin{aligned}
\sum_{x \in C} \nu(x) &= \sum_{x \in C} \sum_{y \in V} \mu(x,y) \\
&= \sum_{x \in C} \sum_{y \in C} \mu(x,y) \\
&= \sum_{y \in C} \sum_{x \in C} \mu(x,y) \\
&= \sum_{y \in C} \sum_{x \in V} \mu(x,y) \\
&= \sum_{y \in C} \pi(y)
\end{aligned}
$$

which proves the lemma. $\qquad\square$

2.2.1 Tensorization

Let $\{G_i\}_{i=1}^n$ be a family of graphs with associated measures π_i and distance functions d_i. Let $G = \prod_{i=1}^n G_i$ with associated measure $\pi(x) = \prod_{i=1}^n \pi_i(x_i)$ and distance function $d(x,y) = \sum_{i=1}^n d_i(x_i, y_i)$ as defined in Section 1.1.1. For $\nu \in P(G)$ let $\nu^i \in P(G_i)$ be defined by:

$$
\nu^i(x_i) = \sum_{(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)} \nu(x_1,\ldots,x_i,\ldots,x_n)
$$

for $x_i \in V_i$. See Talagrand [36] and Alon, Boppana, and Spencer [3] for related tensorization results.

Lemma 2.2.6. *For $\nu, \tilde{\nu} \in P(G)$, $W(\nu, \tilde{\nu}) \geq \sum_{i=1}^n W(\nu^i, \tilde{\nu}^i)$. Furthermore, if $\nu(x_1,\ldots,x_n) = \prod_{i=1}^n \nu^i(x_i)$ and $\tilde{\nu}(x_1,\ldots,x_n) = \prod_{i=1}^n \tilde{\nu}^i(x_i)$, then $W(\nu, \tilde{\nu}) = \sum_{i=1}^n W(\nu^i, \tilde{\nu}^i)$.*

Proof. Let h^i be a solution to Kantorovich's problem with respect to ν^i and $\tilde{\nu}^i$. Let $h : V \to \mathbb{R}$ be defined by

$$
h(x_1, x_2, \ldots, x_n) = \sum_{i=1}^n h^i(x_i).
$$

Note that

$$|h(x_1, x_2, \ldots, x_n) - h(y_1, y_2, \ldots, y_n)| \leq \sum_{i=1}^{n} |h(x_i) - h(y_i)|$$

$$\leq \sum_{i=1}^{n} d_i(x_i, y_i)$$

$$= d(x, y).$$

So h is feasible in Kantorovich's problem with respect to ν and $\tilde{\nu}$. Then:

$$W(\nu, \tilde{\nu})$$

$$\geq \sum_{(x_1, \ldots, x_n)} h(x_1, \ldots, x_n)(\nu(x_1, \ldots, x_n) - \tilde{\nu}(x_1, \ldots, x_n))$$

$$= \sum_{(x_1, \ldots, x_n)} \left(\sum_{k=1}^{n} h^k(x_k) \right) (\nu(x_1, \ldots, x_n) - \tilde{\nu}(x_1, \ldots, x_n))$$

$$= \sum_{k=1}^{n} \sum_{(x_1, \ldots, x_n)} h^k(x_k)(\nu(x_1, \ldots, x_n) - \tilde{\nu}(x_1, \ldots, x_n))$$

$$= \sum_{k=1}^{n} \sum_{x_k} h^k(x_k) \sum_{(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n)} (\nu(x_1, \ldots, x_n) - \tilde{\nu}(x_1, \ldots, x_n))$$

$$= \sum_{k=1}^{n} \sum_{x_k} h^k(x_k)(\nu^k(x_k) - \tilde{\nu}^k(x_k))$$

$$= \sum_{k=1}^{n} W(\nu^i, \tilde{\nu}^i).$$

Next we make the assumption that $\nu(x_1, \ldots, x_n) = \prod_{i=1}^{n} \nu^i(x_i)$ and $\tilde{\nu}(x_1, \ldots, x_n) = \prod_{i=1}^{n} \tilde{\nu}^i(x_i)$. Now we just need to show:

$$W(\nu, \tilde{\nu}) \leq \sum_{i=1}^{n} W(\nu^i, \tilde{\nu}^i).$$

Let μ^i be a solution to Monge's problem with respect to ν^i and $\tilde{\nu}^i$. Let $\mu : V \times V \to \mathbb{R}$ be defined by:

$$\mu((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \prod_{i=1}^{n} \mu^i(x_i, y_i).$$

We will show that μ is feasible in Monge's problem, which means it has first and second marginals ν and $\tilde{\nu}$ respectively. Since the calculation is similar we will only verify the first

marginal. Let $(x_1, \ldots, x_n) \in V$. Then

$$
\begin{aligned}
\sum_{(y_1,\ldots,y_n)\in V} \mu((x_1,\ldots,x_n),(y_1,\ldots,y_n)) &= \sum_{(y_1,\ldots,y_n)\in V} \prod_{i=1}^{n} \mu^i(x_i,y_i) \\
&= \prod_{i=1}^{n} \sum_{y_i \in V_i} \mu^i(x_i,y_i) \\
&= \prod_{i=1}^{n} \nu^i(x_i) \\
&= \nu(x).
\end{aligned}
$$

So now we have:

$$
\begin{aligned}
W(\nu,\tilde{\nu}) \\
\leq \quad & \sum_{x,y\in V} \mu(x,y)d(x,y) \\
= \quad & \sum_{\substack{(x_1,\ldots,x_n)\in V \\ (y_1,\ldots,y_n)\in V}} \prod_{i=1}^{n} \mu^i(x_i,y_i) \sum_{k=1}^{n} d_k(x_k,y_k) \\
= \quad & \sum_{k=1}^{n} \sum_{\substack{(x_1,\ldots,x_n)\in V \\ (y_1,\ldots,y_n)\in V}} \prod_{i=1}^{n} \mu^i(x_i,y_i) d_k(x_k,y_k) \\
= \quad & \sum_{k=1}^{n} \sum_{x_k,y_k\in V_k} \mu^k(x_k,y_k)d_k(x_k,y_k) \sum_{\substack{(x_1,\ldots,x_{k-1},x_{k+1},\ldots,x_n) \\ (y_1,\ldots,x_{k-1},x_{k+1},\ldots,y_n)}} \prod_{\substack{i=1 \\ i\neq k}}^{n} \mu^i(x_i,y_i) \\
= \quad & \sum_{k=1}^{n} \sum_{x_k,y_k\in V_k} \mu^k(x_k,y_k)d_k(x_k,y_k) \\
= \quad & \sum_{k=1}^{n} W(\nu^i,\tilde{\nu}^i)
\end{aligned}
$$

$\square$

2.2.2 Derivative of the Wasserstein Distance

We begin with a graph $G = (V,E)$ with associated measure π and assume that d is the graph distance. We denote derivatives from the right and left as $\frac{d}{dt}^{+}$ and $\frac{d}{dt}^{-}$ respectively.

Proposition 2.2.7. *Suppose $\nu_t^1, \nu_t^2 \in P(G)$ for each t in some real interval I around a. Suppose further that $\frac{d}{dt}^{+}\nu_t^1$ and $\frac{d}{dt}^{+}\nu_t^2$ exist at time $t = a$. Then $\frac{d}{dt}^{+}W(\nu_t^1,\nu_t^2)\Big|_{t=a}$ exists*

and equals:

$$\sum_{x \in V} g(x) \left(\left. \frac{d}{dt}^+ \nu_t^1(x) \right|_{t=a} - \left. \frac{d}{dt}^+ \nu_t^2(x) \right|_{t=a} \right)$$

for some solution g to Kantorovich's problem with respect to ν_a^1 and ν_a^2. Furthermore, an identical statements holds for derivatives from the left.

Proof. We will prove the statement for the derivatives from the right, as the proof for the derivatives from the left is the same. Let G be the set of integer valued Lipschitz functions on V that have zero as their minimum value. Note that G is a finite set. For each real $t \in I$, let $g_t \in G$ be a solution to Kantorovich's problem with respect to ν_t^1 and ν_t^2. Such a solution always exists by Lemmas 2.1.2 and 2.1.3. Let $G' \subset G$ be defined by $g \in G'$ if and only if for every $\epsilon > 0$ there exists $a < t < a + \epsilon$ with $g_t = g$. By Lemma 2.2.4 each $g \in G'$ is also a solution to Kantorovich's problem with respect to ν_a^1 and ν_a^2. Let $g \in G'$. Let $\langle h_i \rangle_{i=1}^\infty$ be a sequence of positive real numbers with $\lim_{i \to \infty} h_i = 0$ and $g_{a+h_i} = g$ for each integer $i \geq 1$. If $\left. \frac{d}{dt}^+ W(\nu_t^1, \nu_t^2) \right|_{t=a}$ exists, then it is equal to:

$$\lim_{i \to \infty} \frac{W(\nu_{a+h_i}^1, \nu_{a+h_i}^2) - W(\nu_a^1, \nu_a^2)}{h_i}$$

$$= \sum_{x \in V} g(x) \lim_{i \to \infty} \left(\frac{\nu_{a+h_i}^1(x) - \nu_a^1(x)}{h_i} - \frac{\nu_{a+h_i}^2(x) - \nu_a^2(x)}{h_i} \right)$$

$$= \sum_{x \in V} g(x) \left(\left. \frac{d}{dt}^+ \nu_t^1(x) \right|_{t=a} - \left. \frac{d}{dt}^+ \nu_t^2(x) \right|_{t=a} \right) \tag{19}$$

To show that $\lim_{s \to 0^+} \frac{W(\nu_{a+s}^1, \nu_{a+s}^2) - W(\nu_a^1, \nu_a^2)}{s}$ is equal to (19), it suffices to show that for every sequence of positive real numbers $\langle \epsilon_i \rangle_{i=1}^\infty$ with $\lim_{i \to \infty} \epsilon_i = 0$, there exists a subsequence $\langle \epsilon_{i_j} \rangle_{j=1}^\infty$ such that $\lim_{i \to \infty} \frac{W(\nu_{a+\epsilon_{i_j}}^1, \nu_{a+\epsilon_{i_j}}^2) - W(\nu_a^1, \nu_a^2)}{\epsilon_{i_j}}$ is equal to (19). So we let $\langle \epsilon_i \rangle_{i=1}^\infty$ be a sequence of positive real numbers with $\lim_{i \to \infty} \epsilon_i = 0$. Let $\langle \epsilon_{i_j} \rangle_{j=1}^\infty$ be a subsequence of $\langle \epsilon_i \rangle_{i=1}^\infty$ with the property that for some $\tilde{g} \in G'$, $g_{a+\epsilon_{i_j}} = \tilde{g}$ for every j. It suffices to show that

$$\sum_{x \in V} \tilde{g}(x) \left(\left. \frac{d}{dt}^+ \nu_t^1(x) \right|_{t=a} - \left. \frac{d}{dt}^+ \nu_t^2(x) \right|_{t=a} \right) = \sum_{x \in V} g(x) \left(\left. \frac{d}{dt}^+ \nu_t^1(x) \right|_{t=a} - \left. \frac{d}{dt}^+ \nu_t^2(x) \right|_{t=a} \right) \tag{20}$$

If $\tilde{g} = g$ we are done. Otherwise let $c = \min_{x \in V} g(x) - \tilde{g}(x)$. For non-negative integers k, let

$$C_k = \{x \in V : g(x) - \tilde{g}(x) \geq c + k\}.$$

Let $k \geq 1$ be an integer. Let $x_1 \in V \setminus C_k$ and $x_2 \in C_k$. Now $g(x_1) - \tilde{g}(x_1) < c + k$ and $g(x_2) - \tilde{g}(x_2) \geq c + k$. So

$$
\begin{aligned}
g(x_1) - g(x_2) &< \tilde{g}(x_1) + c + k - g(x_2) \\
&\leq \tilde{g}(x_1) + c + k - \tilde{g}(x_2) - c - k \\
&\leq d(x_1, x_2).
\end{aligned}
$$

Also,

$$
\begin{aligned}
\tilde{g}(x_2) - \tilde{g}(x_1) &< \tilde{g}(x_2) - g(x_1) + c + k \\
&\leq g(x_2) - c - k - g(x_1) + c + k \\
&\leq d(x_1, x_2).
\end{aligned}
$$

For each i, let μ_{a+h_i} be a solution to Monge's problem with respect to $\nu^1_{a+h_i}$ and $\nu^2_{a+h_i}$. For each j, let $\mu_{a+\epsilon_{i_j}}$ be a solution to Monge's problem with respect to $\nu^1_{a+\epsilon_{i_j}}$ and $\nu^2_{a+\epsilon_{i_j}}$. Then by Lemma 2.1.4 we get that $\mu_{a+h_i}(x, y) = 0$ when $x \in V \setminus C_k$ and $y \in C_k$. Also $\mu_{a+\epsilon_{i_j}}(x, y) = 0$ when $x \in C_k$ and $y \in V \setminus C_k$. So

$$
\begin{aligned}
\sum_{x \in V \setminus C_k} \nu^1_{a+h_i}(x) &= \sum_{x \in V \setminus C_k} \sum_{y \in V} \mu_{a+h_i}(x, y) \\
&= \sum_{x \in V \setminus C_k} \sum_{y \in V \setminus C_k} \mu_{a+h_i}(x, y) \\
&= \sum_{y \in V \setminus C_k} \sum_{x \in V \setminus C_k} \mu_{a+h_i}(x, y) \\
&\leq \sum_{y \in V \setminus C_k} \sum_{x \in V} \mu_{a+h_i}(x, y) \\
&= \sum_{y \in V \setminus C_k} \nu^2_{a+h_i}(y).
\end{aligned}
$$

This also means that

$$\sum_{x \in C_k} \nu^1_{a+h_i}(x) \geq \sum_{x \in C_k} \nu^2_{a+h_i}(x).$$

40

Similarly we have

$$\sum_{x \in C_k} \nu^1_{a+\epsilon_{i_j}}(x) = \sum_{x \in C_k} \sum_{y \in V} \mu_{a+\epsilon_{i_j}}(x,y)$$

$$= \sum_{x \in C_k} \sum_{y \in C_k} \mu_{a+\epsilon_{i_j}}(x,y)$$

$$= \sum_{y \in C_k} \sum_{x \in C_k} \mu_{a+\epsilon_{i_j}}(x,y)$$

$$\leq \sum_{y \in C_k} \sum_{x \in V} \mu_{a+\epsilon_{i+j}}(x,y)$$

$$= \sum_{y \in C_k} \nu^2_{a+\epsilon_{i_j}}(y).$$

This also means that

$$\sum_{x \in V \setminus C_k} \nu^1_{a+\epsilon_{i_j}}(x) \geq \sum_{x \in V \setminus C_k} \nu^2_{a+\epsilon_{i_j}}(x).$$

By similar methods, since g and $\tilde{g}$ are both solution of Kantorovich's problem with respect to ν^1_a and ν^2_a, we get that

$$\sum_{x \in C_k} \nu^1_a(x) = \sum_{x \in C_k} \nu^2_a(x)$$

Now $\left.\frac{d}{dt}^+ \nu^1_t(x)\right|_{t=a}$ and $\left.\frac{d}{dt}^+ \nu^2_t(x)\right|_{t=a}$ exist for each $x \in V$. Hence

$$\lim_{s \to 0^+} \sum_{x \in C_k} \frac{\nu^1_{a+s}(x) - \nu^1_a(x)}{s} \quad \text{and} \quad \lim_{s \to 0^+} \sum_{x \in C_k} \frac{\nu^2_{a+s}(x) - \nu^2_a(x)}{s}$$

exist. Let

$$L = \sum_{x \in C_k} \left(\left.\frac{d}{dt}^+ \nu^1_t(x)\right|_{t=a} - \left.\frac{d}{dt}^+ \nu^2_t(x)\right|_{t=a} \right).$$

Then

$$\lim_{s \to 0^+} \sum_{x \in C_k} \frac{\nu^1_{a+s}(x) - \nu^2_{a+s}(x)}{s}$$

$$= \lim_{s \to 0^+} \sum_{x \in C_k} \frac{\nu^1_{a+s}(x) - \nu^2_{a+s}(x)}{s} + \frac{\nu^1_a(x) - \nu^2_a(x)}{s}$$

$$= \sum_{x \in C_k} \lim_{s \to 0^+} \frac{\nu^1_{a+s}(x) - \nu^1_a(x)}{s} + \frac{\nu^2_{a+s}(x) - \nu^2_a(x)}{s}$$

$$= L$$

And so we show that $L = 0$.

$$0 \leq \lim_{i \to \infty} \sum_{x \in C_k} \frac{\nu^1_{a+h_i}(x) - \nu^2_a(x)}{h_i} = L = \lim_{j \to \infty} \sum_{x \in C_k} \frac{\nu^1_{a+\epsilon_{i_j}}(x) - \nu^2_a(x)}{\epsilon_{i_j}} \leq 0$$

Now we can show (20):

$$\sum_{x \in V} \tilde{g}(x) \left(\frac{d}{dt}^{+} \nu_t^1(x) \Big|_{t=a} - \frac{d}{dt}^{+} \nu_t^2(x) \Big|_{t=a} \right)$$

$$= \sum_{l=0}^{\infty} \sum_{x \in C_l \backslash C_{l+1}} \tilde{g}(x) \left(\frac{d}{dt}^{+} \nu_t^1(x) \Big|_{t=a} - \frac{d}{dt}^{+} \nu_t^2(x) \Big|_{t=a} \right)$$

$$= \sum_{l=0}^{\infty} \sum_{x \in C_l \backslash C_{l+1}} (g(x) - c - l) \left(\frac{d}{dt}^{+} \nu_t^1(x) \Big|_{t=a} - \frac{d}{dt}^{+} \nu_t^2(x) \Big|_{t=a} \right)$$

$$= \sum_{x \in V} (g(x) - c) \left(\frac{d}{dt}^{+} \nu_t^1(x) \Big|_{t=a} - \frac{d}{dt}^{+} \nu_t^2(x) \Big|_{t=a} \right)$$

$$- \sum_{l=0}^{\infty} \sum_{x \in C_l \backslash C_{l+1}} l \left(\frac{d}{dt}^{+} \nu_t^1(x) \Big|_{t=a} - \frac{d}{dt}^{+} \nu_t^2(x) \Big|_{t=a} \right)$$

$$= \sum_{x \in V} (g(x) - c) \left(\frac{d}{dt}^{+} \nu_t^1(x) \Big|_{t=a} - \frac{d}{dt}^{+} \nu_t^2(x) \Big|_{t=a} \right)$$

$$- \sum_{l=1}^{\infty} \sum_{x \in C_l} \left(\frac{d}{dt}^{+} \nu_t^1(x) \Big|_{t=a} - \frac{d}{dt}^{+} \nu_t^2(x) \Big|_{t=a} \right)$$

$$= \sum_{x \in V} g(x) \left(\frac{d}{dt}^{+} \nu_t^1(x) \Big|_{t=a} - \frac{d}{dt}^{+} \nu_t^2(x) \Big|_{t=a} \right) \qquad \square$$

The following corollary follows immediately from the previous proposition, so we state it without further proof. We mention it because with Lemma 2.2.5 we have sufficient conditions for the existence of the derivative of the Wasserstein distance.

Corollary 2.2.8. *Suppose* $\nu_t^1, \nu_t^2 \in P(G)$ *for each* t *in some real interval* I *around* a, *and that* $\frac{d}{dt}\nu_t^1$ *and* $\frac{d}{dt}\nu_t^2$ *exist at time* $t = a$. *Suppose further that up to translation there exists a unique solution* g_a *to Kantorovich's problem with respect to* ν_a^1 *and* ν_a^2. *Then* $\frac{d}{dt}W(\nu_t^1, \nu_t^2)\big|_{t=a}$ *exists and equals:*

$$\sum_{x \in V} g_a(x) \left(\frac{d}{dt}\nu_t^1(x) \Big|_{t=a} - \frac{d}{dt}\nu_t^2(x) \Big|_{t=a} \right).$$

2.3 Dual Formulations of the Transportation Inequalities

This section begins by proving the equivalence of the dual formulations of both the transportation inequality and the variance transportation inequalities as stated in Section 1.4.1. We start with the variance transportation inequality because it is simpler.

Proposition 2.3.1. *Let $\hat{c}^2$ be the smallest constant for which the variance transportation inequality holds for all ν absolutely continuous with respect to π. Let c^2 be the maximum value of the variance of a Lipschitz function on V. Then $\hat{c}^2 = c^2$.*

Proof. We have

$$\hat{c}^2 = \sup_{\nu} \frac{W^2(\nu, \pi)}{\mathrm{Var}\left[\frac{d\nu}{d\pi}\right]}$$

where the supremum is over ν absolutely continuous with respect to π. Let $\epsilon > 0$. Let $\tilde{\nu} \in P(G)$ with $\tilde{\nu}$ absolutely continuous with respect to π and

$$\hat{c}^2 \leq \frac{W^2(\tilde{\nu}, \pi)}{\mathrm{Var}\left[\frac{d\tilde{\nu}}{d\pi}\right]} + \epsilon$$

Let g be an optimal solution to Kantorovich's problem with respect to $\tilde{\nu}$ and π. Since Kantorovich's problem is translation invariant (see Lemma 2.1.2), we may also assume that $\sum_{x \in V} g(x)\pi(x) = 0$. Let f be the density of $\tilde{\nu}$ with respect to π. Then

$$
\begin{aligned}
\hat{c}^2 &\leq \frac{W^2(\tilde{\nu}, \pi)}{\mathrm{Var}(f)} + \epsilon \\
&= \frac{\left(\sum_{x \in V} g(x)(\tilde{\nu}(x) - \pi(x))\right)^2}{\sum_{x \in V}\left(f(x) - \sum_{y \in V} f(y)\pi(y)\right)^2 \pi(x)} + \epsilon \\
&= \frac{\left(\sum_{x \in V} g(x)(f(x) - 1)\pi(x)\right)^2}{\sum_{x \in V}(f(x) - 1)^2 \pi(x)} + \epsilon \\
&\leq \sum_{x \in V} g(x)^2 \pi(x) + \epsilon \\
&= \mathrm{Var}(g) + \epsilon \\
&\leq \max\{\mathrm{Var}(\tilde{g}) : \tilde{g} \text{ is Lipschitz}\} + \epsilon
\end{aligned}
$$

where the second inequality is by the Cauchy-Schwartz inequality. Since ϵ is arbitrary, we have $\hat{c}^2 \leq c^2$.

To show the reverse inequality, let g be a Lipschitz function which attains the maximum variance and for which $\sum_{x \in V} g(x)\pi(x) = 0$. Let $d\nu_\delta = (1 + \delta g)d\pi$ for some δ small enough so that $\nu_\delta(x) > 0$ for each $x \in V$. Let h_δ be a solution to Kantorovich's problem with respect to ν_δ and π. Then since g is feasible in Kantorovich's problem with respect to ν_δ

and π we have:

$$\hat{c}^2 \geq \frac{W^2(\nu_\delta, \pi)}{\mathrm{Var}(1 + \delta g)}$$

$$= \frac{\left(\sum_{x \in V} h_\delta(x)(\nu_\delta(x) - \pi(x))\right)^2}{\mathrm{Var}_\pi(1 + \delta g)}$$

$$\geq \frac{\left(\sum_{x \in V} g(x)(\nu_\delta(x) - \pi(x))\right)^2}{\mathrm{Var}_\pi(1 + \delta g)}$$

$$= \frac{\left(\sum_{x \in V} g(x)((1 + \delta g(x))\pi(x) - \pi(x))\right)^2}{\sum_{x \in V} \left((1 + \delta g(x)) - \sum_{x \in V}(1 + \delta g(x))\pi(x)\right)^2 \pi(x)}$$

$$= \frac{\delta^2 \left(\sum_{x \in V} g(x)^2 \pi(x)\right)^2}{\delta^2 \sum_{x \in V} g(x)^2 \pi(x)}$$

$$= \mathrm{Var}(g)$$

$$= c^2.$$

And so we have $\hat{c}^2 = c^2$. $\qquad\qquad\square$

Now we re-derive the equivalence by Bobkov and Götze [7] of the transportation inequality and its dual, paying close attention to the state of optimality in both formulations. Proposition 2.3.2 below states the result of [7] and the following Proposition 2.3.3 is our refinement of it.

Proposition 2.3.2 (Bobkov-Götze). *Let σ be a positive real number. Then the following two statements are equivalent.*

1. *$\mathrm{E}\left[e^{t(f-\mathrm{E}[f])}\right] \leq e^{\sigma^2 t^2/2}$ for every Lipschitz function f and real number t.*

2. *$W_1^2(\nu, \pi) \leq 2\sigma^2 D(\nu || \pi)$ for every measure ν absolutely continuous with respect to π.*

Proposition 2.3.3. *Suppose that σ is a positive real number for which the two statements in Proposition 2.3.2 are true. Then we have:*

(a) *Suppose that there exists a Lipschitz function f and real number $t > 0$ with the property that $\mathrm{E}\left[e^{t(f-\mathrm{E}[f])}\right] = e^{\sigma^2 t^2/2}$. Define ν by $d\nu = e^{t(f-\mathrm{E}[f])-\sigma^2 t^2/2}d\pi$. Then we have $\nu \in P(G)$ with $\nu \neq \pi$ and $W_1^2(\nu, \pi) = 2\sigma^2 D(\nu || \pi)$. Furthermore, f is a solution to Kantorovich's problem with respect to ν and π and $t^2 = \frac{2}{\sigma^2} D(\nu || \pi)$.*

(b) Suppose σ is a positive real number for which the above two statements are true. Now suppose there exists $\nu \in P(G)$ with $\nu \neq \pi$ and $W_1^2(\nu, \pi) = 2\sigma^2 D(\nu||\pi)$. Let f be an solution to Kantorovich's problem with respect to ν and π. Then f and ν are related by $d\nu = e^{t(f-\mathrm{E}[f])-\sigma^2 t^2/2} d\pi$. And we have $\mathrm{E}\left[e^{t(f-\mathrm{E}[f])}\right] = e^{\sigma^2 t^2/2}$ for $t = \sqrt{\frac{2}{\sigma^2} D(\nu||\pi)}$.

Before proving Proposition 2.3.3, we prove a useful little corollary.

Corollary 2.3.4. *Suppose $\nu \in P(G)$ with $\nu \neq \pi$ and $W_1^2(\nu, \pi) = 2\sigma^2 D(\nu||\pi)$. Then up to translation there exists a unique solution f to Kantorovich's problem with respect to ν and π. And for each $x, y \in V$, $f(x) > f(y)$ if and only if $\frac{\nu(x)}{\pi(x)} > \frac{\nu(y)}{\pi(y)}$.*

Proof of Corollary 2.3.4. Suppose f and g are solutions to Kantorovich's problem with respect to ν and π. Then by Proposition 2.3.2 they are related by $e^{t(f-\mathrm{E}[f])-\sigma^2 t^2/2} = e^{t(g-\mathrm{E}[g])-\sigma^2 t^2/2}$. Hence $f - g = E[g - f]$ which is a constant, proving the first part. The second part follows directly from the fact that $\frac{\nu(x)}{\pi(x)} = e^{t(f(x)-\mathrm{E}[f])-\sigma^2 t^2/2}$ for each $x \in V$. $\square$

We need the following two lemmas in our proof of Proposition 2.3.3. The first is a version of a well known inequality whose proof we include for completeness.

Lemma 2.3.5 (Young's Inequality).

$$uv \leq u \, log \, u - u + e^v, \quad u \geq 0, \quad v \in \mathbb{R} \tag{21}$$

And equality occurs if and only if $u = e^v$.

Proof. If $u = 0$, the statement is true because we define $0 \log 0 = 0$ by continuity (or convention). Note that equality cannot occur if $u = 0$. To obtain the inequality when $u > 0$, fix $v \in \mathbb{R}$. Define $r(u) = uv - u \, log \, u + u - e^v$ for $u > 0$. Then $r'(u) = v - log \, u$. And $r''(u) = -\frac{1}{u} < 0$ for $u > 0$. So $r(u)$ is strictly concave down and has a maximum at $u = e^v$. Hence for all $u > 0$, $r(u) \leq r(e^v) = 0$, which gives us the inequality. Finally, since $r(u)$ is strictly concave, $r(u) < 0$ for $u \neq e^v$. $\square$

The next is a technical lemma based on Young's inequality.

45

Lemma 2.3.6.

$$E[e^h] \leq 1 \iff E[gh] \leq \text{Ent}[g] \textit{ for every density } g. \tag{22}$$

And if either side is true, then $E[gh] = \text{Ent}[g]$ *for some density* g *if and only if* $E[e^h] = 1$ *and* $g = e^h$.

Proof. Suppose g is a probability density on V with respect to π and that $h : V \to \mathbb{R}$. For $x \in V$, apply Young's inequality with $u = g(x)$ and $v = h(x)$ to get:

$$g(x)h(x) \leq g(x) \log g(x) - g(x) + e^{h(x)}. \tag{23}$$

Note that equality holds if and only if $g(x) = e^{h(x)}$. Then we can take expectations of both sides to get:

$$E[gh] \leq \text{Ent}[g] - 1 + E[e^h] \tag{24}$$

where there is equality if and only if $g = e^h$ (i.e. $g(x) = e^{h(x)}$ for all $x \in V$).

If $E[e^h] \leq 1$, then

$$E[gh] \leq \text{Ent}[g] \tag{25}$$

with equality if and only if $E[e^h] = 1$ and $g = e^h$. So $E[e^h] \leq 1$ implies that $E[gh] \leq \text{Ent}[g]$ for every density g, with equality if and only if $E[e^h] = 1$ and $g = e^h$.

Now suppose that for some h we have $E[gh] \leq \text{Ent}[g]$ for every density g. Choose $c > 0$ so that $E[ce^h] = 1$. Let $g = ce^h$. Then g is a density and $E[gh] \leq \text{Ent}[g]$ tells us that

$$cE[he^h] \leq cE[e^h(\log c + h)] \tag{26}$$

This implies that $(\log c)E[e^h] \geq 0$, so $c \geq 1$, and hence $E[e^h] \leq 1$. Then by the previous paragraph, we have $E[gh] = \text{Ent}[g]$ if and only if $E[e^h] = 1$ and $g = e^h$. Hence the lemma. $\square$

At last we get to the proof of Proposition 2.3.3

Proof of Proposition 2.3.3.

Part (a). First suppose there exists a positive real number σ such that for all real t and Lipschitz f we have

$$E\left[e^{t(f - E[f])}\right] \leq e^{\sigma^2 t^2/2} \tag{27}$$

46

or equivalently

$$E\left[e^{t(f-E[f])-\sigma^2 t^2/2}\right] \le 1. \tag{28}$$

Now suppose there exists a Lipschitz function $\tilde{f}$ and a real number $\tilde{t} > 0$ with the property that

$$E\left[e^{\tilde{t}(\tilde{f}-E[\tilde{f}])-\sigma^2\tilde{t}^2/2}\right] = 1. \tag{29}$$

Note that $\tilde{f}$ cannot be a constant function since $\tilde{t} \ne 0$. Now for any real t and Lipschitz f we can set $h = t(f - E[f]) - \sigma^2 t^2/2$ and use the preliminary result above to get:

$$E\left[\left(t(f - E[f]) - \sigma^2 t^2/2\right) g\right] \le \text{Ent}[g] \tag{30}$$

for every density g. Let $\tilde{g} = e^{\tilde{t}(\tilde{f}-E[\tilde{f}])-\sigma^2\tilde{t}^2/2}$. Then there is equality when $f = \tilde{f}$, $t = \tilde{t}$, and $g = \tilde{g}$. Simplifying and rearranging, we get that for all Lipschitz f and $t > 0$:

$$E[fg - f] \le \frac{\sigma^2 t}{2} + \frac{1}{t}\,\text{Ent}[g] \tag{31}$$

for every density g, with equality when $f = \tilde{f}$, $t = \tilde{t}$, and $g = \tilde{g}$. Now for a fixed non-constant density g consider the function

$$\phi_g(t) = \frac{\sigma^2 t}{2} + \frac{1}{t}\,\text{Ent}[g] \tag{32}$$

defined on positive t. We have

$$\phi_g'(t) = \frac{\sigma^2}{2} - \frac{1}{t^2}\,\text{Ent}[g] \tag{33}$$

which is zero if and only if

$$t = t^*(g) = \frac{\sqrt{2\,\text{Ent}[g]}}{\sigma}. \tag{34}$$

Furthermore we have

$$\phi_g''(t) = \frac{2}{t^3}\,\text{Ent}[g] > 0 \tag{35}$$

for every positive t. Hence $t^*(g)$ is the unique minimum of $\phi_g(t)$. Now

$$\phi_g(t^*(g)) = \sqrt{2\sigma^2\,\text{Ent}[g]}. \tag{36}$$

So for every Lipschitz f and $t > 0$ we have

$$E[fg - f] \le \sqrt{2\sigma^2\,\text{Ent}[g]} \le \frac{\sigma^2 t}{2} + \frac{1}{t}\,\text{Ent}[g]. \tag{37}$$

47

for every density g, with equality both places when $f = \tilde{f}$, $t = \tilde{t}$, and $g = \tilde{g}$. Since $\tilde{f}$ is not a constant function and $\tilde{t} \neq 0$ we get that $\tilde{g}$ is not a constant density. So $t^*(\tilde{g})$ is the unique minimum of $\phi_{\tilde{g}}(t)$ giving us $\tilde{t} = t^*(\tilde{g}) = \frac{\sqrt{2\,\mathrm{Ent}[\tilde{g}]}}{\sigma}$, since $\tilde{t}$ also minimizes $\phi_{\tilde{g}}(t)$. Let $d\tilde{\nu} = \tilde{g}\,d\pi$. Then in terms of probability measures ν instead of densities g, we have that for all Lipschitz f:

$$\sum_{x \in V} f(x)(\nu(x) - \pi(x)) \leq \sqrt{2\sigma^2 \mathrm{D}(\nu||\pi)}. \tag{38}$$

for every probability measure ν absolutely continuous with respect to π. There is equality when $f = \tilde{f}$ and $\nu = \tilde{\nu}$. Finally this tells us that

$$\mathrm{W}_1^2(\nu, \pi) \leq 2\sigma^2 \mathrm{D}(\nu||\pi) \tag{39}$$

for every ν absolutely continuous with respect to π. There is equality when $\nu = \tilde{\nu}$ and in this case $\tilde{f}$ is an optimal solution to Kantorovich's problem with respect to $\tilde{\nu}$ and π. And $\tilde{t}^2 = \frac{2}{\sigma^2} D(\tilde{\nu}||\pi)$.

Part (b). We start by assuming that there exists a positive real number σ with the property that for all probability measures ν absolutely continuous with respect to π we have

$$\mathrm{W}^2(\nu, \pi) \leq 2\sigma^2 \mathrm{D}(\nu||\pi). \tag{40}$$

Next suppose there exists a probability measure $\tilde{\nu} \neq \pi$ with

$$\mathrm{W}^2(\tilde{\nu}, \pi) = 2\sigma^2 \mathrm{D}(\tilde{\nu}||\pi). \tag{41}$$

Let $\tilde{f}$ be an optimal solution to Kantorovich's problem with respect to $\tilde{\nu}$ and π. Then we get

$$\sum_{x \in V} f(x)(\nu(x) - \pi(x)) \leq \sqrt{2\sigma^2 \mathrm{D}(\nu||\pi)} \tag{42}$$

for every Lipschitz f and ν absolutely continuous with respect to π, with equality if $f = \tilde{f}$ and $\nu = \tilde{\nu}$. Let $\tilde{g}$ be the density of $\tilde{\nu}$ with respect to π. Note that $\tilde{g}$ is not a constant function since $\tilde{\nu} \neq \pi$. Then we can rewrite this in terms of densities g with respect to π instead of measures ν getting:

$$\mathrm{E}\,[fg - f] \leq \sqrt{2\sigma^2 \,\mathrm{Ent}[g]} \tag{43}$$

for every Lipschitz f and density g, with equality if $f = \tilde{f}$ and $g = \tilde{g}$. Equivalently we can write:

$$\mathrm{E}\left[(f - \mathrm{E}[f])g\right] \leq \sqrt{2\sigma^2 \operatorname{Ent}[g]} \tag{44}$$

for every Lipschitz f and density g, with equality if $f = \tilde{f}$ and $g = \tilde{g}$. Furthermore,

$$\mathrm{E}\left[(f - \mathrm{E}[f])g\right] \leq \sqrt{2\sigma^2 \operatorname{Ent}[g]} \leq \frac{\sigma^2 t}{2} + \frac{1}{t}\operatorname{Ent}[g] \tag{45}$$

for every Lipschitz f, density g, and $t > 0$. Let $\tilde{t} = \frac{\sqrt{2\operatorname{Ent}[\tilde{g}]}}{\sigma}$, and note that $\tilde{t} > 0$. Then we have equality everywhere if $f = \tilde{f}$, $g = \tilde{g}$, and $t = \tilde{t}$. So we get

$$\mathrm{E}\left[\left(t(f - \mathrm{E}[f]) - \frac{\sigma^2 t^2}{2}\right)g\right] \leq \operatorname{Ent}[g]. \tag{46}$$

for every Lipschitz f, density g, and $t > 0$, with equality when $f = \tilde{f}$, $g = \tilde{g}$, and $t = \tilde{t}$. Let $h = t(f - \mathrm{E}[f]) - \frac{\sigma^2 t^2}{2}$. Then by our preliminary result we have:

$$\mathrm{E}\left[e^{t(f - \mathrm{E}[f]) - \frac{\sigma^2 t^2}{2}}\right] \leq 1 \tag{47}$$

for every Lipschitz f and $t > 0$, with equality when $f = \tilde{f}$ and $t = \tilde{t}$. And we have $\tilde{g} = e^{\tilde{t}(\tilde{f} - \mathrm{E}[\tilde{f}]) - \frac{\sigma^2 \tilde{t}^2}{2}}$. Finally we have

$$\mathrm{E}\left[e^{t(f - \mathrm{E}[f])}\right] \leq e^{\frac{\sigma^2 t^2}{2}} \tag{48}$$

for every Lipschitz f and real number t, with equality when $f = \tilde{f}$ and $t = \tilde{t}$. $\qquad \square$

Finally we prove the equivalence of the infimum convolution inequality and the quadratic cost transportation inequality as stated in Section 1.4.3. The proof is mentioned in [26] as an easy repeat of the proof of Proposition 2.3.2, but we spell it out for completeness.

Proposition 2.3.7. *The quadratic cost transportation inequality:*

$$W_{\frac{1}{t}c}(\nu, \pi) \leq D(\nu \| \pi)$$

holds with constant t for every ν absolutely continuous with respect to π if and only if the infimum convolution inequality:

$$\sum_{x \in V} e^{Q_{\frac{1}{t}c} f(x)} \pi(x) \leq e^{\sum_{x \in V} f(x)\pi(x)} \tag{49}$$

holds with constant t for every function f on V.

49

Proof. We start with 49 which is equivalent to

$$\mathrm{E}\left[e^{Q_{\frac{1}{t}c}f(x)-\mathrm{E}[f]}\right] \leq 1$$

for every f. By Lemma 2.3.6 this is equivalent to

$$\mathrm{E}\left[(Q_{\frac{1}{t}c}f - \mathrm{E}[f])g\right] \leq \mathrm{Ent}[g]$$

for every f and every density g. And this is equivalent to:

$$\mathrm{E}[(Q_{\frac{1}{t}c}f)g] - \mathrm{E}[f] \leq \mathrm{Ent}[g]$$

for every f and density g. Writing this in terms of a probability measure ν, instead of the density g, this is equivalent to:

$$\sum_{x \in V} Q_{\frac{1}{t}c}f(x)\nu(x) - \sum_{x \in V} f(x)\pi(x) \leq D(\nu||\pi)$$

for every f and probability measure ν absolutely continuous with respect to π. Taking the supremum over f of the left hand side, we see this is equivalent to:

$$W_{\frac{1}{t}c}(\nu, \pi) \leq D(\nu||\pi)$$

for every ν absolutely continuous with respect to π. $\qquad\square$

CHAPTER III

RELATIONSHIPS BETWEEN THE INEQUALITIES

Figure 3 provides an overview of the relationship between the inequalities in which we are interested, each of which is explained in Section 1.4. Section 3.1 gives proofs along with precise descriptions of the implications, including any assumptions hidden in the figure. Next, Section 3.2 explores the implications under different distances on the underlying graph structure. Finally, in Section 3.3 we take a look at what the implications say about the fastest mixing Markov chain problem. The main contribution here is the implication that $\rho_0(G) \leq \frac{1}{2\sigma^2(G)}$, and the bounds on maximal variance mentioned in the fastest mixing Markov chain section.

3.1 Descriptions and Proofs

Let L be the generator of a Markov chain on the graph $G = (V, E)$ with associated measure π and distance function d as described in Section 1.1. Throughout this chapter we assume that d is an actual metric.

3.1.1 Modified Log-Sobolev Implies Poincaré

The fact that the modified log-Sobolev inequality (11) implies the Poincaré inequality (10) with $\lambda_1 = \rho_0(L)$ is shown in [10]. The implication can be seen by taking functions $\frac{1}{n}f$
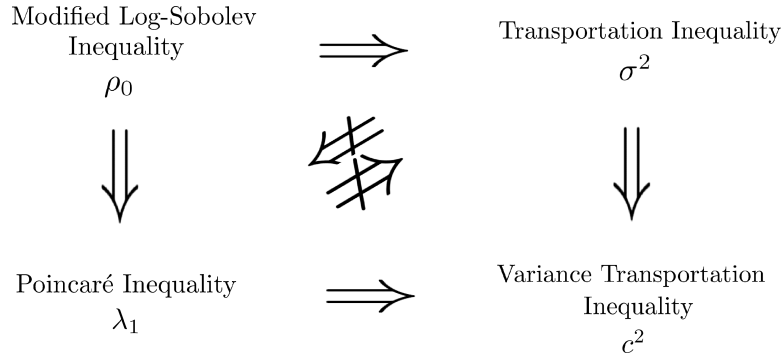
Figure 3: Implications Between Inequalities

in (11) and letting n approach infinity. This implication is equivalent to the inequality $\rho_0(L) \leq \lambda_1(L)$. Note that $\rho_0(L)$ and $\lambda_1(L)$ both depend on the specific Markov generator L and not just on the underlying graph structure of G.

3.1.2 Transportation Implies Variance Transportation

Next we prove that the transportation inequality (6) implies the variance transportation inequality (7) with $c^2 = \sigma^2(G)$. This is equivalent to showing the inequality $c^2(G) \leq \sigma^2(G)$. First recall the general inequality $E[f] \operatorname{Ent}[f] \leq \operatorname{Var}[f]$. When $d\nu = f d\pi$, this gives us $D(\nu \| \pi) \leq \operatorname{Var}[f]$. Hence for the optimal constant $\sigma^2(G)$ we have:

$$W^2(\nu, \pi) \leq 2\sigma^2(G) D(\nu \| \pi) \leq 2\sigma^2(G) \operatorname{Var}\left[\frac{d\nu}{d\pi}\right]$$

for each ν absolutely continuous with respect to π. This shows that $c^2(G) \leq 2\sigma^2(G)$. But we can do better by using the "dual" forms of the transportation inequality and variance transportation inequality as described in Section 1.4.1 (see Propositions 2.3.1 and 2.3.2 for proofs of the equivalence of the dual formulations). In the following proposition we formalize this inequality which was noted in Section 1.4.1. The proof we give here was noted in [8].

Proposition 3.1.1. *The subgaussian constant $\sigma^2(G)$ is at least as large as the spread constant $c^2(G)$.*

Proof. The dual formulation of the subgaussian constant gives us

$$E\left[e^{t(f - Ef)}\right] \leq e^{\sigma^2(G) t^2/2}$$

for every Lipschitz f and real number t. Hence we have:

$$E\left[\frac{e^{t(f - Ef)} - 1}{t^2/2}\right] \leq \frac{e^{\sigma^2 t^2/2} - 1}{t^2/2}$$

for every Lipschitz f and non-zero real number t. If we take the limit as $t \to \infty$, the left hand side of this inequality becomes $\operatorname{Var}(f)$ and the right hand side becomes σ^2. Since the inequality holds for every Lipschitz f, we get $c^2(G) \leq \sigma^2(G)$ using the dual formulation of the spread constant. $\square$

The question of when the subgaussian constant and the spread constant are equal is an interesting one. In Chapter 4 we show that the subgaussian constant and the spread constant are equal on cycles with an even number of vertices and different on cycles with an odd number of vertices, assuming that the associated measures are the uniform probability measures. Lemma 4.1.1 of that chapter gives one general technique for proving that the subgaussian and spread constants are different. Consider the following example found in [8]. It shows that the subgaussian constant need not even be of the same order as the spread constant.

Example 3.1.2 (Two Vertex Path). *Let P_2 be the path on two vertices with vertex set $\{1,2\}$. Suppose the associated measure π is given by $\pi(1) = p$ and $\pi(2) = q$, where $q = 1 - p$. Then $c^2(P_2) = pq$ and $\sigma^2(P_2) = \frac{p-q}{2(\log(p)-\log(q))}$, when d is the graph distance. So $c^2(P_2) \ll \sigma^2(P_2)$ as $p \to 0$ (or as $p \to 1$).*

Finally, note that both the transportation and the variance transportation inequalities depend only on the graph structure of G and are independent of any Markov chain on G.

3.1.3 Modified Log-Sobolev Implies Transportation

Next we show that the modified log-Sobolev inequality (11) implies the transportation inequality (6) with $\sigma^2 = \frac{1}{2\rho_0(L)}$, which is equivalent to the inequality $\sigma^2(G) \leq \frac{1}{2\rho_0(L)}$. The modified log-Sobolev inequality depends on the specific generator L of the Markov chain on G, while the transportation inequality only depends on the underlying graph structure of G. For our proof to work, we must assume:

$$\sum_{y \in V} d^2(x,y)L(x,y) \leq 1, \tag{50}$$

for every $x \in V$. This condition gives a tighter link between the graph structure and the Markov chain on G. In Section 3.2 we look more closely at this constraint through some examples. We formalize the implication in the following proposition, whose proof follows closely a proof by Otto and Villani [30] that the quadratic transportation inequality is implied by the (usual) log-Sobolev inequality in $\mathbb{R}^n$.

Proposition 3.1.3. *Assume that the constraint (50) holds between the distance function d and the Markov generator L. Suppose the modified log-Sobolev inequality:*

$$\rho_0 \operatorname{Ent}(f) \leq \frac{1}{2} \mathcal{E}(f, \log f)$$

holds for every density f with respect to π. Then the transportation inequality:

$$W^2(\nu, \pi) \leq 2 \left(\frac{1}{2\rho_0} \right) D(\nu || \pi)$$

holds for every probability measure ν absolutely continuous with respect to π.

Proof. Let ν be a probability measure absolutely continuous with respect to π. Let ν_t be the Markov chain with generator L and initial distribution ν. Let f_t be the density of ν_t with respect to π. As noted in Section 1.1.2, $\nu_t(x)$ is differentiable as a function of $t \in [0, \infty)$ for each $x \in V$. Let g_t be a solution to Kantorovich's problem with respect to ν_t and π for each $t \in [0, \infty)$. We need the inequality

$$\frac{d^+}{dt} W(\nu_t, \pi) \geq \sum_{x \in V} g_t(x) L f_t(x) \pi(x) = -\mathcal{E}(g_t, f),$$

where g_t is a solution to Kantorovich's problem with respect to ν_t and π. We will show the inequality now, although it is actually satisfied with an equality by an extension of Lemma 2.2.7 to distances other than the graph distance. Let $a \in [0, \infty)$. Then

$$
\begin{aligned}
\frac{d^+}{dt} W(\nu_t, \pi) \bigg|_{t=a} &= \lim_{t \to a} \frac{W(\nu_t, \pi) - W(\nu_a, \pi)}{t - a} \\
&= \lim_{t \to a} \frac{\sum_{x \in V} g_t(x)(\nu_t(x) - \pi(x)) - \sum_{x \in V} g_a(x)(\nu_a(x) - \pi(x))}{t - a} \\
&\geq \lim_{t \to a} \frac{\sum_{x \in V} g_a(x)(\nu_t(x) - \pi(x)) - \sum_{x \in V} g_a(x)(\nu_a(x) - \pi(x))}{t - a} \qquad (51) \\
&= \lim_{t \to a} \frac{\sum_{x \in V} g_a(x)(\nu_t(x) - \nu_a(x))}{t - a} \\
&= \sum_{x \in V} g_a(x) \frac{d^+}{dt} \nu_t(x) \bigg|_{t=a} \\
&= \sum_{x \in V} g_a(x) L f_a(x) \pi(x).
\end{aligned}
$$

The proof will now consist of showing that the derivative of $W(\nu_t, \pi)$ is greater than the derivative of $\sqrt{\frac{1}{\rho_0} D(\nu_t || \pi)}$. As t approaches infinity, the Wasserstein distance between ν_t

54

and π and the relative entropy of ν_t with respect to π both approach zero. Together these facts show that for every t, and in particular at $t = 0$, we get $W(\nu_t, \pi) \le \sqrt{\frac{1}{\rho_0} D(\nu_t || \pi)}$.

We start by proving an inequality that we use below. Here we use reversibility and the constraint (50) on the distance function.

$$\sum_{x \in M} \sum_{y \in M} 2(f_t(y) + f_t(x)) d^2(x, y) L(x, y) \pi(x)$$

$$= 2 \sum_{x \in M} \sum_{y \in M} f_t(y) d^2(x, y) L(x, y) \pi(x)$$

$$+ 2 \sum_{x \in M} \sum_{y \in M} f_t(x) d^2(x, y) L(x, y) \pi(x)$$

$$= 2 \sum_{y \in M} f_t(y) \pi(y) \sum_{x \in M} d^2(y, x) L(y, x)$$

$$+ 2 \sum_{x \in M} f_t(x) \pi(x) \sum_{y \in M} d^2(x, y) L(x, y)$$

$$\le 4$$

Now we can bound the derivative of $W(\nu_t, \pi)$ from below (using reversibility in the first

equality):

$$\frac{d}{dt}^+ W(\nu_t, \pi)$$

$$\geq -\mathcal{E}(g_t, f_t)$$

$$= -\frac{1}{2} \sum_{x \in M} \sum_{y \in M} (g_t(y) - g_t(x))(f_t(y) - f_t(x))L(x,y)\pi(x)$$

$$\geq -\frac{1}{2} \sum_{x \in M} \sum_{y \in M} d(x,y)|f_t(y) - f_t(x)|L(x,y)\pi(x) \tag{52}$$

$$= -\frac{1}{2} \sum_{x \in M} \sum_{y \in M} d(x,y) \left| \sqrt{f_t(y)} - \sqrt{f_t(x)} \right| \left(\sqrt{f_t(y)} + \sqrt{f_t(x)} \right) L(x,y)\pi(x)$$

$$\geq -\frac{1}{2} \left(\sum_{x \in M} \sum_{y \in M} \left(\sqrt{f_t(y)} - \sqrt{f_t(x)} \right)^2 L(x,y)\pi(x) \right)^{1/2} \tag{53}$$

$$\left(\sum_{x \in M} \sum_{y \in M} \left(\sqrt{f_t(y)} + \sqrt{f_t(x)} \right)^2 d^2(x,y)L(x,y)\pi(x) \right)^{1/2}$$

$$\geq -\frac{1}{2} \left(\frac{1}{4} \sum_{x \in M} \sum_{y \in M} (f_t(y) - f_t(x))(\log f_t(y) - \log f_t(x))L(x,y)\pi(x) \right)^{1/2} \tag{54}$$

$$\left(\sum_{x \in M} \sum_{y \in M} 2(f_t(y) + f_t(x))d^2(x,y)L(x,y)\pi(x) \right)^{1/2}$$

$$\geq -\frac{1}{2} \left(\sum_{x \in M} \sum_{y \in M} (f_t(y) - f_t(x))(\log f_t(y) - \log f_t(x))L(x,y)\pi(x) \right)^{1/2} \tag{55}$$

$$= -\frac{1}{\sqrt{2}} \sqrt{\mathcal{E}(f_t, \log f_t)}$$

The inequality in (52) comes from the Lipschitz property of g_t. We use Hölder's inequality in (53). In (54) we use the inequality $\mathcal{E}(e^{f/2}, e^{f/2}) \leq \frac{1}{4}\mathcal{E}(e^f, f)$ observed in [17]. And finally in (55) we use the inequality show in (51).

Next we bound the derivative of $\sqrt{\frac{1}{\rho_0} D(\nu_t||\pi)}$ from above, using the modified log-Sobolev inequality and the derivative of $D(\nu_t, \pi)$ as noted in Section 1.4.2:

$$\frac{d}{dt}\sqrt{\frac{D(\nu_t||\pi)}{\rho_0}} = -\frac{\frac{1}{2}\mathcal{E}(f_t, \log f_t)}{\sqrt{\rho_0 D(\nu_t||\pi)}}$$

$$\leq -\frac{1}{\sqrt{2}} \sqrt{\mathcal{E}(f_t, \log f_t)}$$

$$\leq \frac{d}{dt}^+ W(\nu_t, \pi)$$

56

Let

$$\phi(t) = \sqrt{\frac{D(\nu_t||\pi)}{\rho_0}} - W(\nu_t, \pi)$$

Then $\frac{d}{dt}^+ \phi(t) \leq 0$. Therefore

$$0 = \lim_{t\to\infty} \phi(t) \leq \phi(0) = -W(\nu_0, \pi) + \sqrt{\frac{D(\nu_0||\pi)}{\rho_0}}.$$

And from this we get the result:

$$W^2(\nu_0, \pi) \leq \frac{1}{\rho_0} D(\nu_0||\pi).$$

as desired. □

3.1.4 Poincaré Implies Variance Transportation

Now we show that the Poincaré inequality (10) implies the variance transportation inequality (7) with $c^2 = \frac{1}{2\lambda_1(L)}$, which is equivalent to showing the inequality $c^2(G) \leq \frac{1}{2\lambda_1(L)}$. Here again we have the Poincaré inequality depending on the specific Markov generator, while the variance transportation inequality depends only on the underlying graph structure. So we make an assumption on the distance function that relates the two. We again stipulate that the distance function is an actual metric and further that:

$$\sum_{x,y\in V} d^2(x,y)L(x,y)\pi(x) \leq 1. \tag{56}$$

Note that this assumption is weaker than the assumption (50) needed for Proposition 3.1.3. We formalize the implication in the following proposition and then give two proofs. The first proof is short, well known, and uses the dual form of the variance transportation inequality. The second proof follows the same lines as the proof of the previous proposition, but is easier. See Chapter 6 for a discussion of extending this proof to a family of implications.

Proposition 3.1.4. *Suppose the Poincaré inequality:*

$$\lambda_1 \operatorname{Var}(f) \leq \mathcal{E}(f, f)$$

holds for every density f with respect to π. Then the variance transportation inequality:

$$W^2(\nu, \pi) \leq \left(\frac{1}{2\lambda_1}\right) \operatorname{Var}\left(\frac{d\nu}{d\pi}\right)$$

holds for every probability measure ν absolutely continuous with respect to π.

First Proof. Suppose f is a Lipschitz function. Then

$$\begin{aligned}
\mathcal{E}(f,f) &= \frac{1}{2} \sum_{x,y \in V} (f(x) - f(y))^2 L(x,y)\pi(x) \\
&\leq \frac{1}{2} \sum_{x,y \in V} d^2(x,y) L(x,y)\pi(x) \\
&\leq \frac{1}{2}
\end{aligned}$$

by the condition (56). So

$$\lambda_1(L) = \inf_{f:\mathrm{Var}_\pi f \neq 0} \frac{\mathcal{E}(f,f)}{\mathrm{Var}\, f} \leq \inf_{\substack{f:\mathrm{Var} f \neq 0 \\ f \in Lip(G)}} \frac{\mathcal{E}(f,f)}{\mathrm{Var}\, f} \leq \inf_{\substack{f:\mathrm{Var} f \neq 0 \\ f \in Lip(G)}} \frac{1}{2\,\mathrm{Var}\, f} = \frac{1}{2c^2(G)}. \qquad \square$$

Second Proof. Here we cut to the core of the proof, where it is different from the proof of the previous proposition.

$$\begin{aligned}
\frac{d}{dt}^{+} W(\nu_t, \pi) \\
\geq\ & -\mathcal{E}(g_t, f_t) \\
=\ & -\frac{1}{2} \sum_{x \in M} \sum_{y \in M} (g_t(y) - g_t(x))(f_t(y) - f_t(x)) L(x,y)\pi(x) \\
\geq\ & -\frac{1}{2} \sum_{x \in M} \sum_{y \in M} d(x,y)|f_t(y) - f_t(x)| L(x,y)\pi(x) \\
\geq\ & -\frac{1}{2} \left(\sum_{x \in M} \sum_{y \in M} |f_t(y) - f_t(x)|^2 L(x,y)\pi(x) \right)^{1/2} \\
& \left(\sum_{x \in M} \sum_{y \in M} d^2(x,y) L(x,y)\pi(x) \right)^{1/2} \\
=\ & -\frac{1}{\sqrt{2}} \sqrt{\mathcal{E}(f_t, f_t)}
\end{aligned}$$

Then recalling from Section 1.4.2 that $\frac{d}{dt}\mathrm{Var}(f_t) = -2\mathcal{E}(f_t, f_t)$ we have:

$$\begin{aligned}
\frac{d}{dt} \sqrt{\frac{\mathrm{Var}(f_t)}{2\lambda_1}} &= -\frac{1}{\sqrt{2\lambda_1}} \frac{\mathcal{E}(f_t, f_t)}{\sqrt{\mathrm{Var}(f_t)}} \\
&\leq -\frac{1}{\sqrt{2}} \sqrt{\mathcal{E}(f_t, f_t)} \\
&\leq \frac{d}{dt}^{+} W(\nu_t, \pi)
\end{aligned}$$

Since $\lim_{t \to \infty} W(\nu_t, \pi) = \lim_{t \to \infty} \mathrm{Var}(f_t) = 0$, without repeating the mechanics of the proof

of the previous proposition we get:

$$W(\nu_0, \pi) \leq \sqrt{\frac{\mathrm{Var}(f_0)}{2\lambda_1}}$$

which is what we wanted. □

3.1.5 Transportation Does Not Imply Poincaré

In $\mathbb{R}^n$, the quadratic cost transportation inequality implies the Poincaré inequality [30, 9], which leads us to consider the possibility that the transportation inequality could imply the Poincaré inequality in our discrete setting. We answer this question with a qualified no. If d is taken to be the graph distance, then the transportation inequality does not imply the Poincaré inequality. By this we mean that for every positive constant ϵ we can find a graph G with associated measure π and graph distance d and the generator L of a Markov chain on G so that $\lambda_1(L) < \epsilon \frac{1}{\sigma^2(G)}$. We consider a Markov chain on a two vertex path as described later in Example 3.2.1. For that chain we have $\lambda_1(L) = s+t$ while $\frac{1}{2\sigma^2(P_2)} = \frac{1}{4}$. So $\lambda_1(L) \ll \frac{1}{2\sigma^2(P_2)}$ as $s, t \to 0$. The dumbbell graph of Example 3.2.4 provides a less artificial example, as the transition rates are not arbitrarily sent to zero.

3.1.6 Poincaré Does not Imply Transportation

We know that the modified log-Sobolev inequality implies the transportation inequality, so it is reasonable to ask if the weaker Poincaré inequality also implies the transportation inequality. As in the previous section, we answer in the negative when d is the graph distance. To do this we find a natural family of Markov generators $\{L_i\}_{i=1}^\infty$ on a family of graphs $\{G_i\}_{i=1}^\infty$ for which $\lambda_1(L_i) \gg \frac{1}{2\sigma^2(G_i)}$ as $i \to \infty$. Since $\frac{1}{2c^2(G_i)} \geq \lambda_1(L_i)$ for each i, this also gives us a family of graphs for which $c^2(G_i) \ll \sigma^2(G_i)$ as $i \to \infty$.

We will need the following lower bound on $\sigma^2(G)$, similar to the lower bound on $\frac{1}{\lambda_1(L)}$ by Alon and Milman [4].

Lemma 3.1.5. *Suppose $G = (V, E)$ is a graph with associated measure π and distance function d. Suppose $\pi^* = \min_{x \in V} \pi(x)$ is strictly positive. Then $\sigma^2(G) \geq \frac{D^2}{32 \log \frac{1}{\pi^*}}$.*

Proof. Let $x, y \in V$ with $d(x, y) = D$. Let

$$A_x = \{v \in V : d(x, v) \leq d(v, y)\} \quad \text{and} \quad A_y = \{v \in V : d(x, v) \geq d(v, y)\}$$

Then $\pi(A_x) \geq \frac{1}{2}$ or $\pi(A_y) \geq \frac{1}{2}$. Without loss of generality suppose $\pi(A_x) \geq \frac{1}{2}$. Let $v^* \in A_x$ with the property that $d(y, A_x) = d(y, v^*)$. Then

$$D = d(x, y) \leq d(x, v^*) + d(v^*, y) \leq 2d(v^*, y) = 2d(y, A_x)$$

giving $d(y, A_x) \geq D/2$. So $\{v \in V : d(v, A_x) \geq D/2\}$ is not empty. Then by (8) we get:

$$\pi^* \leq \pi(\{v \in V : d(v, A_x) \geq D/2\}) \leq \alpha(D/2) \leq e^{-\frac{\left(\frac{D}{2}\right)^2}{8\sigma^2}}.$$

Solving for σ^2 gives the result. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

When π is the uniform measure, this lemma gives the bound $\sigma^2(G) \geq \frac{D^2}{32 \log |V|}$. Suppose $\{G_i\}_{i=1}^{\infty}$ is a family of graphs where G_i has n_i vertices. Let d_i be the graph distance on G_i and associate the uniform probability measure π_i with G_i. Let L_i be the generator of a Markov chain on G_i and let D_i denote the diameter of graph G_i. If there exists $\epsilon > 0$ so that $\lambda_1(L_i) \geq \epsilon$ for each i and if $D_i \gg \sqrt{\log n_i}$ as $i \to \infty$, then we have $\lambda_1(L_i) \gg \frac{1}{2\sigma^2(G_i)}$. A natural example is when $\{G_i\}_{i=1}^{\infty}$ is a family of bounded degree expander graphs with Markov generators L_i. By this we mean there exist positive constants ϵ and k such that the maximum degree of a vertex in G_i is bounded from above by k for each i and $\lambda_1(L_i) \geq \epsilon$ for each i. Then $D_i \geq K \log n_i$ for some constant K, since the family has bounded degree.

3.2   Different Distance Functions

Suppose we are interested in a specific Markov chain on a graph, and we would like to have bounds on the modified log-Sobolev constant or the spectral gap of the chain. As we saw in the previous section, we may use the spread constant to bound the spectral gap from above and the subgaussian constant to bound the modified log-Sobolev constant from above. In this section we look at how good these bounds are, and how they can be improved by using distances other than the graph distance.

We start out by showing that when d is the graph distance, then the implications in Propositions 3.1.3 and 3.1.4 hold under very general conditions. First suppose that the Markov chain generated by L is the continuization of a discrete time chain as defined in

Section 1.1.2. Then L satisfies (2), which we repeat here:

$$\sum_{\substack{y \in V \\ y \neq x}} L(x, y) \leq 1,$$

for every $x \in V$. Suppose further that d is the graph distance, so that $d(x, y) = 1$ whenever $L(x, y) > 0$ and $x \neq y$. Then under these assumptions the condition (50) sufficient for Proposition 3.1.3 to hold and the weaker condition (56) sufficient for Proposition 3.1.4 to hold are both satisfied. Hence for any discrete time Markov chain on G with transition matrix P, if we let $L = P - I$, then $\rho_0(L) \leq \frac{1}{2\sigma^2(G)}$ and $\lambda_1(L) \leq \frac{1}{2c^2(G)}$, where $\sigma^2(G)$ and $c^2(G)$ are calculated using the graph distance.

If we are only concerned with the spectral gap, we can require less. Instead of necessitating that the chain be the continuization of a discrete time chain, we may simply assume that L satisfies the normalization condition given in (1), and repeated here:

$$\sum_{x \in V} \sum_{\substack{y \in V \\ y \neq x}} L(x, y) \pi(x) \leq 1.$$

Then condition (56) sufficient for Proposition 3.1.4 is satisfied when d is the graph distance. So for any continuous time Markov chain on G whose generator satisfies the normalization condition we have $\lambda_1(L) \leq \frac{1}{2c^2(G)}$, where $c^2(G)$ is calculated using the graph distance.

We begin our examples with the two vertex path, on which we can put the simplest possible non-trivial Markov chain. We show that under the graph distance, the inequalities $\rho_0(L) \leq \frac{1}{2\sigma^2(G)}$ and $\lambda_1(L) \leq \frac{1}{2c^2(G)}$ are tight only for very specific Markov chains, while if we allow any distance function satisfying conditions (50) or (56), then the inequalities can be made tight for more general chains.

Example 3.2.1 (Markov Chain on Two Vertex Path). *Let P_2 be the two point path of Example 3.1.2 with associated measure π defined in that example. Define the Markov generator L by:*

$$L = \begin{bmatrix} -s & s \\ t & -t \end{bmatrix}$$

The stationary measure $\tilde{\pi}$ of the Markov chain generated by L is given by $\tilde{\pi}(1) = \frac{t}{s+t}$ and $\tilde{\pi}(2) = \frac{s}{s+t}$. In order for L to be a Markov chain on P_2 (as defined in Section 1.1.2), it

must respect the measure π associated with the graph. This means that $\tilde{\pi} = \pi$ (i.e. $p = \frac{t}{s+t}$ and $q = \frac{s}{s+t}$).

Bounds on the spectral gap and the modified log-Sobolev constant for this chain are given in [10] as:

$$\frac{s+t}{2} + \sqrt{st} \le \rho_0(L) \le \lambda_1(L) = s + t$$

Note that there is equality when $s = t$. The spread constant, being the maximum variance of a Lipschitz function on P_2, is:

$$c^2(P_2) = pq\,d(1,2)^2 = \frac{st}{(s+t)^2}d(1,2)^2$$

The subgaussian constant is calculated in [8] as:

$$\sigma^2(P_2) = \frac{p-q}{2(\log p - \log q)}d(1,2)^2 = \frac{1}{(s+t)}\frac{s-t}{2(\log s - \log t)}d(1,2)^2$$

To understand this scenario, we first simplify it by assuming that d is the graph distance and that $s = t$. Then we have $\rho_0(L) = \lambda_1(L) = 2s$, while $\sigma^2(P_2) = c^2(P_2) = \frac{1}{4}$ (the value of $\sigma^2(P_2)$ is the limit as s approaches t in the above formula). When $s = 1$, the inequalities $\rho_0(L) \le \frac{1}{2\sigma^2(P_2)}$ and $\lambda_1(L) \le \frac{1}{2c^2(P_2)}$ are tight. As s approaches zero, the bounds given by $\sigma^2(P_2)$ and $c^2(P_2)$ become meaningless, and for $s > 1$, the inequalities are false. If we continue to assume that $s = t$, but do not arbitrarily choose the graph distance, (50) and (56) are both satisfied exactly when $d(1,2) \le \frac{1}{\sqrt{s}}$. Setting $d(1,2) = \frac{1}{\sqrt{s}}$, we get $\sigma^2(P_2) = c^2(P_2) = \frac{1}{4s}$. Then $\rho_0(L) = \lambda_1(L) = \frac{1}{2\sigma^2(P_2)} = \frac{1}{2c^2(P_2)}$, so the implication is tight for all values of s.

Now we go back to the general setting. We start by looking at the Poincaré to variance transportation implication. Condition (56) is satisfied when $d(1,2) \le \sqrt{\frac{s+t}{2st}}$. Setting $d = \sqrt{\frac{s+t}{2st}}$, we get $c^2(P_2) = \frac{1}{2\lambda_1(L)} = s + t$, so the implication is tight for all values of s and t. The situation for the modified log-Sobolev to transportation implication is not so nice. Condition (50) is satisfied when $d(1,2) \le \min(\frac{1}{\sqrt{s}}, \frac{1}{\sqrt{t}})$. To be concrete we will assume that $t < s$. Then when we set $d(1,2) = \frac{1}{\sqrt{s}}$, we get $\rho_0(L) \le \frac{1}{2\sigma^2(P_2)} = (s+t)\frac{(\log s - \log t)}{s-t}s$. This does give a better bound on $\rho_0(L)$ when $s < 1$ than we get from $\sigma^2(P_2)$ calculated with the graph distance, and it is guaranteed to be an upper bound on $\rho_0(L)$ no matter

what s and t are. On the other hand, from the spread constant we get the bound $\rho_0(L) \leq \lambda_1(L) \leq \frac{1}{2c^2(P_2)} = s + t$, which is a better bound on $\rho_0(L)$ since $\frac{(\log s - \log t)}{s-t} s \geq 1$. One may conjecture that only the weaker condition (56) is necessary for the modified log-Sobolev inequality to imply the transportation inequality. But if we calculate $\sigma^2(P_2)$ with the distance $d = \sqrt{\frac{s+t}{2st}}$ which satisfies (56), we get $\sigma^2(P_2) = \frac{(s+t)^2}{2st} \frac{\log s - \log t}{s-t}$. For $s = 1$ and $t > 1$ we have $\frac{1}{2\sigma^2(P_2)} \ll \frac{s+t}{2} + \sqrt{st} \leq \rho_0(L)$ as $t \to \infty$, so the implication does not always hold. But we make a modified conjecture.

Conjecture 3.2.2. *If L satisfies (2), so that the chain generated by L is the continuization of a discrete time chain, and if further d and L satisfy (56), then the modified log-Sobolev inequality implies the transportation inequality with $\sigma^2 = \frac{1}{2\rho_0(L)}$.*

Next we look at Markov chains on the complete graph. The situation here is somewhat similar to the two vertex graph, but more difficult.

Example 3.2.3 (Markov Chain on the Complete Graph). *Let K_n be the complete graph on n vertices with associated probability measure π. Let L be the generator of a Markov chain on K_n with $L(x, y) = \pi(y)$ for each $x, y \in V$ with $x \neq y$. Then $L(x, x) = 1 - \pi(x)$. Note that the chain satisfies (2), and hence is the continuization of a discrete time chain.*

Again, the bounds on ρ_0 and λ_1 are given in [10] while the values of σ^2 and c^2 are given in [8]. Let $\pi^* = \min_{x \in V} \pi(x)$. Then

$$\frac{1}{2} + \sqrt{\pi^*(1 - \pi^*)} \leq \rho_0(L) \leq \lambda_1(L) = 1.$$

Let $p^* = \min\{\pi(A) : \pi(A) \geq 1/2\}$ and $q^* = 1 - p$. Then calculating the subgaussian and spread constants using the graph distance we have:

$$c^2(K_n) = p^* q^* \quad \text{and} \quad \sigma^2(K_n) = \frac{p^* - q^*}{2(\log p^* - \log q^*)}.$$

where $\sigma^2(K_n) = 1/4$ when $p^* = q^*$.

It $p^* = q^*$, as when n is even and π is the uniform measure, then $\rho_0(L) \leq \lambda_1(L) = 1 < 2 = \frac{1}{2c^2(K_n)} = \frac{1}{2\sigma^2(K_n)}$. When $p^* \neq q^*$ we actually get the inequality $\rho_0(L) \leq \lambda_1(L) = 1 < 2 < \frac{1}{2\sigma^2(K_n)} < \frac{1}{2c^2(K_n)}$. So $\sigma^2(K_n)$ gives a better bound on $\rho_0(L)$ than $c^2(K_n)$ (as it always

does when they are calculated using the graph distance). But neither seems to accurately capture what is going on.

We may try to improve the bound on $\lambda_1(L)$ using $c^2(K_n)$ calculated using another distance function. The condition (56) allows us to bound the variance of a Lipschitz function by:

$$
\begin{aligned}
\text{Var}(f) &= \frac{1}{2} \sum_{x,y \in V} (f(y) - f(x))^2 \pi(x)\pi(y) \\
&\leq \frac{1}{2} \sum_{x,y \in V} d(x,y)^2 \pi(x)\pi(y) \qquad\qquad (57) \\
&= \frac{1}{2} \sum_{x,y \in V} d(x,y)^2 L(x,y)\pi(x) \\
&\leq \frac{1}{2}. \qquad\qquad (58)
\end{aligned}
$$

If we could find a distance function d to make (58) tight and a Lipschitz function f to make (57) tight, then we would have $c^2(K_n) = 1/2$ and hence $\lambda_1(L) = \frac{1}{2c^2(K_n)}$. Although finding a distance function which makes (58) tight is easy to do, given a particular d, making (57) tight is not always possible. Consider again the case where n is even and π is the uniform measure. Then conditions (56) and even the stricter (50) are satisfied when $d(x,y) = \sqrt{\frac{n}{n-1}}$ for each $x \neq y$. With this distance $\sigma^2(K_n) = c^2(K_n) = \frac{n}{n-1}\frac{1}{4}$, which is an improvement over the graph distance, but not by much. It is not clear if we can do better than this. When π is not uniform, the problem is even more difficult.

Example 3.2.4 (Markov Chain on the Dumbbell Graph). *Consider the dumbbell graph $B = (V, E)$ with an even number of vertices n. B consists of two complete graphs on $n/2$ vertices connected by one edge $\{v, v'\}$. We are interested in the Markov generator L on B in which the transition rate r is the same between every two adjacent vertices. Then the chain's stationary distribution is the uniform distribution.*

Here we use the bounds:

$$
\rho_0(L) \leq \lambda_1(L) \leq \frac{\mathcal{E}(f,f)}{\text{Var}(f)} \quad \text{and} \quad \frac{2}{D^2} \leq \frac{1}{2\sigma^2(B)} \leq \frac{1}{2c^2(B)} \leq \frac{2}{d(v,v')^2}
$$

where D is the diameter of G and f is any function on V. If we let f be the indicator

64

function of one side of the dumbbell, and if d is the graph distance we have:

$$\rho_0(L) \leq \lambda_1(L) \leq \frac{r}{n} \quad \text{and} \quad \frac{2}{9} \leq \frac{1}{2\sigma^2(B)} \leq \frac{1}{2c^2(B)} \leq 2,$$

where we recall that r is the transition rate between adjacent vertices. Both of the conditions (2) and (1) imply that $r = O(\frac{1}{n})$. So we have $\rho_0(L) \leq \lambda_1(L) = O(\frac{1}{n^2})$ while $\frac{1}{2\sigma^2(B)}$ and $\frac{1}{2c^2(B)}$ are bounded below by a constant. So $\rho_0(L) \ll \frac{1}{2\sigma^2(B)}$ and $\lambda_1(L) \ll \frac{1}{2c^2(B)}$, as $n \to \infty$.

If we use condition (56) which guarantees that $\lambda_1(L) \leq \frac{1}{2c^2(B)}$ we are able to take $d(v, v') = \sqrt{\frac{n}{2r}}$ (and $d(x, y) = 0$ for all other adjacent vertices x and y) which gives us $\frac{1}{2c^2(B)} = \frac{4r}{n}$ which is at least the same order as our upper bound on $\lambda_1(L)$.

Using condition (56) which guarantees that $\rho_0(L) \leq \frac{1}{2\sigma^2(B)}$ we cannot get $d(v, v')$ larger than $\sqrt{\frac{1}{r}}$, so finding a good upper bound on $\rho_0(L)$ using $\sigma^2(B)$ is more difficult that simply using $c^2(B)$.

These examples seem to suggest that unless we can prove Conjecture 3.2.2, if we want an upper bound on $\rho_0(L)$, we are better off using $\lambda_1(L)$ as an upper bound and using $\frac{1}{2c^2(G)}$ to find the upper bound on λ_1. But as we showed in Section 3.1.6, families of expander graphs provide an example where $\frac{1}{2\sigma^2(G_n)} \ll \lambda_1(L_n)$ as $n \to \infty$, even when $\sigma^2(G_n)$ is calculated using the graph distance. Since $\frac{1}{2c^2(G_n)}$ is bounded below by $\lambda_1(L_n)$, in this example $\frac{1}{2\sigma^2(G_n)}$ provides a much better bound on $\rho_0(L_n)$ than we could get using $c^2(G_n)$.

3.3 Fastest Mixing Markov Process Problem

In this section we look at the opposite of the problem in the previous section. Now we begin with a graph $G = (V, E)$ with associated measure π and distance function d for which we would like to know bounds on the subgaussian or spread constants. As we saw earlier, whenever L is the generator of a Markov chain on G, if (2) is satisfied, then $\rho_0 \leq \frac{1}{2\sigma^2}$ and $\lambda_1 \leq \frac{1}{2c^2}$, where σ^2 and c^2 are calculated using the graph distance. If only (1) is satisfied, then $\lambda_1 \leq \frac{1}{2c^2}$, where c^2 is calculated using the graph distance. The values of ρ_0 and λ_1 control the speed at which a Markov chain decays to its stationary measure. This leads us to the problem of finding the fastest mixing Markov chain on G in the sense of finding the one that maximize ρ_0 or λ_1.

Recently, Sun, Boyd, Xiao, and Diaconis [35] examined the problem of finding the fastest continuous time Markov chain on a graph in the sense of maximizing λ_1, following work by the last three authors [12] on finding fastest mixing discrete time Markov chains. For the continuous time chain, they show that finding the generator of a fastest mixing Markov chain on a graph is equivalent to a semi-definite program, whose dual is equivalent to the maximum variance unfolding problem. The maximum variance unfolding problem is given as:

$$
\begin{array}{ll}
\text{maximize:} & \sum_{x \in V} \|F(x)\|^2 \pi(x) \\
\text{subject to:} & F(x) \in \mathbb{R}^{|V|} \qquad\qquad x \in V \\
& \|F(x) - F(y)\| \le d(x, y), \quad \{x, y\} \in E \\
& \sum_{x \in V} F(x) \pi(x) = 0,
\end{array}
\tag{59}
$$

where $\|v\| = \sqrt{v \cdot v}$ for $v \in \mathbb{R}^{|V|}$. Note that this is simply the $|V|$-dimensional relaxation of the spread constant.

Considering Markov generators that satisfy the normalization condition 1, we denote the spectral gap of the fastest Markov chain on G (in the sense of having the largest spectral gap) as $\lambda_1^*(G)$. This means that $\lambda_1^*(G)$ is also the value of the maximum variance unfolding problem. Now we already have $2c^2(G) \le \frac{1}{\lambda_1^*}$, and Assaf Naor showed us that using the Johnson-Lindenstrauss Lemma [23] on the maximum variance unfolding problem we get that the inequality is tight up to a factor of $\log |V|$.

Proposition 3.3.1. $\frac{1}{\lambda_1^*(G)} = O(c^2(G) \log |V|)$.

Proof. The Johnson-Lindenstrauss Lemma [23] guarantees the existence of a map $h : \mathbb{R}^{|V|} \to \mathbb{R}^d$ with $d = O(\log |V|)$, with

$$
\frac{1}{2} \|F(x) - F(y)\| \le \|h(F(x)) - h(F(y))\| \le \|F(x) - F(y)\| \text{ for all } x, y \in V.
$$

This map maintains the constraint that $\|h(F(x)) - h(F(y))\| \le 1$, for $\{x, y\} \in E$, and returns vectors in $\mathbb{R}^d$, with no worse than a factor of 4 loss in the optimal variance. For

$$
\frac{1}{4} \sum_{x,y \in V} \|F(x) - F(y)\|^2 \pi(x) \pi(y) \le \sum_{x,y \in V} \|h(F(x)) - h(F(y))\|^2 \pi(x) \pi(y).
$$

The constraint about the mean of the configuration being zero should not matter, since a constant can be subtracted from any solution without affecting $\|F(x) - F(y)\|$. Finally to complete the proposition, observe that

$$
\begin{aligned}
\frac{1}{4d}\frac{1}{\lambda_1^*} = \frac{1}{8d}\sum_{x,y\in V}\|F(x) - F(y)\|^2\pi_i\pi_j \\
\leq \frac{1}{2d}\sum_{x,y\in V}\sum_{k=1}^{d}|h(F(x))(k) - h(F(y))(k)|^2\pi(x)\pi(y) \\
\leq \max_k\frac{1}{2}\sum_{x,y\in V}|h(F(x))(k) - h(F(y))(k)|^2\pi(x)\pi(y) \\
\leq c^2(G). \qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \square
\end{aligned}
$$

This bound on $\frac{1}{\lambda_1^*}$ provides an interesting existence guarantee for fast mixing Markov chains on a graph. Consider a graph G on n vertices with the uniform probability measure and diameter bounded above by a polynomial in the log of n. The total variation mixing time of a Markov chain with generator L on G is bounded above by $k\frac{1}{\lambda_1(L)}\log n$, for some constant k. This bound on the total variation mixing time, together with the previous result and the fact the $c^2(G)$ is bounded above by a fraction of the diameter squared, shows that the mixing time of a fastest chain on G is polynomial in $\log n$.

Further extensions by A. Naor of this work appear in [29], including the following results. For planar graphs, the spread constant can be estimated efficiently up to a constant factor. It can be estimated efficiently for graphs with a doubling constant λ up to a factor of $\log\lambda\log\log\lambda$. As the previous lemma shows, in general the spread constant can be estimated efficiently up to a factor of $\log|V|$, since semidefinite programs have efficient algorithms. Further, there exist graphs showing that this $\log|V|$ bound is tight up to a factor of $\log\log|V|$. Finally, in general there is no efficient algorithm for calculating the spread constant of a graph up to a small constant factor.

CHAPTER IV

CONCENTRATION ON THE DISCRETE TORUS

In this chapter, we show that $\sigma^2(G) = c^2(G)$ if there is no probability measure ν other than π for which $W^2(\nu, \pi) = 2\sigma^2 D(\nu||\pi)$. We then use this fact to find the subgaussian constant of the even cycle exactly by showing that $\sigma^2(C_{2k}) = c^2(C_{2k})$, for every $k \geq 1$. For the odd cycle we use the transportation inequality definition of the subgaussian constant to show that $c^2(C_{2k+1}) < \sigma^2(C_{2k+1}) = c^2(C_{2k+1})(1 + o(1))$, where $o(1)$ disappears as the number of vertices goes to infinity.

The above has superficial similarity to a recent result of Chen and Sheu [15], who found the exact value of the log-Sobolev constant ρ of the even cycle. They prove that for the even cycle, ρ equals the spectral gap λ_1, using the fact that the inequality $\rho \leq \lambda_1$ is actually an equality if there is no function f for which the log-Sobolev inequality is satisfied with equality.

Prior to this work, the exact values of the subgaussian constant were computed for a few graphs in [8], including the 2-point space with arbitrary probability measure, the completely connected graph and the path of arbitrary length with uniform probability measure. They reduce the problem of finding $\sigma^2(C_3)$ to finding the subgaussian constant on a (nonuniformly) weighted path of length two. And the value of $\sigma^2(C_4)$ is known, as C_4 is the product of two copies of a 2-vertex path. But computing the subgaussian constant of cycles of with more than four vertices remained open. One goal of this chapter is to find the asymptotically correct value of $\sigma^2(C_n)$, irrespective of the parity of n.

We are further motivated by the work of Riordan [31], who in 1998 solved the isoperimetric problem on the discrete torus consisting of a product of even cycles, by finding an ordering on the torus for which the initial segments are sets of smallest surface area. He notes that his proof cannot extend to products of cycles which include ones of odd length because the extremal sets are not necessarily nested (as in the cube of the 3-cycle, for

example). Finding tight bounds on the subgaussian constant of a cycle, and using tensoring, gives a concentration result for the discrete torus without needing to go through the isoperimetric problem.

The results for the cycle are proved in Section 4.3, while much of the work for the last proposition of Section 4.3 is contained in a collection of lemmas in Section 4.2. We begin with section 4.1 which contains general lemmas concerning the transportation inequality, with special emphasis on the state in which the inequality obtains equality.

4.1 Technical Lemmas about the Transportation Inequality

Let $G = (V, E)$ be a graph with associated measure π and distance function d. The following lemma is useful for proving that the subgaussian constant and the spread constant are different. Recall that $P(G)$ denotes the set of probability measures on V and $\| \cdot \|$, denotes the l_1 norm.

Lemma 4.1.1. *Let f be a Lipschitz function with $\mathrm{E}[f] = 0$ and $\mathrm{Var}[f] = c^2(G)$. If $\mathrm{E}[f^3] \neq 0$ then $\sigma^2(G) > c^2(G)$.*

Proof. Define

$$F(\nu) = \frac{D(\nu \| \pi)}{\left(\sum_{x \in V} f(x)(\nu(x) - \pi(x)) \right)^2}$$

on the subset D of $P(G)$ for which the denominator is not zero. Then

$$\inf_{\nu \in D} F(\nu) \geq \inf_{\nu \in P(G) \setminus \{\pi\}} \frac{D(\nu \| \pi)}{W^2(\nu, \pi)} = \frac{1}{2\sigma^2(G)}.$$

For positive ϵ small enough that $|\epsilon f(x)| < 1$ for every $x \in V$, define the measure ν_ϵ by

$d\nu_\epsilon = (1 + \epsilon f)d\pi$. Consider the following limit:

$$
\begin{aligned}
\lim_{\epsilon \to 0} F(\nu_\epsilon) &= \lim_{\epsilon \to 0} \frac{\sum_{x \in V}(1 + \epsilon f(x)) \log(1 + \epsilon f(x))\pi(x)}{\left(\sum_{x \in V} f(x)[(1 + \epsilon f(x))\pi(x) - \pi(x)]\right)^2} \\
&= \lim_{\epsilon \to 0} \frac{\sum_{x \in V}(1 + \epsilon f(x)) \log(1 + \epsilon f(x))\pi(x)}{\epsilon^2 \left(\sum_{x \in V} f(x)^2 \pi(x)\right)^2} \\
&= \lim_{\epsilon \to 0} \frac{\sum_{x \in V}(1 + \epsilon f(x)) \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} \epsilon^k f(x)^k \pi(x)}{\epsilon^2 \mathrm{E}[f^2]^2} \\
&= \frac{\frac{1}{2} E[f^2] + \lim_{\epsilon \to 0} \sum_{x \in V} \sum_{k=2}^{\infty} \left(\frac{(-1)^{k+2}}{k+1} + \frac{(-1)^{k+1}}{k}\right) \epsilon^{k-1} f(x)^{k+1} \pi(x)}{\mathrm{E}[f^2]^2} \\
&= \frac{1}{2\mathrm{E}[f^2]} \\
&= \frac{1}{2c^2(G)}.
\end{aligned}
$$

Let I be an open interval around 0 small enough so that $|\epsilon f(x)| < 1$ for every $x \in V$ and $\epsilon \in I$. Define $H : I \to \mathbb{R}$ by

$$
H(\epsilon) = \begin{cases} F(\nu_\epsilon), & \epsilon \neq 0 \\ \frac{1}{2c^2(G)}, & \epsilon = 0 \end{cases}
$$

H is continuous at 0 by the previous limit. Let us calculate the derivative of H at 0.

$$
\begin{aligned}
\frac{d}{d\epsilon} H(\epsilon)|_{\epsilon=0} &= \lim_{t \to 0} \frac{H(t) - H(0)}{t} \\
&= \lim_{t \to 0} \frac{F(\nu_t) - \frac{1}{2\mathrm{E}[f^2]}}{t} \\
&= \lim_{t \to 0} \frac{\frac{\frac{1}{2}E[f^2] + \sum_{x \in V} \sum_{k=2}^{\infty} \left(\frac{(-1)^{k+2}}{k+1} + \frac{(-1)^{k+1}}{k}\right) t^{k-1} f(x)^{k+1} \pi(x)}{\mathrm{E}[f^2]^2} - \frac{1}{2\mathrm{E}[f^2]}}{t} \\
&= \frac{-\frac{1}{6} \mathrm{E}[f^3] + \lim_{t \to 0} \sum_{x \in V} \sum_{k=3}^{\infty} \left(\frac{(-1)^{k+2}}{k+1} + \frac{(-1)^{k+1}}{k}\right) t^{k-2} f(x)^{k+1} \pi(x)}{\mathrm{E}[f^2]^2} \\
&= \frac{-\frac{1}{6} \mathrm{E}[f^3]}{\mathrm{E}[f^2]^2}.
\end{aligned}
$$

Now suppose $\mathrm{E}[f^3] \neq 0$. Then $\frac{d}{d\epsilon} H(\epsilon)|_{\epsilon=0} \neq 0$, which implies there exists $\epsilon \neq 0$ with $H(\epsilon) < H(0)$. This means there exists $\nu \in P(G)$ with $\nu \neq \pi$ and $F(\nu) < \frac{1}{2c^2(G)}$. Hence $\sigma^2(G) > c^2(G)$. $\qquad\square$

The next key lemma gives a sufficient condition for obtaining equality in the transportation inequality.

Lemma 4.1.2. *If $\sigma^2(G) \neq c^2(G)$ then there exists $\nu \in P(G)$ with $\nu \neq \pi$ and $W^2(\nu, \pi) = 2\sigma^2(G)D(\nu\|\pi)$.*

Proof. Define $F : P(G) \setminus \{\pi\} \to \mathbb{R}$ by

$$F(\nu) = \frac{D(\nu\|\pi)}{W^2(\nu, \pi)},$$

so that

$$\frac{1}{2\sigma^2(G)} = \inf_{\nu \in P(G) \setminus \{\pi\}} F(\nu).$$

To prove the lemma we must show that the infimum is attained if $\sigma^2(G) > c^2(G)$.

As mentioned in Section 1.4, $D(\nu\|\pi)$ is a continuous function of $\nu \in P(G)$. Lemma 2.2.3 shows that $W(\nu, \pi)$ is a continuous function of ν. Since $W(\nu, \pi) = 0$ if and only if $\nu = \pi$, F is continuous on $P(G) \setminus \{\pi\}$.

At this point, if $P(G) \setminus \{\pi\}$ were compact, we would be done. We will show that if ν is near π then $F(\nu)$ is too large to be relevant to the infimum. Since $\sigma^2(G) \neq c^2(G)$, there exists $\epsilon > 0$ such that $\sigma^2(G) > (1 + \epsilon)c^2(G)$. Let $m = \min\{\pi(x) : x \in V \text{ and } \pi(x) \neq 0\}$. Then let K and δ_1 be positive real numbers with

$$\frac{1}{2} - 3\delta_1 \geq K = \frac{1}{2(1 + \epsilon)}.$$

Next let $\delta_2 > 0$ small enough so that $m - \delta_2 > 0$ and

$$\frac{\delta_2}{m - \delta_2} \leq \delta_1.$$

Let $\nu \in P(G) \setminus \{\pi\}$ with $\|\nu - \pi\| \leq \delta_2$. Let $a(x) = 1 - \nu(x)/\pi(x)$ for $x \in V$. Then $\|a\| \leq \frac{1}{m}\|\nu - \pi\| \leq \frac{\delta_2}{m}$. Let f be a solution to Kantorovich's problem respect to ν and π

with $\mathrm{E}[f] = 0$. Then

$$
\begin{aligned}
F(\nu) &= \frac{D(\nu\|\pi)}{W^2(\nu,\pi)} \\[2mm]
&= \frac{\sum_{x\in V} \frac{\nu(x)}{\pi(x)} \log\left(\frac{\nu(x)}{\pi(x)}\right)\pi(x)}{\left(\sum_{x\in V} f(x)(\nu(x)-\pi(x))\right)^2} \\[2mm]
&= \frac{\sum_{x\in V}[1-(1-\frac{\nu(x)}{\pi(x)})]\log\left([1-(1-\frac{\nu(x)}{\pi(x)})]\right)\pi(x)}{\left(\sum_{x\in V} f(x)\pi(x)(\frac{\nu(x)}{\pi(x)}-1)\right)^2} \\[2mm]
&= \frac{\sum_{x\in V}(1-a(x))\log\left(1-a(x)\right)\pi(x)}{\left(\sum_{x\in V} f(x)a(x)\pi(x)\right)^2}
\end{aligned}
$$

For each $x \in V$ we have $|a(x)| \le \|a\| \le \frac{\delta_2}{m} < 1$, so we may use the Taylor expansion of $\log(1-a(x))$ to get $\log(1-a(x)) = -a(x) - \frac{1}{2}a(x)^2 + R_3(-a(x))$, where $R_3(-a(x))$ is the remainder term. We use standard bounds on the remainder term in the Taylor expansion of $\log(1+x)$, (see Salas, Hille, and Etgen [32] for example) to obtain $|R_3(-a(x))| \le a(x)^2\frac{|a(x)|}{1-|a(x)|} \le a(x)^2\frac{\frac{1}{m}\delta_2}{1-\frac{1}{m}\delta_2} = a(x)^2\frac{\delta_2}{m-\delta_2} \le a(x)^2\delta_1$. Since $1-a(x)$ is positive we have:

$$
\begin{aligned}
(1-a(x))\log(1-a(x)) &= (1-a(x))[-a(x)-(1/2)a(x)^2+R_3(-a(x))] \\[1mm]
&\ge (1-a(x))[-a(x)-(1/2)a(x)^2-|R_3(-a(x))|] \\[1mm]
&\ge (1-a(x))[-a(x)-(1/2)a(x)^2-\delta_1 a(x)^2] \\[1mm]
&= -a(x)+(1/2-\delta_1)a(x)^2+(1/2+\delta_1)a(x)^3 \\[1mm]
&\ge -a(x)+(1/2-\delta_1)a(x)^2-(1/2+\delta_1)|a(x)|a(x)^2 \\[1mm]
&\ge -a(x)+(1/2-\delta_1)a(x)^2-(1/2+\delta_1)\frac{\delta_2}{m}a(x)^2 \\[1mm]
&\ge -a(x)+(1/2-\delta_1)a(x)^2-(1/2+\delta_1)\delta_1 a(x)^2 \\[1mm]
&\ge -a(x)+(1/2-3\delta_1)a(x)^2 \\[1mm]
&\ge -a(x)+Ka(x)^2
\end{aligned}
$$

72

So we have:

$$
\begin{aligned}
F(\nu) &= \frac{\sum_{x\in V}(1-a(x))\log\left(1-a(x)\right)\pi(x)}{\left(\sum_{x\in V}f(x)a(x)\pi(x)\right)^2}\\
&\geq \frac{\sum_{x\in V}(-a(x)+Ka(x)^2)\pi(x)}{\left(\sum_{x\in V}f(x)a(x)\pi(x)\right)^2}\\
&= \frac{K\sum_{x\in V}a(x)^2\pi(x)}{\left(\sum_{x\in V}f(x)a(x)\pi(x)\right)^2}\\
&\geq \frac{K}{\sum_{x\in V}f(x)^2\pi(x)} \qquad\qquad (60)\\
&= \frac{K}{\mathrm{Var}[f]}\\
&\geq \frac{K}{c^2(G)}\\
&= \frac{1}{2(1+\epsilon)c^2(G)}
\end{aligned}
$$

where the inequality in (60) is the Cauchy-Schwarz inequality. Let $B(\pi,\delta_2) = \{\mu \in P(G):$ $\|\mu-\pi\| < \delta_2\}$. $P(G)\backslash B(\pi,\delta_2)$ is closed and hence compact (since it is a subset of a compact set). Let $\nu_i \in P(G)\setminus\{\pi\}$ so that $F(\nu_i) \to 1/(2\sigma^2(G))$ as $i\to\infty$. Now

$$
\frac{1}{2\sigma^2(G)} < \frac{1}{2(1+\epsilon)c^2(G)}
$$

so there exists an integer N so that for all integers $i\geq N$ we have

$$
F(\nu_i) < \frac{1}{2(1+\epsilon)c^2(G)}.
$$

So $\nu_i \in P(G)\setminus B(\pi,\delta_2)$ for all integers $i\geq N$. Hence

$$
\inf_{\nu\in P(G)\backslash B(\pi,\delta_2)} F(\nu) = \frac{1}{2\sigma^2},
$$

Since $P(G)\setminus B(\pi,\delta_2)$ is compact and F is continuous on $P(G)\setminus B(\pi,\delta_2)$, the infimum is attained. Hence there exists $\nu \neq \pi$ with $W^2(\nu,\pi) = 2\sigma^2(G)D(\nu\|\pi)$. $\qquad\square$

Lemma 4.1.3. *Suppose there exists $\nu \in P(G)$ with $\nu \neq \pi$ and $W^2(\nu,\pi) = 2\sigma^2 D(\nu\|\pi)$. Then for every $C \subsetneq V$ we have:*

- *If $\sum_{x\in C}\nu(x) \geq \sum_{x\in C}\pi(x)$ then there exists a vertex $x \in C$ and a vertex $y \notin C$ with $f(x) - f(y) = d(x,y)$.*

- If $\sum_{x \in C} \nu(x) \leq \sum_{x \in C} \pi(x)$ then there exists a vertex $x \in C$ and a vertex $y \notin C$ with
$f(x) - f(y) = -d(x, y)$.

We note that if the distance d under consideration is the graph distance, then the vertices x and y may be taken to be neighbors.

Proof. Let f be a solution to Kantorovich's problem with respect to ν and π. For $C \subsetneq V$, define

$$f_\epsilon(x) = \begin{cases} f(x), & x \notin C \\ f(x) + \epsilon, & x \in C \end{cases}$$

Then

$$K(f_\epsilon) = \sum_{x \in V} f(x)(\nu(x) - \pi(x)) + \epsilon \sum_{x \in C} (\nu(x) - \pi(x)),$$

where the function K is defined in Section 1.3. First suppose that $\sum_{x \in C} \nu(x) > \sum_{x \in C} \pi(x)$. Since the coefficient of ϵ is positive and f is optimal, we must have that $f_\epsilon \notin Lip(G)$ for any positive ϵ. Then there exists $x \in C$ and $y \notin C$ with $f(x) - f(y) = d(x, y)$. Now suppose that $\sum_{x \in C} \nu(x) < \sum_{x \in C} \pi(x)$. The coefficient of ϵ is now negative and f is optimal, so we must have that $f_\epsilon \notin Lip(G)$ for any negative ϵ. Then there exists $x \in C$ and $y \notin C$ with $f(x) - f(y) = -d(x, y)$. Finally suppose that $\sum_{x \in C} \nu(x) = \sum_{x \in C} \pi(x)$. Then $K(f_\epsilon) = K(f)$ for every ϵ. Hence f_ϵ is a solution to Kantorovich's problem with respect to ν and π whenever f_ϵ is a Lipschitz function. Since C is a strict subset of V, f_ϵ is not a translation of f for any $\epsilon \neq 0$. By Corollary 2.3.4, f is the unique solution to Kantorovich's problem up to translation, so f_ϵ is not Lipschitz for any $\epsilon \neq 0$. The conclusion then follows. $\square$

The following lemma is inspired by Alon, Boppana, and Spencer's Theorem 2.1 concerning optimality of the spread constant [3]. For this lemma we assume we are using the graph distance.

Lemma 4.1.4. *Suppose that there exists $\nu \in P(G)$ with $\nu \neq \pi$ and $W^2(\nu, \pi) = 2\sigma^2 D(\nu||\pi)$. Then for any solution f to Kantorovich's problem with respect to ν and π, after a possible translation f will be integer valued and have the property that for some $U \subset V$, $f(x) = \pm d(x, U)$ for all $x \in V$.*

Proof. Let f be a solution to Kantorovich's problem with respect to ν and π. We start by showing that a translation of f will be integer valued. Consider the graph G_f with vertex set V and edge set $E_f \subset E$ where $\{x, y\} \in E_f$ if and only if $\{x, y\} \in E$ and $|f(x) - f(y)| = 1$. Then Lemma 4.1.3 shows that G_f is connected. Hence a translation of f will be integer valued. For the next part we assume f is integer valued and consider the following set:

$$U = \left\{ x \in V : \nu(x) \geq \pi(x) \text{ and } \frac{\nu(x)}{\pi(x)} \leq \frac{\nu(y)}{\pi(y)} \text{ for all } y \in V \text{ with } \nu(y) \geq \pi(y) \right\}$$

If $x, y \in U$, then $\frac{\nu(x)}{\pi(x)} = \frac{\nu(y)}{\pi(y)}$ and hence $f(x) = f(y)$ by Corollary 2.3.4. By translating f we may assume that $f(x) = 0$ for all $x \in U$. Let $O = \{|f(x)| : x \in V\}$. Then O contains every integer between 0 and some maximum value. The proof of the lemma will now be by induction on $|f(x)|$. For the base case let $x \in V$ with $|f(x)| = 0$. Then $f(x) = f(u)$ for some $u \in U$. Hence $\frac{\nu(x)}{\pi(x)} = \frac{\nu(u)}{\pi(u)}$ again by Corollary 2.3.4. So $x \in U$ and $d(x, U) = 0$, showing that $f(x) = d(x, U)$. Now let $m \in O$ and assume that $f(x) = \pm d(x, U)$ for any $x \in V$ with $|f(x)| \leq m$. Suppose $m + 1 \in O$. Let $z \in V$ with $|f(z)| = m + 1$. Suppose $f(z) > 0$. Since $f(z) > f(u)$ for some $u \in U$, we have $\frac{\nu(z)}{\pi(z)} > \frac{\nu(u)}{\pi(u)} \geq 1$ by Corollary 2.3.4. By Lemma 4.1.3, since $\nu(z) > \pi(z)$ there exists a neighbor x of z with $f(z) - f(x) = 1$. Then $f(x) = m$ and by the induction hypothesis, $f(x) = d(x, U)$. So $f(z) = d(z, x) + d(x, U) \geq d(z, U)$, by the triangle inequality. Let $u \in U$ with $d(z, u) = d(z, U)$. Then since $f \in Lip(G)$ we have $d(z, U) = d(z, u) \geq |f(z) - f(u)| = f(z) \geq d(z, U)$. Hence $f(z) = d(z, U)$. Now assume that $f(z) < 0$. Since $f(z) < f(u)$ for some $u \in U$, we have by Corollary 2.3.4 that $\frac{\nu(z)}{\pi(z)} < \frac{\nu(u)}{\pi(u)}$. So $\nu(z) < \pi(z)$ by the definition of U. Hence there exists a neighbor x of z with $f(z) - f(x) = -1$. So $f(x) = -m$ and $f(x) = -d(x, U)$ by the induction hypothesis. This means that $f(z) = -(d(z, x) + d(x, U)) \leq -d(z, U)$. Let $u \in U$ with $d(z, u) = d(z, U)$. Then since f is Lipschitz we have $d(z, U) = d(z, u) \geq |f(z) - f(u)| = -f(z) \geq d(z, U)$. This give us $f(z) = -d(z, U)$, which ends the induction step. Therefore, $f(x) = \pm d(x, U)$ for all $x \in V$. $\qquad\square$

4.2 Specific Purpose Lemmas

We start with a quick inequality that we need in the next lemma.

Lemma 4.2.1. $(x + y - 1)\log(x + y - 1) \geq x\log(x) + y\log(y)$ *for* $(x, y) \in A$ *where* $A = \{(x, y) \in \mathbb{R}^2 : (x + y \geq 1) \wedge (x, y \geq 1 \vee x, y \leq 1)\}$.

Proof. Let $f(x, y) = x\log(x) + y\log(y) - (x + y - 1)\log(x + y - 1)$. We must show that $f(x, y) \leq 0$ on A. Note that $f(1, y) = f(x, 1) = 0$ for all $(x, y) \in A$. It suffices to show that $\frac{\partial f}{\partial x}(x, y) \geq 0$ for $(x, y) \in A$ with $x, y \leq 1$ and $\frac{\partial f}{\partial x}(x, y) \leq 0$ for $(x, y) \in A$ with $x, y \geq 1$.

$$\frac{\partial f}{\partial x} = \log\left(\frac{x}{x + y - 1}\right),$$

so the lemma follows by noting that $\frac{x}{x+y-1} \geq 1$ for $(x, y) \in A$ with $x, y \leq 1$ and $\frac{x}{x+y-1} \leq 1$ for $(x, y) \in A$ with $x, y \geq 1$. $\qquad\square$

For the following two lemmas, let $G = (V, E)$ with $z \in V$ and $z_1, z_2 \notin V$. Let π be a probability measure on V and let $\tilde{\pi}$ be a probability measure on $\tilde{V} = (V \setminus \{z\}) \cup \{z_1, z_2\}$. Assume that $\tilde{\pi}(x) = k\pi(x)$ for $x \in V \setminus \{z\}$ and that $\tilde{\pi}(z_1) = \tilde{\pi}(z_2) = k\pi(z)$, where k is the constant necessary to make $\tilde{\pi}$ a probability measure. Let us note that $k = \frac{1}{1+\pi(z)}$, giving us $k = \frac{n}{n+1}$ when π is the uniform measure on V.

Lemma 4.2.2. *Let f be a probability density on V with respect π. Let g be a probability density on $\tilde{V}$ with respect to $\tilde{\pi}$. Assume that $g(x) = f(x)$ for $x \in V \setminus \{z\}$ and $f(z) = g(z_1) + g(z_2) - 1$. If $g(z_1), g(z_2) \leq 1$ or $g(z_1), g(z_2) \geq 1$, then $\mathrm{Ent}_\pi[f] \geq \frac{1}{k}\mathrm{Ent}_{\tilde{\pi}}[g]$.*

Proof.

$$
\begin{aligned}
\mathrm{Ent}_\pi[f] &= \sum_{x \in V}^{n} f(x)\log(f(x))\pi(x) \\
&= f(z)\log(f(z))\pi(z) + \sum_{x \in V \setminus \{z\}}^{n} f(x)\log(f(x))\pi(x) \\
&= (g(z_1) + g(z_2) - 1)\log(g(z_1) + g(z_2) - 1)\pi(z) + \sum_{x \in V \setminus \{z\}}^{n} g(x)\log(g(x))\pi(x) \\
&\geq g(z_1)\log(g(z_1))\pi(z) + g(z_2)\log(g(z_2))\pi(z) + \sum_{x \in V \setminus \{z\}}^{n} g(x)\log(g(x))\pi(x) \\
&= \frac{1}{k}\left[g(z_1)\log(g(z_1))\tilde{\pi}(z_1) + g(z_2)\log(g(z_2))\tilde{\pi}(z_2) + \sum_{x \in V \setminus \{z\}}^{n} g(x)\log(g(x))\tilde{\pi}(x)\right] \\
&= \frac{1}{k}\mathrm{Ent}_{\tilde{\pi}}[g].
\end{aligned}
$$

76

The lone inequality comes from Lemma 4.2.1 $\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

Let $\nu \in P(G)$ and let μ be a solution to Monge's problem with respect to ν and π. Assume z has the following properties:

- If $\nu(z) \geq \pi(z)$ then $x \in V$ with $\mu(x, z) > 0$ implies that $x = z$.

- If $\nu(z) \leq \pi(z)$ then $x \in V$ with $\mu(z, x) > 0$ implies that $x = z$.

Note that Lemma 2.1.5 guarantees that we can always find a μ so that these properties are satisfied for any z. Suppose $\tilde{G} = (\tilde{V}, \tilde{E})$ is a graph with distance function $\tilde{d}$ satisfying the following conditions:

1. $\tilde{d}(x, y) \geq d(x, y)$ for every $x, y \in V \setminus \{z\}$.

2. $\tilde{d}(x, y) = d(x, y)$ for every $x, y \in V \setminus \{z\}$ with $\mu(x, y) > 0$.

3. $\tilde{d}(x, z_1) \geq d(x, z)$ and $\tilde{d}(x, z_2) \geq d(x, z)$ for every $x \in V \setminus \{z\}$.

4. $\tilde{d}(x, z_1) = d(x, z)$ or $\tilde{d}(x, z_2) = d(x, z)$ for every $x \in V \setminus \{z\}$.

Then we get the following result.

Lemma 4.2.3. *There exists $\tilde{\nu} \in P(\tilde{G})$ satisfying the following properties:*

1. $W(\tilde{\nu}, \tilde{\pi}) = kW(\nu, \pi)$.

2. $\frac{\tilde{\nu}(x)}{\tilde{\pi}(x)} = \frac{\nu(x)}{\pi(x)}$ *for every* $x \in V \setminus \{z\}$.

3. $\frac{\tilde{\nu}(z_1)}{\tilde{\pi}(z_1)} + \frac{\tilde{\nu}(z_2)}{\tilde{\pi}(z_2)} - 1 = \frac{\nu(z)}{\pi(z)}$

4. *If* $\nu(z) \geq \pi(z)$ *then* $\tilde{\nu}(z_1) \geq \tilde{\pi}(z_1)$ *and* $\tilde{\nu}(z_2) \geq \tilde{\pi}(z_2)$. *If* $\nu(z) \leq \pi(z)$ *then* $\tilde{\nu}(z_1) \leq \tilde{\pi}(z_1)$ *and* $\tilde{\nu}(z_2) \leq \tilde{\pi}(z_2)$.

Proof. Let $V_1 = \{x \in V \setminus \{z\} : d(x,z) = \tilde{d}(x,z_1)\}$. Let $V_2 = (V \setminus \{z\}) \setminus V_1$. Then $d(x,z) = \tilde{d}(x,z_2)$ for all $x \in V_2$ by Condition 4. Define $\tilde{\mu} : \tilde{V} \times \tilde{V} \to \mathbb{R}$ by

$$
\begin{array}{ll}
\tilde{\mu}(x,y) = k\mu(x,y) & x, y \in V \setminus \{z\} \\[2mm]
\tilde{\mu}(z_1, x) = k\mu(z,x) & x \in V_1 \\[2mm]
\tilde{\mu}(x, z_1) = k\mu(x,z) & x \in V_1 \\[2mm]
\tilde{\mu}(z_1, x) = 0 & x \in V_2 \\[2mm]
\tilde{\mu}(x, z_1) = 0 & x \in V_2 \\[2mm]
\tilde{\mu}(z_2, x) = k\mu(z,x) & x \in V_2 \\[2mm]
\tilde{\mu}(x, z_2) = k\mu(x,z) & x \in V_2 \\[2mm]
\tilde{\mu}(z_2, x) = 0 & x \in V_1 \\[2mm]
\tilde{\mu}(x, z_2) = 0 & x \in V_1 \\[2mm]
\tilde{\mu}(z_1, z_1) = k\mu(z,z) + k\sum_{x \in V_2} \mu(x,z) \\[2mm]
\tilde{\mu}(z_2, z_2) = k\mu(z,z) + k\sum_{x \in V_1} \mu(x,z) \\[2mm]
\tilde{\mu}(z_2, z_1) = 0 \\[2mm]
\tilde{\mu}(z_1, z_2) = 0
\end{array}
$$

First let us verify that $\tilde{\mu}$ has $\tilde{\pi}$ as a second marginal. We start with $y \in \tilde{V} \setminus \{z_1, z_2\}$. Then

$$
\begin{aligned}
\sum_{x \in \tilde{V}} \tilde{\mu}(x,y) &= \sum_{x \in V \setminus \{z\}} \tilde{\mu}(x,y) + \tilde{\mu}(z_1, y) + \tilde{\mu}(z_2, y) \\[2mm]
&= k \left[\sum_{x \in V \setminus \{z\}} \mu(x,y) + \mu(z,y) \right] \\[2mm]
&= k \sum_{x \in V} \mu(x,y) \\[2mm]
&= k\pi(y) \\[2mm]
&= \tilde{\pi}(y).
\end{aligned}
$$

78

Next we will check z_1 and z_2:

$$\sum_{x \in \tilde{V}} \tilde{\mu}(x, z_1) = \sum_{x \in V_1} \tilde{\mu}(x, z_1) + \sum_{x \in V_2} \tilde{\mu}(x, z_1) + \tilde{\mu}(z_1, z_1) + \tilde{\mu}(z_2, z_1)$$

$$= k \left[\sum_{x \in V_1} \mu(x, z) + 0 + \mu(z, z) + \sum_{x \in V_2} \mu(x, z) + 0 \right]$$

$$= k \sum_{x \in V} \mu(x, z)$$

$$= k\pi(z)$$

$$= \tilde{\pi}(z_1).$$

Similarly we have:

$$\sum_{x \in \tilde{V}} \tilde{\mu}(x, z_2) = \sum_{x \in V_1} \tilde{\mu}(x, z_2) + \sum_{x \in V_2} \tilde{\mu}(x, z_2) + \tilde{\mu}(z_1, z_2) + \tilde{\mu}(z_2, z_2)$$

$$= k \left[0 + \sum_{x \in V_2} \mu(x, z) + 0 + \mu(z, z) + \sum_{x \in V_1} \mu(x, z) \right]$$

$$= k \sum_{x \in V} \mu(x, z)$$

$$= k\pi(z)$$

$$= \tilde{\pi}(z_2).$$

Now define $\tilde{\nu} : \tilde{V} \to \mathbb{R}$ by $\tilde{\nu}(x) = \sum_{y \in \tilde{V}} \tilde{\mu}(x, y)$. Let us verify Property 2 of the lemma. Let $x \in V \setminus \{z\}$.

$$\tilde{\nu}(x) = \sum_{y \in \tilde{V}} \tilde{\mu}(x, y)$$

$$= \sum_{y \in V \setminus \{z\}} \tilde{\mu}(x, y) + \tilde{\mu}(x, z_1) + \tilde{\mu}(x, z_2)$$

$$= k \left[\sum_{y \in V \setminus \{z\}} \mu(x, y) + \mu(x, z) \right]$$

$$= k \sum_{y \in V} \mu(x, y)$$

$$= k\nu(x)$$

$$= \frac{\tilde{\pi}(x)}{\pi(x)} \nu(x).$$

And to check Property 3 we calculate:

$$
\begin{aligned}
\tilde{\nu}(z_1) + \tilde{\nu}(z_2) \; &= \; \sum_{y \in \tilde{V}} \tilde{\mu}(z_1, y) + \sum_{y \in \tilde{V}} \tilde{\mu}(z_2, y) \\
&= \; \sum_{y \in V_1} \tilde{\mu}(z_1, y) + \sum_{y \in V_2} \tilde{\mu}(z_1, y) + \tilde{\mu}(z_1, z_1) + \tilde{\mu}(z_1, z_2) \\
&\quad + \sum_{y \in V_1} \tilde{\mu}(z_2, y) + \sum_{y \in V_2} \tilde{\mu}(z_2, y) + \tilde{\mu}(z_2, z_1) + \tilde{\mu}(z_2, z_2) \\
&= \; k \left[\sum_{y \in V_1} \mu(z, y) + 0 + \mu(z, z) + \sum_{x \in V_2} \mu(x, z) + 0 \right] \\
&\quad + k \left[0 + \sum_{y \in V_2} \mu(z, y) + 0 + \mu(z, z) + \sum_{x \in V_1} \mu(x, z) \right] \\
&= \; k \sum_{x \in V} \mu(z, x) + k \sum_{x \in V} \mu(x, z) \\
&= \; k[\nu(z) + \pi(z)].
\end{aligned}
$$

This is what we want after dividing both sides by $k\pi(z)$ and recalling that $\tilde{\pi}(z_1) = \tilde{\pi}(z_2) = k\pi(z)$. Now we will verify Property 4.

$$
\begin{aligned}
\tilde{\nu}(z_1) \; &= \; \sum_{y \in \tilde{V}} \tilde{\mu}(z_1, y) \\
&= \; \sum_{y \in V_1} \tilde{\mu}(z_1, y) + \sum_{y \in V_2} \tilde{\mu}(z_1, y) + \tilde{\mu}(z_1, z_1) + \tilde{\mu}(z_1, z_2) \\
&= \; k \left[\sum_{y \in V_1} \mu(z, y) + 0 + \mu(z, z) + \sum_{x \in V_2} \mu(x, z) + 0 \right] \\
&= \; k \left[\sum_{y \in V_1} \mu(z, y) - \sum_{x \in V_1} \mu(x, z) + \sum_{x \in V} \mu(x, z) \right] \\
&= \; k \left[\sum_{y \in V_1} \mu(z, y) - \sum_{x \in V_1} \mu(x, z) + \pi(z) \right]
\end{aligned}
$$

If $\nu(z) \geq \pi(z)$, then $\mu(x, z) > 0$ implies that $x = z$ and so

$$
\tilde{\nu}(z_1) = k \left[\sum_{y \in V_1} \mu(z, y) + \pi(z) \right] \geq k\pi(z) = \tilde{\pi}(z_1).
$$

If $\nu(z) \leq \pi(z)$, then $\mu(z, y) > 0$ implies that $y = z$ and so

$$
\tilde{\nu}(z_1) = k \left[-\sum_{x \in V_1} \mu(x, z) + \pi(z) \right] \leq k\pi(z) = \tilde{\pi}(z_1).
$$

80

Similarly for z_2 we get:

$$
\begin{aligned}
\tilde{\nu}(z_2) &= \sum_{y \in \tilde{V}} \tilde{\mu}(z_2, y) \\
&= \sum_{y \in V_1} \tilde{\mu}(z_2, y) + \sum_{y \in V_2} \tilde{\mu}(z_2, y) + \tilde{\mu}(z_2, z_1) + \tilde{\mu}(z_2, z_2) \\
&= k \left[0 + \sum_{y \in V_2} \mu(z, y) + 0 + \mu(z, z) + \sum_{x \in V_1} \mu(x, z) \right] \\
&= k \left[\sum_{y \in V_2} \mu(z, y) - \sum_{x \in V_2} \mu(x, z) + \sum_{x \in V} \mu(x, z) \right] \\
&= k \left[\sum_{y \in V_2} \mu(z, y) - \sum_{x \in V_2} \mu(x, z) + \pi(z) \right]
\end{aligned}
$$

If $\nu(z) \geq \pi(z)$, then $\mu(x, z) > 0$ implies that $x = z$ and so

$$
\tilde{\nu}(z_2) = k \left[\sum_{y \in V_2} \mu(z, y) + \pi(z) \right] \geq k\pi(z) = \tilde{\pi}(z_2).
$$

If $\nu(z) \leq \pi(z)$, then $\mu(z, y) > 0$ implies that $y = z$ and so

$$
\tilde{\nu}(z_2) = k \left[-\sum_{x \in V_2} \mu(x, z) + \pi(z) \right] \leq k\pi(z) = \tilde{\pi}(z_2).
$$

Before verifying Property 1, we will use properties 2 and 3 to show that indeed $\tilde{\nu}$ is a probability measure on $\tilde{V}$. From the definition of $\tilde{\nu}$ we get that $\tilde{\nu}(x) \geq 0$ for all $x \in \tilde{V}$. And

$$
\begin{aligned}
\sum_{x \in \tilde{V}} \tilde{\nu}(x) &= \tilde{\nu}(z_1) + \tilde{\nu}(z_2) + \sum_{x \in V \setminus \{z\}} \tilde{\nu}(x) \\
&= k[\nu(z) + \pi(z)] + k \sum_{x \in V \setminus \{z\}} \nu(x) \\
&= k \left(\pi(z) + \sum_{x \in V} \nu(x) \right) \\
&= k \left(\pi(z) + 1 \right) \\
&= k \left(\pi(z) + \pi(z) + \sum_{x \in V \setminus \{z\}} \pi(x) \right) \\
&= \tilde{\pi}(z_1) + \tilde{\pi}(z_2) + \sum_{x \in \tilde{V} \setminus \{z_1, z_2\}} \tilde{\pi}(x) \\
&= 1.
\end{aligned}
$$

81

Now we only need to verify Property 1. Let f be an optimal solution to Kantorovich's problem on G with respect to ν and π. Define $\tilde{f} : \tilde{V} \to \mathbb{R}$ by $\tilde{f}(x) = f(x)$ for $x \in \tilde{V} \setminus \{z_1, z_2\}$ and $\tilde{f}(z_1) = \tilde{f}(z_2) = f(z)$. First let us verify that $\tilde{f}$ is Lipschitz with respect to $\tilde{d}$. Suppose $x, y \in \tilde{V} \setminus \{z_1, z_2\}$. Then by Condition 1 we have:

$$|\tilde{f}(x) - \tilde{f}(y)| = |f(x) - f(y)| \leq d(x, y) \leq \tilde{d}(x, y).$$

By Condition 3, for $i \in \{1, 2\}$ and for all $x \in \tilde{V} \setminus \{z_1, z_2\}$ we have:

$$|\tilde{f}(x) - \tilde{f}(z_i)| = |f(x) - f(z)| \leq d(x, z) \leq \tilde{d}(x, z_i).$$

Finally $|\tilde{f}(z_1) - \tilde{f}(z_2)| = 0 \leq \tilde{d}(z_1, z_2)$. So $\tilde{f}$ is Lipschitz with respect to $\tilde{d}$. Now we will use Lemma 2.1.4 to show that $\tilde{f}$ is a solution to Kantorovich's problem and $\tilde{\mu}$ is a solution to Monge's problem both with respect to $\tilde{\nu}$ and $\tilde{\pi}$. Suppose $x, y \in \tilde{V} \setminus \{z_1, z_2\}$ with $\tilde{\mu}(x, y) > 0$. Then by the definition of $\tilde{\mu}$ we must also have $\mu(x, y) > 0$. So by the definition of $\tilde{f}$, Lemma 2.1.4, and Condition 2 we get:

$$\tilde{f}(x) - \tilde{f}(y) = f(x) - f(y) = d(x, y) = \tilde{d}(x, y)$$

If $\tilde{\mu}(x, z_i) > 0$ for some $x \in \tilde{V} \setminus \{z_1, z_2\}$ and $i \in \{1, 2\}$, then $\mu(x, z) > 0$ and $x \in V_i$ so that

$$\tilde{f}(x) - \tilde{f}(z_i) = f(x) - f(z) = d(x, z) = \tilde{d}(x, z_i).$$

Similarly, if $\tilde{\mu}(z_i, y) > 0$ for some $y \in \tilde{V} \setminus \{z_1, z_2\}$ and $i \in \{1, 2\}$, then $\mu(z, y) > 0$ and $y \in V_i$ so that

$$\tilde{f}(z_i) - \tilde{f}(y) = f(z) - f(y) = d(z, y) = \tilde{d}(z_i, y).$$

Finally we note that $\tilde{\mu}(z_1, z_2) = \tilde{\mu}(z_2, z_1) = 0$, and $\tilde{f}(z_i) - \tilde{f}(z_i) = 0 = \tilde{d}(z_i, z_i)$ for $i \in \{1, 2\}$. Hence for $x, y \in \tilde{V}$, $\tilde{\mu}(x, y) > 0$ implies that $\tilde{f}(x) - \tilde{f}(y) = \tilde{d}(x, y)$. So by Lemma 2.1.4 $\tilde{\mu}$ is a solution to Monge's problem and $\tilde{f}$ is a solution to Kantorovich's problem, both on $\tilde{G}$ with respect to $\tilde{\nu}$ and $\tilde{\pi}$. And now we can finish the verification of property 1 (using properties

2 and 3):

$$W(\tilde{\nu}, \tilde{\pi}) = \sum_{x \in \tilde{V}} \tilde{f}(x)(\tilde{\nu}(x) - \tilde{\pi}(x))$$

$$= \sum_{x \in V \setminus \{z\}} \tilde{f}(x)(\tilde{\nu}(x) - \tilde{\pi}(x)) + \tilde{f}(z_1)(\tilde{\nu}(z_1) - \tilde{\pi}(z_1)) + \tilde{f}(z_2)(\tilde{\nu}(z_2) - \tilde{\pi}(z_2))$$

$$= \sum_{x \in V \setminus \{z\}} f(x)(k\nu(x) - k\pi(x)) + f(z)(k\nu(z) + k\pi(z) - k\pi(z) - k\pi(z))$$

$$= k \sum_{x \in V} f(x)(\nu(x) - \pi(x))$$

$$= kW(\nu, \pi). \qquad \qquad \square$$

For a solution μ to Monge's problem with respect to ν and π, define the equivalence relation $\sim_\mu$ on V to be the smallest equivalence relation for which $x \sim_\mu y$ if $\mu(x, y) > 0$ or $\mu(y, x) > 0$. Let $\{V_i\}_{i=1}^m$ be the equivalence classes generated by $\sim_\mu$. Let $G_i = (V_i, E_i)$ for $i \in [m]$ be the subgraphs of G induced by V_i. Let π_i be a probability measure on V_i defined by $\pi_i(x) = k_i \pi(x)$ for $x \in V_i$, where k_i is the appropriate constant that makes π_i a probability measure. We will note that if π is the uniform measure on V, then π_i is the uniform measure on V_i and $k_i = \frac{n}{|V_i|}$. Let d_i denote the distance function on G_i defined by $d_i(x, y) = d(x, y)$ for $x, y \in V_i$. Define $\nu_i \in P(G_i)$ by $\nu_i(x) = k_i \nu(x)$, for $x \in V_i$. Before continuing let us verify that indeed ν_i is a probability measure on G_i.

$$\begin{aligned}
\sum_{x \in V_i} \nu_i(x) &= \sum_{x \in V_i} k_i \nu(x) \\
&= \sum_{x \in V_i} k_i \sum_{y \in V} \mu(x, y) \\
&= \sum_{x \in V_i} k_i \sum_{y \in V_i} \mu(x, y) \\
&= \sum_{y \in V_i} k_i \sum_{x \in V_i} \mu(x, y) \\
&= \sum_{y \in V_i} k_i \sum_{x \in V} \mu(x, y) \\
&= \sum_{y \in V_i} k_i \pi(y) \\
&= \sum_{y \in V_i} \pi_i(y) \\
&= 1.
\end{aligned}$$

83

Then we get the following lemma.

Lemma 4.2.4.

$$W(\nu, \pi) = \sum_{i=1}^{m} \frac{1}{k_i} W(\nu_i, \pi_i)$$

Proof. Let μ be a solution to Monge's problem with respect to ν and π and let f be a solution to Kantorovich's problem with respect to ν and π. Define $\mu_i : V_i \times V_i \to \mathbb{R}$ by

$$\mu_i(x, y) = k_i \mu(x, y), \text{ for } x, y \in V_i$$

and $f : V_i \to \mathbb{R}$ by

$$f_i(x) = f(x), \text{ for } x \in V_i.$$

Let us compute the marginals of μ_i. Suppose $z \in V_i$. Then

$$
\begin{aligned}
\sum_{x \in V_i} \mu_i(x, z) &= \sum_{x \in V_i} k_i \mu(x, z) \\
&= \sum_{x \in V} k_i \mu(x, z) \\
&= k_i \pi(z) \\
&= \pi_i(z).
\end{aligned}
$$

$$
\begin{aligned}
\sum_{x \in V_i} \mu_i(z, x) &= \sum_{x \in V_i} k_i \mu(z, x) \\
&= \sum_{x \in V} k_i \mu(z, x) \\
&= k_i \nu(z) \\
&= \nu_i(z)
\end{aligned}
$$

By Lemma 2.1.4, μ_i is a solution to Monge's problem and f_i is a solution to Kantorovich's problem both on G_i with respect to ν_i and π_i. We may now verify the statement of the

lemma:

$$\sum_{i \in [m]} \frac{1}{k_i} W(\nu_i, \pi_i) = \sum_{i \in [m]} \frac{1}{k_i} \sum_{x,y \in V_i} d_i(x,y)\mu_i(x,y)$$

$$= \sum_{i \in [m]} \frac{1}{k_i} \sum_{x,y \in V_i} d(x,y)k_i\mu(x,y)$$

$$= \sum_{i \in [m]} \sum_{x,y \in V_i} d(x,y)\mu(x,y)$$

$$= \sum_{x,y \in V} d(x,y)\mu(x,y)$$

$$= W(\nu, \pi). \qquad \square$$

4.3 The Subgaussian Constant of a Cycle

For this entire section we assume that the measure π associated with the graph of interest in the uniform probability measure and d is the graph distance.

Lemma 4.3.1. *Suppose f is an integer valued Lipschitz function on the vertices of a cycle $C = (V, E)$. Then there exist $a, b \in V$ and a permutation p of V that satisfy the following properties:*

- *$f(p(x))$ is Lipschitz.*

- *$f(p(x))$ is non-decreasing along the two internally disjoint paths from a to b.*

Proof. Let $m = \min_{x \in V} f(x)$ and $M = \max_{x \in V} f(x)$. Let $w_1, w_2 \in V$ with $f(w_1) = m$ and $f(w_2) = M$. Let $[m, M]$ denote the integers between and including m and M. Since f is integer valued and Lipschitz, and because C is a connected graph, $f(V) = [m, M]$. Suppose c is an integer with $m < c < M$ and let $x_1 \in V$ with $f(x_1) = c$. Since $C - x_1$ is still a connected graph and f is Lipschitz on $C - x_1$, we also have $f(V \setminus \{x_1\}) = [m, M]$ and so there exists $x_2 \in V \setminus \{x_1\}$ with $f(x_2) = c$. Hence we can find $V_1, V_2 \subset V$ with $V_1 \cup V_2 = V$, $V_1 \cap V_2 = \{w_1, w_2\}$, and $f(V_1) = f(V_2) = [m, M]$. From this we can form paths P_1 and P_2 (not necessarily subgraphs of C) with vertex sets V_1 and V_2 respectively with the property that f is non-decreasing on each path from w_1 to w_2. It also follows that f is Lipschitz on each path. We can then form the cycle $J = (V, E(P_1) \cup E(P_2))$, which has the property

that f is Lipschitz and non-decreasing on the two internally disjoint paths from w_1 to w_2. Let p be an isomorphism between C and J, and let $a = p^{-1}(w_1)$ and $b = p^{-1}(w_2)$. Then a, b and p are the desired vertices and permutation. $\square$

Lemma 4.3.2. *Let $C = (V, E)$ be a cycle and suppose there exists $\nu \in P(C)$ with $\nu \neq \pi$ and $W^2(\nu, \pi) = 2\sigma^2 D(\nu||\pi)$. Then there exists $z \in V$ with the property that one of the functions $f(x) = d(x, z)$ or $f(x) = -d(x, z)$ is a solution to Kantorovich's problem with respect to ν and π.*

Proof. By Lemma 4.1.4, let f be an integer valued solution to Kantorovich's problem with respect to ν and π. Then there exist vertices a and b and a permutation p of V satisfying the properties of Lemma 4.3.1. Let $\tilde{f}(x) = f(p(x))$. Let $\tilde{\nu}(x) = \nu(p(x))$. Since the image of V under $\tilde{\nu}$ is equal as a multiset to the image of V under ν, we have $D(\tilde{\nu}||\pi) = D(\nu||\pi)$. Also,

$$
\begin{aligned}
W(\nu, \pi) &= \sum_{x \in C} f(x)(\nu(x) - \pi(x)) \\
&= \sum_{x \in C} f(p(x))(\nu(p(x)) - \pi(p(x))) \\
&= \sum_{x \in C} \tilde{f}(x)(\tilde{\nu}(x) - \pi(x)) \\
&\leq W(\tilde{\nu}, \pi).
\end{aligned}
$$

This leads us to:

$$
\frac{1}{2\sigma^2} = \frac{D(\nu||\pi)}{W^2(\nu, \pi)} \geq \frac{D(\tilde{\nu}||\pi)}{W^2(\tilde{\nu}, \pi)} \geq \frac{1}{2\sigma^2},
$$

and so the inequalities must actually be equalities. This means that $\tilde{\nu}$ gives equality in the transportation inequality and $\tilde{f}$ is a solution to Kantorovich's problem with respect to $\tilde{\nu}$ and π.

Let P_1 and P_2 be the two internally disjoint paths from a to b. Since $\tilde{f}$ is non-decreasing along P_1 and P_2 from a to b, we have $\tilde{f}(a) \leq \tilde{f}(x) \leq \tilde{f}(b)$ for every $x \in V$. Since $\tilde{f}$ is Lipschitz, for any integer c with $\tilde{f}(a) < c < \tilde{f}(b)$ there must exist $x_1 \in P_1$ and $x_2 \in P_2$ with $\tilde{f}(x_1) = \tilde{f}(x_2) = c$. Let $i \in \{1, 2\}$ and suppose to the contrary that there exist vertices x' and x'' both in P_i with $\tilde{f}(x') = \tilde{f}(x'') = c$. Since $\tilde{f}$ is non-decreasing along
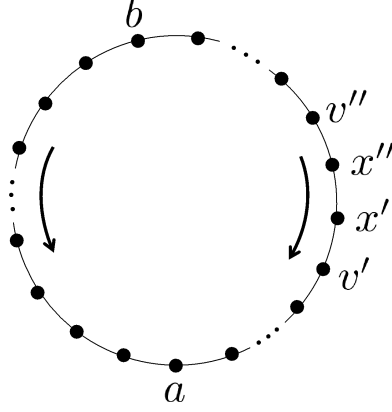
Figure 4: Distance From a Point

P_i there exist adjacent x' and x'' with this property. Without loss of generality we can assume that $aP_i x' x'' P_i b$. By Corollary 2.3.4, $\tilde{\nu}(x') = \tilde{\nu}(x'')$ since $\tilde{f}(x') = \tilde{f}(x'')$. Let v' be the neighbor of x' other than x'' and let v'' be the neighbor of x'' other than x', so that we have $aP_i v' x' x'' v'' P_i b$ (see Figure 4). Then $\tilde{f}(v') \leq \tilde{f}(x') = \tilde{f}(x'') \leq \tilde{f}(v'')$ since $\tilde{f}$ is non-decreasing along P_i. If $\tilde{\nu}(x') = \tilde{\nu}(x'') \geq \pi(x'')$ then by Lemma 4.1.3 there must exist a neighbor y'' of x'' with $f(y'') < f(x'')$. But x'' has only two neighbors, so this is a contradiction. Similarly, if $\pi(x') \leq \tilde{\nu}(x') = \tilde{\nu}(x'')$ then there exists a neighbor y' of x' with $f(y') > f(x')$ which is again a contradiction. Hence for every integer c with $\tilde{f}(a) < c < \tilde{f}(b)$, for each $i \in \{1, 2\}$, there exists exactly one vertex $x \in P_i$ with $\tilde{f}(x) = c$. Next assume that there exist three adjacent vertices $r \sim s \sim t \in V$ with $\tilde{f}(r) = \tilde{f}(s) = \tilde{f}(t)$ (which is only possible if this joint value is $\tilde{f}(a)$ or $\tilde{f}(b)$). By Lemma 4.1.3, s must have a neighbor x with $\tilde{f}(x) > \tilde{f}(s)$ or $\tilde{f}(x) < \tilde{f}(s)$, which is a contradiction. Now there are an even number of vertices for which $\tilde{f}$ attains values strictly between $\tilde{f}(a)$ and $\tilde{f}(b)$, and there are at most two vertices which attain the maximum value, $\tilde{f}(b)$, and at most two which attain the minimum value, $\tilde{f}(a)$. So if $|V|$ is odd, there exists only one vertex on which the maximum value of $\tilde{f}$ is attained or one vertex on which the minimum value of $\tilde{f}$ is attained. By translating $\tilde{f}$ so that respectively either the maximum or minimum value is zero, we get that $\tilde{f}(x) = -d(x, b)$ or $\tilde{f}(x) = d(x, a)$. If $|V|$ is even then we must show there cannot be two vertices on which $\tilde{f}$ attains the maximum value and two vertices on which $\tilde{f}$ attains the minimum value. Let P be a path in C along which $\tilde{f}$ is strictly increasing, starting from one of the vertices on

which $\tilde{f}$ attains the minimum value and ending on one of the vertices on which $\tilde{f}$ attains the maximum value. Then by Lemma 4.1.3 either the vertex of P which attains the maximum value of $\tilde{f}$ must have a neighbor outside P with a different value of $\tilde{f}$ or the vertex of P which attains the minimum value of $\tilde{f}$ must have a neighbor outside P with a different value of $\tilde{f}$. Hence there can only be one vertex on which $\tilde{f}$ attains the maximum value or one vertex on which $\tilde{f}$ attains the minimum value. But since $|V|$ is even, $\tilde{f}$ attains both the maximum and minimum values at only one vertex. By translating $\tilde{f}$ we can choose to get either $\tilde{f}(x) = -d(x, b)$ or $\tilde{f}(x) = d(x, a)$. Now $f(V)$ is equal to $\tilde{f}(V)$ as a multiset. But up to rotations of the cycle and translations of the function, there is only one integer valued Lipschitz function on C with this image as a multiset. So f is just a translation and a rotation of $\tilde{f}$. Hence after a possible translation, $f(x) = d(x, z)$ for some vertex $z \in V$ or $f(x) = -d(x, z)$ for some vertex $z \in V$. $\qquad\square$

We now prove the main result in three propositions. For the first proof, we employ the technique used by Bobkov, Houdré, and Tetali [8] to show that $\sigma^2(C_4) = c^2(C_4)$.

Proposition 4.3.3. *If C is a cycle with an even number of vertices, then $\sigma^2(C) = c^2(C)$.*

Proof. Let $C = (V, E)$ be a cycle on $2n$ vertices. Let π be the uniform measure on V so that $\pi(x) = \frac{1}{2n}$ for every $x \in V$. Assume to the contrary that $\sigma^2 \neq c^2$. Let x_0 be an arbitrary vertex in V. Let $f(x) = d(x, x_0)$. From Lemmas 4.1.2 and 4.3.2 and Proposition 2.3.2 we know that there exists a $t \neq 0$ such that $\mathrm{E}\left[e^{t(f - \mathrm{E}f)}\right] = e^{\sigma^2 t^2/2}$. So $\sigma^2(C)$ is actually the smallest constant s so that for this particular f, $\mathrm{E}\left[e^{t(f - \mathrm{E}f)}\right] \leq e^{st^2/2}$ for every real number t. Let $L_f(t) = \log \mathrm{E}\left[e^{t(f - \mathrm{E}f)}\right]$. Now $\mathrm{E}[f] = n/2$. So

$$
\begin{aligned}
L_f(t) &= \log\left[\frac{1}{2n}\left(e^{t(0 - n/2)} + 2\sum_{i=1}^{n-1} e^{t(i - n/2)} + e^{t(n - n/2)}\right)\right] \\
&= \begin{cases} \log\left[\frac{1}{2n}\left(e^{\frac{nt}{2}} + e^{-\frac{nt}{2}} - \frac{2e^{-\frac{nt}{2}}(e^t - e^{nt})}{e^t - 1}\right)\right] & t \neq 0 \\ 0 & t = 0 \end{cases}
\end{aligned}
$$

Consider the function

$$
\phi(t) = L_f(t) - \frac{st^2}{2}.
$$

88

Then $\sigma^2(C)$ is the smallest constant s for which $\phi(t) \leq 0$ for every real number t. We will consider the following derivatives of ϕ:

$$\phi'(t) = L_f'(t) - s^2 t, \quad \phi''(t) = L_f''(t) - s^2, \quad \text{and} \quad \phi'''(t) = L_f'''(t)$$

Since $L_f(t)$ is an even function, $\phi(t)$ is also an even function, so $\phi'(t)$ is an odd function and $\phi'(0) = 0$. We also have that $\phi(0) = 0$. Then in order to have $\phi(t) \leq 0$ for all real t, we must have $\phi''(0) \leq 0$, implying $s \geq L_f''(0)$. But, in fact, we will show that if we set $s = L_f''(0)$ then $\phi(t) \leq 0$ for all real t. So the smallest constant s for which $\phi(t) \leq 0$ for all real t is $L_f''(0)$, meaning that $\sigma^2(C) = L_f''(0)$.

To show that $\phi(t) \leq 0$ for all real t when $s = L_f''(0)$ we first restrict ourselves to $t \geq 0$ since $\phi(t)$ is an even function. Then we show that $L_f''(t) < L_f''(0)$ for every $t > 0$. Hence $\phi''(0) = 0$ and $\phi''(t) < 0$ for all $t > 0$, giving us $\phi'(0) = 0$ and $\phi'(t) < 0$ for all $t > 0$, finally giving us $\phi(0) = 0$ and $\phi(t) < 0$ for all $t > 0$.

Now to show that $\phi''(t) < \phi''(0)$ for all $t > 0$, we note that $\phi'''(0) = 0$ again because $\phi(t)$ is an even function. Then we show that $\phi'''(t) < 0$ for all $t > 0$. Now for $t \neq 0$ we have:

$$
\begin{aligned}
\phi'''(t) = L_f'''(t) \quad = \quad & \frac{1}{(e^t - 1)^3 (e^t + 1)^3 (e^{nt} - 1)^3} \cdot \\
& \Big(2e^t + 12e^{3t} + 2e^{5t} + 36e^{(3+2n)t} + 6e^{(5+2n)t} + 6e^{(1+2n)t} \\
& + 3n^3 e^{(2+2n)t} + 3n^3 e^{(4+2n)t} + n^3 e^{(6+2n)t} + n^3 e^{(6+n)t} \\
& - 12e^{(3+3n)t} - 36e^{(3+n)t} - 6e^{(5+n)t} - 2e^{(5+3n)t} - 6e^{(1+n)t} \\
& - 2e^{(1+3n)t} - n^3 e^{nt} - n^3 e^{2nt} - 3n^3 e^{(4+2n)t} - 3n^3 e^{(4+n)t} \Big).
\end{aligned}
$$

Although we are only interested in positive integers $n \geq 2$, for a fixed t, if we allow n to take on any positive real value then $\phi(t)$ is a differentiable function of n. Now for $n = 2$ we have:

$$\phi'''(t) = -\frac{2e^t(e^t - 1)}{(e^t + 1)^3}$$

and so $\phi'''(t) < 0$ for any $t > 0$ in this case. Finally we show that for every $t > 0$

$$\frac{\partial}{\partial n} \phi'''(t) < 0,$$

for all $n \geq 2$. Hence $\phi(t) < 0$ for all integers $n \geq 2$ and real $t > 0$.

So for $t > 0$ we calculate:

$$\frac{\partial}{\partial n}\phi'''(t) = \frac{-n^2}{8\sinh\left((nt/2)^4\right)}(nt(2 + \cosh(nt)) - 3\sinh(nt))$$

To show that $\frac{\partial}{\partial n}\phi'''(t) < 0$ for every real $n \geq 2$ and $t > 0$ it suffices to show that

$$\psi(t) = nt(2 + \cosh(nt)) - 3\sinh(nt) > 0$$

for every real $n \geq 2$ and $t > 0$. We'll start by taking some derivatives:

$$\tfrac{d}{dt}\psi(t) = n(2 - 2\cosh(nt) + nt\sinh(nt))$$

$$\tfrac{d^2}{dt^2}\psi(t) = n^2(nt\cosh(nt) - \sinh(nt))$$

$$\tfrac{d^3}{dt^3}\psi(t) = n^4 t\sinh(nt))$$

Now $\psi(t)$, $\frac{d}{dt}\psi(t)$, and $\frac{d^2}{dt^2}\psi(t)$ are zero when $t = 0$, and $\frac{d^3}{dt^3}\psi(t)$ is strictly positive for $t > 0$. Hence $\frac{d^2}{dt^2}\psi(t)$, $\frac{d}{dt}\psi(t)$ and $\psi(t)$ are all strictly positive for $t > 0$.

Now that we have shown that $\sigma^2(C) = L_f''(0)$ we may take our pick of contradictions. First, we have shown that $\phi(t) < 0$ for $t \neq 0$ when $s = \sigma^2(C) = L_f''(0)$. This contradicts the fact that there exists a $t \neq 0$ for which $\mathrm{E}\left[e^{t(f - \mathrm{E}f)}\right] = e^{\sigma^2 t^2/2}$. Or we could note that $L_g''(0) = \mathrm{Var}[g]$ for any g. Bobkov, Houdré, and Tetali [8] showed that $\mathrm{Var}[f] = c^2(C)$ for our particular function f. Hence we have shown that $\sigma^2(C) = c^2(C)$, another contradiction. Therefore, in fact, $\sigma^2(C) = c^2(C)$. □

Proposition 4.3.4. *If C is a cycle with an odd number of vertices, then $\sigma^2(C) > c^2(C)$.*

Proof. Let C be a cycle with $2n + 1$ vertices. Bobkov, Houdré, and Tetali show in [8] that for any vertex x_0, the function $f(x) = d(x, x_0)$ is optimal for the spread constant of this graph, meaning that $\mathrm{Var}[f] = c^2(C)$. Now $\mathrm{E}[f] = \frac{n(1+n)}{1+2n}$. If we set $g(x) = f(x) - \mathrm{E}[f]$, then $\mathrm{Var}[g] = c^2$ and $\mathrm{E}[g] = 0$, but $\mathrm{E}[g^3] = -\frac{n^2(1+n)^2}{2(1+2n)^2} \neq 0$. So by Lemma 4.1.1 we have $\sigma^2(C) > c^2(C)$. □

Proposition 4.3.5. *Suppose that C is a cycle with an odd number of vertices. Then $\sigma^2(C) = c^2(C)(1 + o(1))$, where $o(1)$ goes to 0 as the number of vertices goes to infinity.*

90
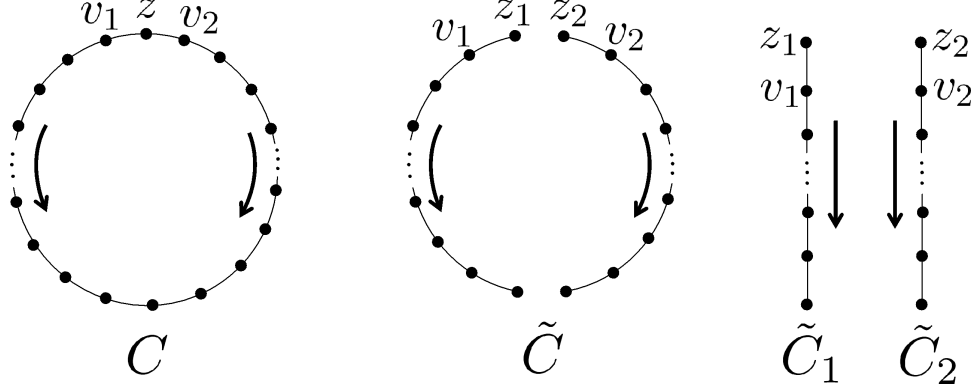
Figure 5: Bounding the Subgaussian on the Odd Cycle

Proof. Let $C = (V, E)$ be a cycle on n vertices, where $n \geq 3$ is an odd integer. From Proposition 4.3.4 we know that $\sigma^2(C) \neq c^2(C)$. Hence, by Lemma 4.1.2, there exists $\nu \in P(G)$ with $\nu \neq \pi$ and $W^2(\nu, \pi) = 2\sigma^2 D(\nu || \pi)$. Let μ be a solution to Monge's problem with respect to ν and π given to us by Lemma 2.1.5. By Lemma 4.3.2, there exists $z \in V$ so that either the function $f(x) = d(x, z)$ or the function $f(x) = -d(x, z)$ is an optimal solution to Kantorovich's problem with respect to ν and π. Let v_1 and v_2 be the two neighbors of z. For $z_1, z_2 \notin V$, let $\tilde{C} = (\tilde{V}, \tilde{E})$ be the graph obtained from C by

- $\tilde{V} = (V \setminus \{z\}) \cup \{z_1, z_2\}$.

- For $x, y \in V \setminus \{z\}$, $\{x, y\} \in \tilde{E}$ if and only if $\{x, y\} \in E$ and $|f(x) - f(y)| = 1$.

- $\{z_1, v_1\}, \{z_2, v_2\} \in \tilde{E}$.

We will verify that $\tilde{C}$ satisfies the conditions before Lemma 4.2.3. For Condition 1, suppose $x, y \in V \setminus \{z\}$. If the distance between x and y in $\tilde{C}$ is infinite, then we are done. Otherwise suppose P is a shortest path between x and y in $\tilde{C}$. Since z_1 and z_2 each have only one neighbor and they are not the endpoints of P, they cannot appear in P. Hence P only contains edges and vertices that appear in C, meaning P is a path in C between x and y. So $\tilde{d}(x, y) \geq d(x, y)$. For Condition 2 suppose that $x, y \in V \setminus \{z\}$ with $\mu(x, y) > 0$. Let P be a shortest path in C between x and y. We will show that P is also a path in $\tilde{C}$. Since $\mu(x, y) > 0$ we have $f(x) - f(y) = d(x, y)$. Suppose to the contrary that z is a vertex in P. We need to do two cases. First suppose $f(\cdot) = d(\cdot, z)$. Then we would have

91

$f(x) - f(y) < f(x) - f(z) \leq d(x,z)$. But then we have $d(x,y) < d(x,z)$ contradicting the fact that z is a vertex in the shortest path from x to y. Next assume $f(\cdot) = -d(\cdot, z)$, then we would have $f(x) - f(y) < f(z) - f(y) \leq d(z,y)$. But then we have $d(x,y) < d(z,y)$ again contradicting the fact that $z \in P$. So in fact $z \notin P$, and P only contains vertices that are also vertices of $\tilde{C}$. Because f is Lipschitz and $d(x,y) = f(x) - f(y)$ we get that for every edge $\{s,t\}$ in P, $|f(s) - f(t)| = 1$. This means that $\{s,t\}$ is also an edge of $\tilde{C}$. So P is also a path in $\tilde{C}$. And now for Conditions 3 and 4. Let $x \in V \setminus \{z\}$. $\tilde{C}$ has two connected components, one containing z_1 and the other containing z_2. By construction, if x is in the component containing z_1 then $\tilde{d}(x, z_1) = d(x, z)$ and $\tilde{d}(x, z_2) = \infty$. Likewise, if x is in the component containing z_2 then $\tilde{d}(x, z_1) = \infty$ and $\tilde{d}(x, z_2) = d(x, z)$.

So Lemma 4.2.3 gives us a probability measure $\tilde{\nu}$ satisfying the four properties of the lemma. If we look at the equivalence relation $\sim_{\tilde{\mu}}$ define before Lemma 4.2.4, we obtain the two graphs $\tilde{C}_1 = (\tilde{V}_1, \tilde{E}_1)$ and $\tilde{C}_2 = (\tilde{V}_2, \tilde{E}_2)$, which are both paths on $(n+1)/2$ vertices. Now we may apply Lemmas 4.2.2, 4.2.3, and 4.2.4 to finish the proposition. Note that in making use of Lemmas 4.2.2 and 4.2.3, we have $k = \frac{n}{n+1}$, while in making use of Lemma

4.2.4, we have $k_1 = k_2 = 2$.

$$
\begin{aligned}
W^2(\nu, \pi) &= \left[\frac{n+1}{n} W(\tilde{\nu}, \tilde{\pi})\right]^2 \\
&= \left(\frac{n+1}{n}\right)^2 \left(\frac{1}{2} W(\tilde{\nu}_1, \tilde{\pi}_1) + \frac{1}{2} W(\tilde{\nu}_2, \tilde{\pi}_2)\right)^2 \\
&\leq \left(\frac{n+1}{n}\right)^2 \left(\frac{1}{2} W^2(\tilde{\nu}_1, \tilde{\pi}_1) + \frac{1}{2} W^2(\tilde{\nu}_2, \tilde{\pi}_2)\right) \\
&\leq \left(\frac{n+1}{n}\right)^2 \left(\sigma^2(\tilde{C}_1) D(\tilde{\nu}_1 || \tilde{\pi}_1) + \sigma^2(\tilde{C}_2) D(\tilde{\nu}_2 || \tilde{\pi}_2)\right) \\
&= \left(\frac{n+1}{n}\right)^2 \sigma^2 \left(P_{\frac{n+1}{2}}\right) (D(\tilde{\nu}_1 || \tilde{\pi}_1) + D(\tilde{\nu}_2 || \tilde{\pi}_2)) \\
&= \left(\frac{n+1}{n}\right)^2 \sigma^2 \left(P_{\frac{n+1}{2}}\right) \\
&\qquad \left(\sum_{x \in \tilde{V}_1} \frac{\tilde{\nu}_1(x)}{\tilde{\pi}_1(x)} \log\left(\frac{\tilde{\nu}_1(x)}{\tilde{\pi}_1(x)}\right) \tilde{\pi}_1(x) + \sum_{x \in \tilde{V}_2} \frac{\tilde{\nu}_2(x)}{\tilde{\pi}_2(x)} \log\left(\frac{\tilde{\nu}_2(x)}{\tilde{\pi}_2(x)}\right) \tilde{\pi}_2(x)\right) \\
&= \left(\frac{n+1}{n}\right)^2 \sigma^2 \left(P_{\frac{n+1}{2}}\right) \\
&\qquad \left(\sum_{x \in \tilde{V}_1} \frac{\tilde{\nu}(x)}{\tilde{\pi}(x)} \log\left(\frac{\tilde{\nu}(x)}{\tilde{\pi}(x)}\right) 2\tilde{\pi}(x) + \sum_{x \in \tilde{V}_2} \frac{\tilde{\nu}(x)}{\tilde{\pi}(x)} \log\left(\frac{\tilde{\nu}(x)}{\tilde{\pi}(x)}\right) 2\tilde{\pi}(x)\right) \\
&= 2\left(\frac{n+1}{n}\right)^2 \sigma^2 \left(P_{\frac{n+1}{2}}\right) D(\tilde{\nu} || \tilde{\pi}) \\
&\leq 2\left(\frac{n+1}{n}\right) \sigma^2(P_{\frac{n+1}{2}}) D(\nu || \pi).
\end{aligned}
$$

Hence

$$
\sigma^2(C_n) \leq \left(\frac{n+1}{n}\right) \sigma^2(P_{\frac{n+1}{2}}) = \frac{n^2 + 3n - 1 - \frac{3}{n}}{48} = c^2(C_n)\left(1 + \frac{3n^3 - 3n^2 - 3n + 3}{n^4 + 2n^2 - 3}\right).
$$

$\square$

CHAPTER V

COUPLING AS AN ANALOG OF RICCI CURVATURE

In continuous spaces, lower bounds on Ricci curvature have yielded various interesting results, such as diameter bounds on manifolds, characterization of compactness of manifolds, estimates on spectral gap and log-Sobolev constant etc. Thus it is natural to attempt to capture a notion of Ricci curvature which works well (at least from an applications viewpoint) in discrete settings. As far as we are aware, there seems to be little published work in this direction. In part the difficulty here might be the possibility that there is no single correct definition of a discrete curvature, and instead that such a notion depends on the applications at hand.

We propose here a notion, motivated by a result of Sturm and Renesse [34], which seems interesting in its own right. Whether our definition corresponds to any discrete curvature or not, it seems natural enough as far as measuring the convergence of finite Markov chains to stationarity, in the Wasserstein distance sense, is concerned. This then raises the question of comparing the new quantity with other functional constants such as ρ_0 (which bounds the rate of decay of relative entropy of the chain) and λ_1 (which governs the decay of variance). Proposition 5.2.1 below obtains one such relation. Finally we note that the quantity R in the above definition below is what one tries to estimate, while employing the coupling technique to bound the rate of convergence of an ergodic Markov chain. Once we realized this, we also found that our Proposition 5.2.1 is implicit in the work of Mu-Fa Chen [16] (see also [6] an account of Chen's proof).

In [33], Schmuckenschläger discussed an analog of Ricci curvature for Markov chains on a discrete state space. In [28], Murali considered a slight modification to his definition, proving several results, and finding the curvature (which we refer to as the Schmuckenschläger Ricci curvature) for several examples. While the definitions of Schmuckenschläger and Murali are functional analytic in nature and based on the Γ_2 functional of Bakry and Emery [5] (see

also Ledoux's paper [25]), our definition is based on the coupling of Markov chains using the mass transportation problem. The definition we take is one of the several equivalent forms of a lower bound on the Ricci curvature of a smooth manifold proven by Sturm and Renesse [34].

We start with a graph $G = (V, E)$ with associated measure π and distance function d. Throughout this chapter we assume that d is the graph distance. Next we let L be the generator of a Markov chain on G as defined in Section 1.1.2. Recall that $P(G)$ is the set of probability measures on V. Throughout this chapter, for any $\nu \in P(G)$, ν_t is the Markov chain on G with generator L and initial distribution ν.

Definition 5.0.6. *The Wasserstein Ricci curvature of a Markov generator L on graph G has lower bound R if for all $\nu, \tilde{\nu} \in P(G)$ we have:*

$$W(\nu_t, \tilde{\nu}_t) \leq e^{-Rt} W(\nu_0, \tilde{\nu}_0)$$

for every $t \geq 0$.

The next lemma makes it easier to calculate lower bounds on the Wasserstein Ricci curvature. Let δ^v denote the measure on V defined by:

$$\delta^v(x) = \begin{cases} 1 & x = v \\ 0 & x \neq v \end{cases}$$

Instead of checking the decay of $W(\nu_t, \tilde{\nu}_t)$ for every $\nu, \tilde{\nu} \in P(G)$, we only need to solve the finite problem of calculating the derivative $W(\delta_t^x, \delta_t^y)$ at $t = 0$ for each $x, y \in V$. We apply this lemma several times in examples in Section 5.3.

Lemma 5.0.7. *The Wasserstein Ricci curvature of the Markov generator L on the graph G has lower bound R if and only if*

$$\left. \frac{d^+}{dt} W(\delta_t^x, \delta_t^y) \right|_{t=0} \leq -RW(\delta^x, \delta^y)$$

for every $x, y \in V$.

Proof. First assume the Wasserstein Ricci curvature has lower bound R. Then by definition, for every $x, y \in V$, we have:

$$W(\delta_t^x, \delta_t^y) \leq e^{-Rt} W(\delta^x, \delta^y).$$

So

$$\left.\frac{d}{dt}^+ W(\delta_t^x, \delta_t^y)\right|_{t=0} = \lim_{t\to 0^+} \frac{W(\delta_t^x, \delta_t^y) - W(\delta^x, \delta^y)}{t}$$

$$\leq \lim_{t\to 0^+} \frac{e^{-Rt}W(\delta^x, \delta^y) - W(\delta^x, \delta^y)}{t}$$

$$= -RW(\delta^x, \delta^y)$$

and we are done with the first direction.

Next we assume that

$$\left.\frac{d}{dt}^+ W(\delta_t^x, \delta_t^y)\right|_{t=0} \leq -RW(\delta^x, \delta^y)$$

for every $x, y \in V$. To show that the Wasserstein Ricci curvature is bounded below by R it suffices to show that for $a \geq 0$ we have:

$$\left.\frac{d}{dt}^+ W(\nu_t, \tilde\nu_t)\right|_{t=a} \leq -kW(\nu_a, \tilde\nu_a).$$

Let μ_t be a solution to Monge's problem with respect to ν_t and $\tilde\nu_t$. We first show that

$$\left.\frac{d}{dt}^+ \sum_{x,y\in V} \mu_t(x,y)W(\delta^x, \delta^y)\right|_{t=a} \leq \left.\frac{d}{dt}^+ \sum_{x,y\in V} \mu_a(x,y)W(\delta_t^x, \delta_t^y)\right|_{t=0}$$

which will be the key to the proof. Now

$$\left.\frac{d}{dt}^+ \sum_{x,y\in V} \mu_t(x,y)W(\delta^x, \delta^y)\right|_{t=a} = \lim_{h\to 0^+} \frac{1}{h}\sum_{x,y\in V} \mu_{a+h}(x,y)W(\delta^x, \delta^y) - \frac{1}{h}\sum_{x,y\in V} \mu_a(x,y)W(\delta^x, \delta^y)$$

and

$$\left.\frac{d}{dt}^+ \sum_{x,y\in V} \mu_a(x,y)W(\delta_t^x, \delta_t^y)\right|_{t=0} = \lim_{h\to 0^+} \frac{1}{h}\sum_{x,y\in V} \mu_a(x,y)W(\delta_h^x, \delta_h^y) - \frac{1}{h}\sum_{x,y\in V} \mu_a(x,y)W(\delta^x, \delta^y)$$

So we just need to show:

$$\sum_{x,y\in V} \mu_{a+h}(x,y)W(\delta^x, \delta^y) \leq \sum_{x,y\in V} \mu_a(x,y)W(\delta_h^x, \delta_h^y)$$

for $h > 0$. Let μ_h^{xy} be a solution to Monge's problem with respect to δ_h^x and δ_h^y. First we show that

$$\tilde\mu(v,w) = \sum_{x,y} \mu_a(x,y)\mu_h^{xy}(v,w)$$

96

has first and second marginals ν_{a+h} and $\tilde{\nu}_{a+h}$ respectively.

$$
\begin{aligned}
\sum_{w \in V} \left(\sum_{x,y \in V} \mu_a(x,y) \mu_h^{xy}(v,w) \right) &= \sum_{x,y \in V} \mu_a(x,y) \sum_{w \in V} \mu_h^{xy}(v,w) \\
&= \sum_{x,y \in V} \mu_a(x,y) \delta_h^x(v) \\
&= \sum_{x \in V} \delta_h^x(v) \nu_a(x) \\
&= \sum_{x \in V} \delta^x P_h(v) \nu_a(x) \\
&= \sum_{x \in V} \sum_{y \in V} \delta^x(y) P_h(y,v) \nu_a(x) \\
&= \sum_{x \in V} P_h(x,v) \nu_a(x) \\
&= \nu_a P_h(v) \\
&= \nu_{a+h}(v)
\end{aligned}
$$

The second marginal is similar. So $\tilde{\mu}$ has the same marginals as μ_{a+h}. Hence recalling that $\sum_{x,y \in V} \mu_{a+h}(x,y) W(\delta^x, \delta^y)$ is just $W(\nu_{a+h}, \tilde{\nu}_{a+h})$ and μ_{a+h} is a solution to Monge's problem with respect to ν_{a+h} and $\tilde{\nu}_{a+h}$, while $\tilde{\mu}$ is just feasible in the problem, we get:

$$
\begin{aligned}
\sum_{x,y \in V} \mu_{a+h}(x,y) W(\delta^x, \delta^y) &\leq \sum_{x,y \in V} \tilde{\mu}(x,y) W(\delta^x, \delta^y) \\
&= \sum_{x,y \in V} \sum_{v,w} \mu_a(v,w) \mu_h^{vw}(x,y) W(\delta^x, \delta^y) \\
&= \sum_{v,w \in V} \mu_a(v,w) \sum_{x,y} \mu_h^{vw}(x,y) W(\delta^x, \delta^y) \\
&= \sum_{v,w \in V} \mu_a(v,w) \sum_{x,y} d(x,y) \mu_h^{vw}(x,y) \\
&= \sum_{v,w \in V} \mu_a(v,w) W(\delta_h^v, \delta_h^w)
\end{aligned}
$$

which is the inequality we wanted to prove. Now we can use this inequality to calculate:

$$
\begin{aligned}
\frac{d^+}{dt} W(\nu_t, \tilde{\nu}_t)\Big|_{t=a} &= \frac{d^+}{dt} \sum_{x,y \in V} d(x,y)\mu_t(x,y)\Big|_{t=a} \\
&= \frac{d^+}{dt} \sum_{x,y \in V} \mu_t(x,y)W(\delta^x, \delta^y)\Big|_{t=a} \\
&\leq \frac{d^+}{dt} \sum_{x,y \in V} \mu_a(x,y)W(\delta_t^x, \delta_t^y)\Big|_{t=0} \\
&\leq -R \sum_{x,y \in V} \mu_a(x,y)W(\delta^x, \delta^y) \\
&= -R \sum_{x,y \in V} d(x,y)\mu_a(x,y) \\
&= -RW(\nu_a, \tilde{\nu}_a)
\end{aligned}
$$

□

5.1   Tensorization

Let $\{G_i\}_{i=1}^n$ be a family of graphs with associated measures π_i and graph distance functions d_i. For each $i \in \{1, 2, \ldots n\}$ let L_i be the generator of a Markov chain on graph G_i. Let $G = \prod_{i=1}^n G_i$ be the product graph and L be the generator of the product chain on G as defined in Section 1.1.2.

Proposition 5.1.1. *If R_i is a lower bound for the Wasserstein Ricci curvature of the Markov generator L_i on graph G_i, then*

$$
\frac{1}{n} \min_i R_i
$$

is a lower bound on the Wasserstein Ricci curvature for the Markov generator L on the graph G.

Proof. Let $(v_1, \ldots, v_n), (w_1, \ldots, w_n) \in V$. By Lemma 5.0.7 we only need to show that

$$
\frac{d^+}{dt} W(\delta_t^{v_1 \ldots v_n}, \delta_t^{w_1 \ldots w_n})\Big|_{t=0} \leq \frac{1}{n}(\min_i R_i)W(\delta^{v_1 \ldots v_n}, \delta^{w_1, \ldots, w_n})
$$

Now

$$W(\delta^{v_1...v_n}, \delta^{w_1,...,w_n}) = d((v_1, \ldots, v_n), (w_1, \ldots, w_n))$$
$$= \sum_{i=1}^{n} d_i(v_i, w_i)$$
$$= \sum_{i=1}^{n} W(\delta^{v_i}, \delta^{w_i}).$$

And we know from Lemma 5.0.7 that for each i we have:

$$\left. \frac{d}{dt}^{+} W(\delta_t^{v_i}, \delta_t^{w_i}) \right|_{t=0} \leq -R_i W(\delta^{v_i}, \delta^{w_i}).$$

We will further show that:

$$W(\delta_t^{v_1...v_n}, \delta_t^{w_1...w_n}) = \sum_{i=1}^{n} W\left(\delta_{\frac{1}{n}t}^{v_i}, \delta_{\frac{1}{n}t}^{w_i} \right) + o(t). \tag{61}$$

Putting these together we get the result we want:

$$\left. \frac{d}{dt}^{+} W(\delta_t^{v_1...v_n}, \delta_t^{w_1...w_n}) \right|_{t=0}$$
$$= \lim_{t \to 0^+} \frac{W(\delta_t^{v_1...v_n}, \delta_t^{w_1...w_n}) - W(\delta^{v_1...v_n}, \delta^{w_1...w_n})}{t}$$
$$= \lim_{t \to 0^+} \frac{\sum_{i=1}^{n} W\left(\delta_{\frac{1}{n}t}^{v_i}, \delta_{\frac{1}{n}t}^{w_i} \right) + o(t) - W(\delta^{v_1...v_n}, \delta^{w_1...w_n})}{t}$$
$$= \lim_{t \to 0^+} \sum_{i=1}^{n} \frac{W\left(\delta_{\frac{1}{n}t}^{v_i}, \delta_{\frac{1}{n}t}^{w_i} \right) - W(\delta^{v_i}, \delta^{w_i})}{t} + \frac{o(t)}{t}$$
$$= \frac{1}{n} \sum_{i=1}^{n} \left. \frac{d}{dt}^{+} W(\delta_t^{v_i}, \delta_t^{w_i}) \right|_{t=0}$$
$$\leq -\frac{1}{n} \sum_{i=1}^{n} R_i W(\delta^{v_i}, \delta^{w_i})$$
$$\leq -\frac{1}{n}(\min_i R_i) W(\delta^{v_1...v_n}, \delta^{w_1...w_n}).$$

So all we need to show is Equation 61. Now

$$\frac{d}{dt}^{+} \delta_t^{v_1\ldots v_n}(x_1,\ldots,x_n)\bigg|_{t=0}$$

$$= \delta^{v_1\ldots v_n} L(x_1,\ldots,x_n)$$

$$= \sum_{(y_1,\ldots,y_n)} \delta^{v_1\ldots v_n}(y_1,\ldots,y_n)L((y_1,\ldots,y_n),(x_1,\ldots,x_n))$$

$$= L((v_1,\ldots,v_n),(x_1,\ldots,x_n))$$

$$= \frac{1}{n}\sum_{i=1}^{n} L_i(v_i,x_i)\prod_{\substack{j=1\\j\neq i}}^{n}\delta^{v_j}(x_j)$$

and

$$\frac{d}{dt}^{+}\delta_t^{v_i}(x_i)\bigg|_{t=0}$$

$$= \delta^{v_i}L(x_i)$$

$$= \sum_{y_i\in V_i}\delta^{v_i}(y_i)L_i(y_i,x_i)$$

$$= L_i(v_i,x_i).$$

Also, for any ν_t we have:

$$\nu_t(x) = \nu_0(x) + t\,\frac{d}{dt}^{+}\nu_t(x)\bigg|_{t=0} + o(t)$$

So

$$\delta_t^{v_1\ldots v_n}(x_1,\ldots,x_n)$$

$$= \delta^{v_1\ldots v_n}(x_1,\ldots,x_n) + t\frac{1}{n}\sum_{i=1}^{n}L_i(v_i,x_i)\prod_{\substack{j=1\\j\neq i}}^{n}\delta^{v_j}(x_j) + o(t)$$

and

$$\delta_t^{v_i}(x_i) = \delta^{v_i}(x_i) + tL_i(v_i,x_i) + o(t).$$

Then

$$\prod_{i=1}^{n}\delta_t^{v_i}(x_i) = \prod_{i=1}^{n}(\delta^{v_i}(x_i) + tL_i(v_i,x_i) + o(t))$$

$$= \prod_{i=1}^{n}\delta^{v_i}(x_i) + \sum_{i=1}^{n}tL_i(v_i,x_i)\prod_{\substack{j=1\\j\neq i}}^{n}\delta^{v_j}(x_j) + o(t)$$

$$= \delta_{nt}^{v_1\ldots v_n}(x_1,\ldots,x_n) + o(t).$$

Hence by Lemmas 2.2.2 and 2.2.6 we get:

$$W(\delta_t^{v_1 \ldots v_n}, \delta_t^{w_1 \ldots w_n}) \;=\; W\left(\prod_{i=1}^{n} \delta_{\frac{1}{n}t}^{v_i}, \prod_{i=1}^{n} \delta_{\frac{1}{n}t}^{w_i}\right) + o(t)$$

$$=\; \sum_{i=1}^{n} W\left(\delta_{\frac{1}{n}t}^{v_i}, \delta_{\frac{1}{n}t}^{w_i}\right) + o(t)$$

which is what we wanted to show. $\qquad\square$

5.2   Lower Bound on Spectral Gap

Proposition 5.2.1. *Suppose R is a lower bound on the Wasserstein Ricci curvature of a Markov chain with generator L on graph G. If $\lambda_1(L)$ is the spectral gap of the chain, then $R \le \lambda_1(L)$.*

Proof. Let f be an eigenfunction of L with eigenvalue $-\lambda_1(L)$, so that $f_t(x) = e^{-\lambda_1(L)t} f(x)$. Note that f_t is also an eigenfunction of $-L$ with eigenvalue $-\lambda_1(L)$ for each $t \ge 0$. Now

$$\sum_{x \in V} f_t(x)\pi(x) = \frac{1}{\lambda_1(L)} \sum_{x \in V} L f_t(x)\pi(x) = 0,$$

so $(1 + f_t)d\pi$ is a probability measure for all $t \ge a$ for some time $a \ge 0$. If g is a solution to Kantorovich's problem with respect to $(1 + f_a)d\pi$ and π at some time a, then g is also a solution to Kantorovich's problem with respect to $(1 + f_t)d\pi$ and π for $t \ge a$. Hence

$$e^{-\lambda_1(L)h} W((1 + f_a)d\pi, \pi) = W((1 + f_{a+h})d\pi, \pi) \le e^{-Rh} W((1 + f_a)d\pi, \pi)$$

for each $h \ge 0$. This gives us $R \le \lambda_1(L)$. $\qquad\square$

5.3   Examples and Comparison with Schmuckenschläger

Now we calculate a lower bound on the Wasserstein Ricci curvature of several example graphs. At the end of this section we give a table comparing the Wasserstein Ricci curvature, the Schmuckenschläger Ricci curvature (as given in [28]), and the spectral gap for each of the examples.

We start with the general two point graph with generator:

$$L = \begin{bmatrix} -s & s \\ t & -t \end{bmatrix}.$$

We will label the vertices 1 and 2 to correspond with the rows (or columns) of the matrix L. Then the stationary distribution of the chain is $\pi(1) = \frac{t}{s+t}$ and $\pi(2) = \frac{s}{s+t}$. Now the function $g(1) = 1$ and $g(0) = 0$ is a solution to Kantorovich's problem with respect to δ_t^1 and δ_t^2 for any t close to zero and the function $g(1) = 0$ and $g(0) = 1$ is a solution to Kantorovich's problem with respect to δ_t^2 and δ_t^1 for any t close to zero. Hence

$$
\begin{aligned}
\frac{d}{dt}^+ W(\delta_t^1, \delta_t^2)\bigg|_{t=0} &= g(1)(\delta^1 L(1) - \delta^2 L(1)) + g(2)(\delta^1 L(2) - \delta^2 L(2)) \\
&= L(1,1) - L(2,1) \\
&= -(s+t) \\
&= -(s+t)W(\delta^1, \delta^2)
\end{aligned}
$$

We similarly get that $\frac{d}{dt}^+ W(\delta_t^2, \delta_t^1)\big|_{t=0} = -(s+t)W(\delta^2, \delta^1)$. Hence the Wasserstein Ricci curvature of the two point space is bounded below by $s + t$.

Next we look at the complete graph on the vertex set $[n] = \{1, 2, \ldots, n\}$ with arbitrary stationary distribution π and generator $L(i, j) = \pi(j)$ for each $i, j \in [n]$ with $i \neq j$. By Lemma 2.2.4, there exists a solution g to Kantorovich's problem with respect to δ^i and δ^j for which

$$
\frac{d}{dt}^+ W(\delta_t^i, \delta_t^j)\bigg|_{t=0} = \sum_{k \in [n]} g(k)(\delta^i L(k) - \delta^j L(k))
$$

Now if g is any solution to Kantorovich's problem with respect to δ^i and δ^j we have:

$$
\begin{aligned}
&\sum_{k \in [n]} g(k)(\delta^i L(k) - \delta^j L(k)) \\
&= \sum_{k \in [n]} g(k)(L(i, k) - L(j, k)) \\
&= g(i)(L(i, i) - L(j, i)) + g(j)(L(i, j) - L(j, j)) \\
&\quad + \sum_{\substack{k \in [n] \\ k \notin \{i,j\}}} g(k)(\pi(k) - \pi(k))) \\
&= g(i)(-\sum_{\substack{k \in [n] \\ k \neq i}} \pi(k) - \pi(i)) + g(j)(\pi(j) + \sum_{\substack{k \in [n] \\ k \neq j}} \pi(k)) \\
&= g(j) - g(i) \\
&= -W(\delta^i, \delta^j).
\end{aligned}
$$

So the Wasserstein Ricci curvature of this chain is bounded below by 1.

Now consider the path on n vertices, where we again take the vertex set to be $[n]$. We consider the chain which moves down (if not already at 1) with rate p and moves up (if not already at n) with rate $(1-p)$. So the chain has generator:

$$
L = \begin{bmatrix}
p-1 & 1-p & 0 & 0 & \cdots & 0 & 0 \\
p & -1 & 1-p & 0 & \cdots & 0 & 0 \\
0 & p & -1 & 1-p & \ddots & \vdots & \vdots \\
0 & 0 & p & -1 & \ddots & 0 & 0 \\
\vdots & \vdots & \ddots & \ddots & \ddots & 1-p & 0 \\
0 & 0 & \cdots & 0 & p & -1 & 1-p \\
0 & 0 & \cdots & 0 & 0 & p & -p
\end{bmatrix}
$$

For any $i,j \in [n]$ with $i > j$ we have $g(i) = i$ is a solution to Kantorovich's problem with respect to δ_t^i and δ_t^j for small enough t. Assuming $i \neq n$ and $j \neq 1$ we have:

$$
\begin{aligned}
\frac{d^+}{dt} W(\delta_t^i, \delta_t^j)\Big|_{t=0} &= \sum_{k \in [n]} g(k)(\delta^i L(k) - \delta^j L(k)) \\
&= \sum_{k \in [n]} k(L(i,k) - L(j,k)) \\
&= (i-1)L(i,i-1) + iL(i,i) + (i+1)L(i,i+1) \\
&\quad - (j-1)L(j,j-1) - jL(j,j) - (j+1)L(j,j+1) \\
&= (i-1)p + i(-1) + (i+1)(1-p) \\
&\quad - (j-1)p - j(-1) - (j+1)(1-p) \\
&= 0.
\end{aligned}
$$

For $j \neq 1$ we have:

$$
\frac{d^+}{dt} W(\delta_t^n, \delta_t^j)\Big|_{t=0} = -(1-p).
$$

For $i \neq n$ we have:

$$
\frac{d^+}{dt} W(\delta_t^i, \delta_t^1)\Big|_{t=0} = -p.
$$

If $i = n$ and $j = 1$ we have:

$$
\frac{d^+}{dt} W(\delta_t^n, \delta_t^1)\Big|_{t=0} = -1.
$$

Table 1: Comparison of Curvature and Spectral Gap

Markov Chain	Wasserstein Ricci	Schmuckenschläger Ricci	Spectral gap
Two Point	$s + t$	$\min((3s + t)/2, (3t + s)/2))$	s+t
Complete	1	$1/2 + \min_{y \in V} \pi(y)$	1
Path $(n = 3)$	$\min(p/2, (1 - p)/2)$	> 0	> 0
Path $(n = 4)$	0	> 0	> 0
Path $(n \geq 5)$	0	0	> 0

So if $n \geq 4$, we have Wasserstein Ricci curvature bounded below by 0. If $n = 3$, then we have Wasserstein Ricci curvature bounded below by $\min\{p/2, (1 - p)/2\}$. If $n = 2$, then we have Wasserstein Ricci curvature bounded below by 1.

CHAPTER VI

CONCLUSION AND FUTURE DIRECTIONS

We have looked at three different aspects of the mass transportation problem in the study of discrete concentration inequalities.

First we examined the transportation and variance transportation inequalities and their relationship with the Poincaré and modified log-Sobolev inequalities. Here we saw that using the graph distance, the subgaussian constant and the spread constant provide upper bounds on the modified log-Sobolev constant and the spectral gap respectively for large classes of Markov chains on a graph, while tight bounds could sometimes be obtained by tailoring the distance function to the specific Markov chain of interest. And in the other direction, we examined the bounds on the spread constant provided by the spectral gap of a fastest mixing Markov chain on the graph.

Next we looked at the specific problem of bounding the subgaussian constant of a cycle, in order to obtain a concentration result on the discrete torus. We used the interplay between the dual formulations of the subgaussian constant to obtain the exact value for even cycles. For odd cycles we used geometric insight provided by the mass transportation formulation of the subgaussian constant to find the asymptotically correct value.

Finally we explored a notion of discrete Ricci curvature given in terms of the mass transportation problem. As it is closely related to the coupling method of bounding mixing time for Markov chains, it provides a lower bound on the spectral gap of a Markov chain. The constant we defined tensorizes properly and is relatively easy to compute on example graphs.

This work has led us to several intriguing questions and directions for future study. We begin with Chapter 3, where we saw that essentially the same proof worked to show both that the modified log-Sobolev inequality implies the transportation inequality and that the Poincaré inequality implies the variance transportation inequality. In [10], Bobkov and

Tetali define the inequality:

$$\alpha(p) \left[\|f\|_p^p - \|f\|_1^p\right] \leq \frac{p}{2}\mathcal{E}(f, f^{p-1}) \tag{62}$$

for $p \in (0, 1]$, which interpolates between the Poincaré inequality and the modified log-Sobolev inequality. This leads naturally to the definition of an inequality that interpolates between the variance transportation inequality and the transportation inequality:

$$W^2(\nu, \pi) \leq \beta(p)\frac{p}{2(p-1)}\mathcal{E}(f, f^{p-1}). \tag{63}$$

One question is whether we can modify the proof technique that works to show that the endpoints of (62) respectively imply the endpoints of (63), to get that (62) implies (63) for each $p \in (1, 2]$. The other obvious question is whether the interpolated transportation inequality is useful for anything.

Another direction we would like to take for further study based on Chapter 3 is to look at the fastest mixing Markov chain problem in terms of the modified log-Sobolev constant. We would like to show a tightness result for the inequality $\rho_0^*(G) \leq \frac{1}{2\sigma^2(G)}$ where $\rho_0^*(G)$ is the maximum of $\rho_0(L)$ over all Markov generators L on G that satisfy some normalization condition, and $\sigma^2(G)$ is calculated using the graph distance.

In Chapter 4, we tried to understand and classify the solutions to Kantorovich's problem with respect to ν and π when ν satisfies the transportation inequality with equality. An interesting question is whether this ν that attains equality has any physical interpretation or contains any useful information. Does it hold any information about the optimal sets in the isoperimetric problem for example?

A more concrete direction of study from Chapter 4 is to better classify those graphs for which the subgaussian constant and the spread constant are equal. In some sense, they will be different for most graphs, as a typical random graph is an expander, for which $c^2 \ll \sigma^2$ as we saw in Section 3.1.6. From Lemma 4.1.1, it appears that some type of symmetry in the graph may be necessary in order for the subgaussian constant to be equal to the spread constant.

The notion of curvature in Chapter 5 opens the door to a seemingly infinite number of questions. A key test for this notion of Ricci curvature is whether or not it implies

normal concentration with constant R (where R is a lower bound on the Wasserstein Ricci curvature). Another good question is whether or not it can be used to prove an analog of Buser's inequality [14], or perhaps inverse diameter squared lower bounds on the spectral gap under a non-negative Wasserstein Ricci curvature assumption. And we would like to know if an assumption on the Wasserstein Ricci curvature of a Markov generator on a graph will allow proofs of the convexity of the relative entropy of the distribution of a Markov chain on the graph. It would be very exciting if we could prove an analog of the HWI inequality of [30] using this notion of curvature.

We have had some success with the transportation inequality and the l_1 Wasserstein distance, and as we have seen there are many possibly fruitful future directions of study involving these ideas. But we also remain hopeful that there is an analog of the L^2 Wasserstein distance from the Euclidean setting that will provide even more powerful functionality.

REFERENCES

[1] ALDOUS, D. and FILL, J., "Reversible Markov chains and random walks on graphs." Monograph in preparation, available at http://www.stat.berkeley.edu/users/aldous/RWG/book.html.

[2] ALON, N., "Eigenvalues and expanders," *Combinatorica*, vol. 6, no. 2, pp. 83–96, 1986. Theory of computing (Singer Island, Fla., 1984).

[3] ALON, N., BOPPANA, R., and SPENCER, J., "An asymptotic isoperimetric inequality," *Geom. Funct. Anal.*, vol. 8, no. 3, pp. 411–436, 1998.

[4] ALON, N. and MILMAN, V. D., "λ_1, isoperimetric inequalities for graphs, and superconcentrators," *J. Combin. Theory Ser. B*, vol. 38, no. 1, pp. 73–88, 1985.

[5] BAKRY, D. and ÉMERY, M., "Diffusions hypercontractives," in *Séminaire de probabilités, XIX, 1983/84*, vol. 1123 of *Lecture Notes in Math.*, pp. 177–206, Berlin: Springer, 1985.

[6] BERGER, N., KENYON, C., MOSSEL, E., and PERES, Y., "Glauber Dynamics on Trees and Hyperbolic Graphs." Preprint available at http://www.stat.berkeley.edu/~peres/.

[7] BOBKOV, S. G. and GÖTZE, F., "Exponential integrability and transportation cost related to logarithmic Sobolev inequalities," *J. Funct. Anal.*, vol. 163, no. 1, pp. 1–28, 1999.

[8] BOBKOV, S. G., HOUDRÉ, C., and TETALI, P., "The subgaussian constant and concentration inequalities." Preprint available at http://www.math.gatech.edu/~tetali.

[9] BOBKOV, S. G., GENTIL, I., and LEDOUX, M., "Hypercontractivity of Hamilton-Jacobi equations," *J. Math. Pures Appl. (9)*, vol. 80, no. 7, pp. 669–696, 2001.

[10] BOBKOV, S. G. and TETALI, P., "Modified log-Sobolev inequalities in discrete settings," *Proc. of the ACM STOC*, pp. 287–296, 2003. Preliminary version, full version available at http://www.math.gatech.edu/~tetali.

[11] BORGS, C., CHAYES, J. T., FRIEZE, A., KIM, J. H., TETALI, P., VIGODA, E., and VU, V. H., "Torpid mixing of some Monte Carlo Markov chain algorithms in statistical physics," in *40th Annual Symposium on Foundations of Computer Science (New York, 1999)*, pp. 218–229, IEEE Computer Soc., Los Alamitos, CA, 1999.

[12] BOYD, S., DIACONIS, P., and XIAO, L., "Fastest mixing Markov chain on a graph," *SIAM REVIEW*, vol. 46, no. 4, pp. 667–689, 2004.

[13] BRENIER, Y., "Décomposition polaire et réarrangement monotone des champs de vecteurs," *C. R. Acad. Sci. Paris Sér. I Math.*, vol. 305, no. 19, pp. 805–808, 1987.

[14] BUSER, P., "A note on the isoperimetric constant," *Ann. Sci. École Norm. Sup. (4)*, vol. 15, no. 2, pp. 213–230, 1982.

[15] CHEN, G.-Y. and SHEU, Y.-C., "On the log-Sobolev constant for the simple random walk on the n-cycle: the even cases," *J. Funct. Anal.*, vol. 202, no. 2, pp. 473–485, 2003.

[16] CHEN, M.-F., "Trilogy of couplings and general formulas for lower bound of spectral gap," in *Probability towards 2000 (New York, 1995)*, vol. 128 of *Lecture Notes in Statist.*, pp. 123–136, New York: Springer, 1998.

[17] DIACONIS, P. and SALOFF-COSTE, L., "Logarithmic Sobolev inequalities for finite Markov chains," *Ann. Appl. Probab.*, vol. 6, no. 3, pp. 695–750, 1996.

[18] FELDMAN, M. and MCCANN, R. J., "Monge's transport problem on a Riemannian manifold," *Trans. Amer. Math. Soc.*, vol. 354, no. 4, pp. 1667–1697 (electronic), 2002.

[19] GANGBO, W. and MCCANN, R. J., "The geometry of optimal transportation," *Acta Math.*, vol. 177, no. 2, pp. 113–161, 1996.

[20] GAO, F. and QUASTEL, J., "Exponential decay of entropy in the random transposition and Bernoulli-Laplace models," *Ann. Appl. Probab.*, vol. 13, no. 4, pp. 1591–1600, 2003.

[21] GRIMMETT, G. R. and STIRZAKER, D. R., *Probability and random processes.* New York: Oxford University Press, third ed., 2001.

[22] GROMOV, M. and MILMAN, V. D., "Generalization of the spherical isoperimetric inequality to uniformly convex Banach spaces," *Compositio Math.*, vol. 62, no. 3, pp. 263–282, 1987.

[23] JOHNSON, W. B. and LINDENSTRAUSS, J., "Extensions of Lipschitz mappings into a Hilbert space," in *Conference in modern analysis and probability (New Haven, Conn., 1982)*, vol. 26 of *Contemp. Math.*, pp. 189–206, Providence, RI: Amer. Math. Soc., 1984.

[24] KANTOROVICH, L. V., "On the translocation of masses," *Dokl. Akad. Nauk SSSR*, vol. 37, no. 7–8, pp. 227–229, 1942.

[25] LEDOUX, M., "On an integral criterion for hypercontractivity of diffusion semigroups and extremal functions," *J. Funct. Anal.*, vol. 105, no. 2, pp. 444–465, 1992.

[26] LEDOUX, M., *The Concentration of Measure Phenomenon*, vol. 89 of *Mathematical Surveys and Monographs.* Providence, RI: American Mathematical Society, 2001.

[27] MONGE, G., "Mémoire sur la théorie des déblais et de remblais," *Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année*, pp. 666–704, 1781.

[28] MURALI, S., *Curvature, Isoperimetry, and Discrete Spin Systems.* PhD thesis, Georgia Institute of Technology, Atlanta, Georgia, 2001.

[29] NAOR, A., SAMMER, M., and TETALI, P., "On approximating the maximal variance problem." In preparation.

[30] OTTO, F. and VILLANI, C., "Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality," *J. Funct. Anal.*, vol. 173, no. 2, pp. 361–400, 2000.

[31] RIORDAN, O., "An ordering on the even discrete torus," *SIAM J. Discrete Math.*, vol. 11, no. 1, pp. 110–127 (electronic), 1998.

[32] SALAS, S., HILLE, E., and ETGEN, G., *Salas and Hille's Calculus: One and Several Variables.* John Wiley & Sons, 8th ed., 1998.

[33] SCHMUCKENSCHLÄGER, M., "Curvature of nonlocal Markov generators," in *Convex geometric analysis (Berkeley, CA, 1996)*, vol. 34 of *Math. Sci. Res. Inst. Publ.*, pp. 189–197, Cambridge: Cambridge Univ. Press, 1999.

[34] STURM, K.-T. and VON RENESSE, M.-K., "Transport inequalities, gradient estimates, entropy and Ricci curvature," *Comm. Pure Appl. Math.*, 2004. to appear.

[35] SUN, J., BOYD, S., XIAO, L., and DIACONIS, P., "The fastest mixing Markov process on a graph and a connection to a maximum variance unfolding problem." Preprint available at http://www.stanford.edu/∼boyd/fmmp.html.

[36] TALAGRAND, M., "Transportation cost for Gaussian and other product measures," *Geom. Funct. Anal.*, vol. 6, no. 3, pp. 587–600, 1996.

[37] VERSHIK, A. M., "The Kantorovich metric: the initial history and little-known applications," *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, vol. 312, no. Teor. Predst. Din. Sist. Komb. i Algoritm. Metody. 11, pp. 69–85, 311, 2004.

[38] VILLANI, C., *Topics in Optimal Transportation*, vol. 58 of *Graduate Studies in Mathematics.* Providence, RI: American Mathematical Society, 2003.