Quantitative Convergence of Quadratically Regularized Linear Programs*

Alberto González-Sanz† Marcel Nutz‡

April 22, 2025

Abstract

Linear programs with quadratic ("ridge") regularization are of recent interest in optimal transport: unlike entropic regularization, the squared-norm penalty gives rise to sparse approximations of optimal transport couplings. More broadly, quadratic regularization is used in overparametrized learning problems to single out a particular solution. It is well known that the solution of a quadratically regularized linear program over any polytope converges stationarily to the minimal-norm solution of the linear program when the regularization parameter tends to zero. However, that result is merely qualitative. Our main result quantifies the convergence by specifying the exact threshold for the regularization parameter, after which the regularized solution also solves the linear program. Moreover, we bound the suboptimality of the regularized solution before the threshold. These results are complemented by a convergence rate for the regime of large regularization. We apply our general results to the setting of optimal transport, where we shed light on how the threshold and suboptimality depend on the number of data points.

Keywords Linear Program, Quadratic Regularization, Optimal Transport
AMS 2020 Subject Classification 49N10; 49N05; 90C25

1 Introduction

Let $\mathbf{c} \in \mathbb{R}^d$ and let $\mathcal{P} \subset \mathbb{R}^d$ be a polytope. Moreover, let $\langle \cdot, \cdot \rangle$ be an inner product on $\mathbb{R}^d$ and $\| \cdot \|$ its induced norm. We study the linear program

$$\text{minimize } \langle \mathbf{c}, \mathbf{x} \rangle \qquad \text{subject to } \mathbf{x} \in \mathcal{P} \tag{LP}$$

and its quadratically regularized counterpart,

$$\text{minimize } \langle \mathbf{c}, \mathbf{x} \rangle + \frac{\|\mathbf{x}\|^2}{\eta} \qquad \text{subject to } \mathbf{x} \in \mathcal{P}. \tag{QLP}$$

*The authors thank Roberto Cominetti, Andrés Riveros Valdevenito and two anonymous referees for helpful comments.

†Columbia University, Department of Statistics, ag4855@columbia.edu.

‡Columbia University, Departments of Statistics and Mathematics, mnutz@columbia.edu. Research supported by NSF Grants DMS-1812661, DMS-2106056, DMS-2407074.

Here $\eta \in (0, \infty)$ is called the inverse regularization parameter (whereas $1/\eta$ is the regularization). In the limit $\eta \to \infty$ of small regularization, (QLP) converges to (LP). More precisely, the unique solution $\mathbf{x}^\eta$ of (QLP) converges to a particular solution $\mathbf{x}^*$ of (LP), namely the solution with smallest norm: $\mathbf{x}^* = \arg\min_{\mathbf{x} \in \mathcal{M}} \|\mathbf{x}\|^2$, where $\mathcal{M}$ denotes the set of minimizers of (LP). Our main goal is to describe how quickly this convergence happens.

Linear programming and regularization are fundamental tools in data science. Many statistical methodologies, including for instance quantile regression [32], statistical depths [31] or multivariate quantiles [30], rely on solving a linear program. Regularization by a quadratic penalty—also called ridge penalty due to its prominent application in ridge regression—is used in many statistical problems (e.g., regularized quantile regression [36]) but also in data science more broadly, for instance in overparametrized learning problems where the aim is to single out a particular solution [5, 8, 54].

The aforementioned convergence of the solution $\mathbf{x}^\eta$ of (QLP) to the minimum-norm solution $\mathbf{x}^*$ of (LP) is stationary: there exists a threshold η^* such that $\mathbf{x}^\eta = \mathbf{x}^*$ for all $\eta \geq \eta^*$. This was first established for linear programs in [40, Theorem 1] and [39, Theorem 2.1], and was more recently rediscovered in the context of optimal transport [18, Property 5]. However, those results are qualitative: they do not give a value or a bound for η^*. We shall characterize the exact value of the threshold η^* (cf. Theorem 2.5), and show how this leads to computable bounds in applications. This exact result raises the question about the speed of convergence as $\eta \uparrow \eta^*$. Specifically, we are interested in the convergence of the error $\mathcal{E}(\eta) = \langle \mathbf{c}, \mathbf{x}^\eta \rangle - \min_{\mathbf{x} \in \mathcal{P}} \langle \mathbf{c}, \mathbf{x} \rangle$ measuring how suboptimal the solution $\mathbf{x}^\eta$ of (QLP) is when plugged into (LP). In Theorem 2.5, we show that $\mathcal{E}(\eta) = o(\eta^* - \eta)$ as $\eta \uparrow \eta^*$ and give an explicit bound for $\mathcal{E}(\eta)/(\eta^* - \eta)$. After observing that the curve $\eta \mapsto \mathbf{x}^\eta$ is piecewise affine, this linear rate can be understood as the slope of the last segment of the curve before ending at $\mathbf{x}^*$. Figure 1 illustrates these quantities in a simple example. Our results for $\eta \to \infty$ are complemented by a convergence rate for the large regularization regime $\eta \to 0$ where $\mathbf{x}^\eta$ tends to $\arg\min_{\mathbf{x} \in \mathcal{P}} \|\mathbf{x}\|^2$; cf. Proposition 2.7.

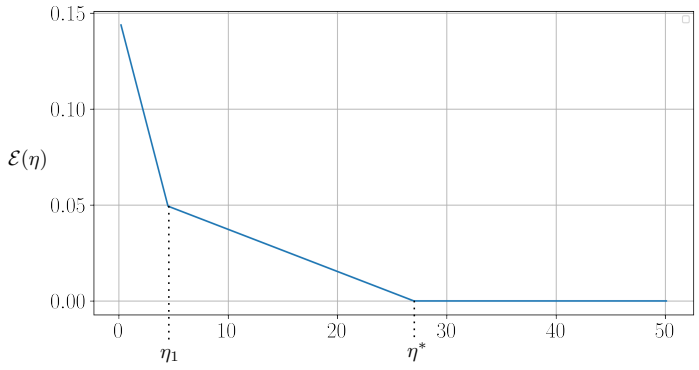

Figure 1: Suboptimality $\mathcal{E}(\eta)$ of (QOT) when $\mu = \nu = \frac{1}{3}\sum_{i=1}^{3} \delta_{i/3}$ and $c(x, y) = \|x - y\|^2$. Theorem 2.5 characterizes the location of η^* and bounds the slope to the left of η^*.

While linear programs and their penalized counterparts go back far into the last century, our interest is fueled by the surge of optimal transport in applications such as machine learning (e.g., [33]), statistics (e.g., [45]), language and image processing (e.g., [3, 47]) and economics (e.g., [24]).

In its simplest form, the optimal transport problem between probability measures μ and ν is

$$\inf_{\gamma \in \Gamma(\mu,\nu)} \int c(x,y) d\gamma(x,y), \tag{OT}$$

where $\Gamma(\mu,\nu)$ denotes the set of couplings; i.e., probability measures γ with marginals μ and ν (see [49, 50] for an in-depth exposition). Here $c(\cdot,\cdot)$ is a given cost function, most commonly $c(x,y) = \|x-y\|^2$. In many applications the marginals represent observed data: data points $\mathbf{X}_1, \dots, \mathbf{X}_N$ and $\mathbf{Y}_1, \dots, \mathbf{Y}_N$ are encoded in their empirical measures $\mu = \frac{1}{N}\sum_i \delta_{\mathbf{X}_i}$ and $\nu = \frac{1}{N}\sum_i \delta_{\mathbf{Y}_i}$. Writing also $\mathbf{c}_{ij} = c(\mathbf{X}_i, \mathbf{Y}_j)$, the problem (OT) is a particular case of (LP) in dimension $d = N \times N$. The general linear program (LP) also includes other transport problems of recent interest, such as multi-marginal optimal transport and Wasserstein barycenters [1], adapted Wasserstein distances [4] or martingale optimal transport [7].

As the optimal transport problem is computationally costly (e.g., [46]), [17] proposed to regularize (OT) by penalizing with Kullback–Leibler divergence (entropy). Then, solutions can be computed using the Sinkhorn–Knopp (or IPFP) algorithm, which has lead to an explosion of high-dimensional applications. Entropic regularization always leads to "dense" solutions (couplings whose support contains all data pairs $(\mathbf{X}_i, \mathbf{Y}_j)$) even though the unregularized problem (OT) typically has a sparse solution. In some applications that is undesirable; for instance, it may correspond to blurrier images in an image processing task [10]. For that reason, [10] suggested the quadratic penalization

$$\inf_{\gamma \in \Gamma(\mu,\nu)} \int c(x,y) d\gamma(x,y) + \frac{1}{\eta} \left\| \frac{d\gamma}{d(\mu \otimes \nu)} \right\|^2_{L^2(\mu\otimes\nu)} \tag{QOT}$$

where $d\gamma/d(\mu\otimes\nu)$ denotes the density of γ with respect to the product measure $\mu\otimes\nu$. See also [22] for a similar formulation of minimum-cost flow problems, the predecessors referenced therein, and [18] for optimal transport with more general convex regularization. Quadratic regularization gives rise to sparse solutions (see [10], and [26, 27, 42, 52] for recent theoretical results). Applications of quadratically regularized optimal transport include manifold learning [53] and image processing [35] while [41] establishes a connection to maximum likelihood estimation of Gaussian mixtures. Computational approaches are developed in [20, 25, 28, 35, 48] whereas [38, 19, 6, 42] study theoretical aspects with a focus on continuous problems. In that context, [37, 21] show Gamma convergence to the unregularized optimal transport problem in the small regularization limit. Those results are straightforward in the discrete case considered in the present work. Conversely, the stationary convergence studied here does not take place in the continuous case.

For linear programs with entropic regularization, [15] established that solutions converge exponentially to the limiting unregularized counterpart. More recently, [51] gave an explicit bound for the convergence rate. The picture for entropic regularization is quite different to quadratic regularization as the convergence is not stationary. For instance, in optimal transport, the support of the regularized solution contains all data pairs for any value of the regularization parameter, collapsing only at the unregularized limit. Nevertheless, our analysis benefits from some of the technical ideas in [51], specifically for the proof of the slope bound (3). The small regularization limit has also attracted a lot of attention in continuous optimal transport (e.g., [2, 9, 14, 16, 34, 43, 44]) which however is technically less related to the present work.

The remainder of this note is organized as follows. Section 2 contains the main results on the general linear program and its quadratic regularization, Section 3 the application to optimal transport. Proofs are gathered in Section 4.

2 Main Results

Throughout, $\emptyset \neq \mathcal{P} \subset \mathbb{R}^d$ denotes a polytope. That is, $\mathcal{P}$ is the convex hull of its extreme points (or vertices) $\exp(\mathcal{P}) = \{\mathbf{v}_1, \ldots, \mathbf{v}_K\}$, which are in turn minimal with the property of spanning $\mathcal{P}$ (see [12] for detailed definitions). We recall the linear program (LP) and its quadratically penalized version (QLP) as defined in the Introduction, and in particular their cost vector $\mathbf{c} \in \mathbb{R}^d$. The set of minimizers of (LP) is denoted

$$\mathcal{M} = \mathcal{M}(\mathcal{P}, \mathbf{c}) = \arg\min_{\mathbf{x} \in \mathcal{P}} \langle \mathbf{c}, \mathbf{x} \rangle;$$

it is again a polytope. To avoid a degenerate problem, we assume throughout that the projection of the origin onto $\mathcal{P}$ is not a minimizer of (LP). (If it is a minimizer of (LP), then it is also the minimizer of (QLP) for any η, so that our problem is trivial.) We abbreviate the objective function of (QLP) as

$$\Phi_\eta(\mathbf{x}) = \langle \mathbf{c}, \mathbf{x} \rangle + \frac{\|\mathbf{x}\|^2}{\eta}.$$

In view of $\Phi_\eta(\mathbf{x}) = \frac{1}{\eta} \left\|\mathbf{x} + \frac{\eta \mathbf{c}}{2}\right\|^2 - \frac{\eta}{4}\|\mathbf{c}\|^2$, minimizing $\Phi_\eta(\mathbf{x})$ over $\mathcal{P}$ is equivalent to projecting $-\eta\mathbf{c}/2$ onto $\mathcal{P}$ in the Hilbert space $(\mathbb{R}^d, \langle \cdot, \cdot \rangle)$. The projection theorem (e.g., [11, Theorem 5.2]) thus implies the following result. We denote by $\mathrm{ri}(C)$ the relative interior of a set $C \subset \mathbb{R}^d$; i.e, the topological interior when C is considered as a subset of its affine hull.

Lemma 2.1. *Given $\eta > 0$,* (QLP) *admits a unique minimizer $\mathbf{x}^\eta$. It is characterized as the unique $\mathbf{x}^\eta \in \mathcal{P}$ such that*

$$\left\langle -\frac{\eta \mathbf{c}}{2} - \mathbf{x}^\eta, \mathbf{x} - \mathbf{x}^\eta \right\rangle \leq 0 \quad \textit{for all } \mathbf{x} \in \mathcal{P}.$$

In particular, if $\mathbf{x}^\eta \in \mathrm{ri}(C)$ for some convex set $C \subset \mathcal{P}$, then also

$$\left\langle -\frac{\eta \mathbf{c}}{2} - \mathbf{x}^\eta, \mathbf{x} - \mathbf{x}^\eta \right\rangle = 0 \quad \textit{for all } \mathbf{x} \in C.$$

Figure 2 illustrates how $\mathbf{x}^\eta$ is obtained as the projection of $-\eta\mathbf{c}/2$. The algorithm of [29] solves the problem of projecting a point onto a polyhedron, hence can be used to find $\mathbf{x}^\eta$ numerically.

Next, we are interested in the error or *suboptimality*

$$\mathcal{E}(\eta) = \langle \mathbf{c}, \mathbf{x}^\eta \rangle - \min_{\mathbf{x} \in \mathcal{P}} \langle \mathbf{c}, \mathbf{x} \rangle \tag{1}$$

measuring how suboptimal the solution $\mathbf{x}^\eta$ of (QLP) is when used as a feasible point in (LP). It follows from the optimality of $\mathbf{x}^\eta$ for (QLP) that $\eta \mapsto \mathcal{E}(\eta)$ is nonincreasing. (Figure 2 illustrates that it need not be strictly decreasing even on $[0, \eta^*]$). The optimality of $\mathbf{x}^\eta$ also implies that $\mathcal{E}(\eta) \leq \eta^{-1}(\|\mathbf{x}^*\|^2 - \|\mathbf{x}^\eta\|^2)$; in fact, an analogous result holds for any regularization. The following improvement is particular to the quadratic penalty and will be important for our main result.

Lemma 2.2. *Let $\mathbf{x}^\eta$ be the unique minimizer of* (QLP) *and let $\mathbf{x}^*$ be any minimizer of* (LP). *Then*

$$\mathcal{E}(\eta) \leq \frac{\|\mathbf{x}^*\|^2 - \|\mathbf{x}^\eta\|^2 - \|\mathbf{x}^* - \mathbf{x}^\eta\|^2}{\eta} \quad \textit{for all } \eta > 0.$$

Remark 2.3. The bound in Lemma 2.2 cannot be improved in general. Indeed, consider the example $\mathcal{P} = [0, 1]$ and $\mathbf{c} = -1$. Then $\mathbf{x}^* = 1$ and $\mathbf{x}^\eta = \eta/2$ for $\eta \in (0, 2]$, whereas $\mathbf{x}^\eta = \mathbf{x}^*$ for $\eta \geq 2$. It is straightforward to check that the inequality in Lemma 2.2 is an equality for all $\eta > 0$.

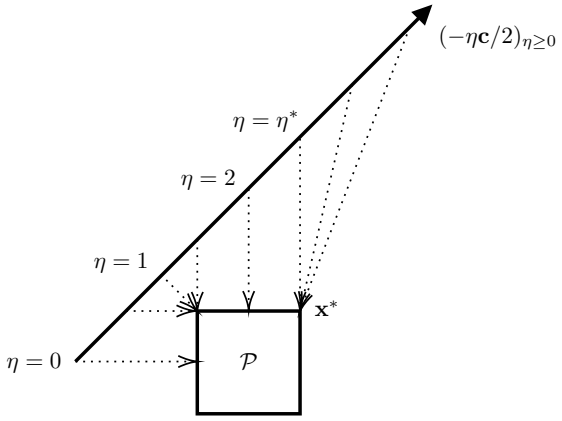

Figure 2: The minimizer $\mathbf{x}^\eta$ of (QLP) is the projection of $-\eta\mathbf{c}/2$ onto $\mathcal{P}$. The curve $\eta \mapsto \mathbf{x}^\eta$ is piecewise affine and converges stationarily to a point $\mathbf{x}^*$; i.e., $\mathbf{x}^\eta = \mathbf{x}^*$ for all $\eta \geq \eta^*$.

The next lemma details the piecewise linear nature of the curve $\eta \mapsto \mathbf{x}^\eta$. This result is known (even for some more general norms, see [23] and the references therein), and so is the stationary convergence [39, Theorem 2.1]. For completeness, we detail a short proof in Section 4.

Lemma 2.4. *Let $\mathbf{x}^\eta$ be the unique minimizer of* (QLP)*. The curve $\eta \mapsto \mathbf{x}^\eta$ is piecewise linear and converges stationarily to $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{M}} \|\mathbf{x}\|^2$ as $\eta \to \infty$. That is, there exist $n \in \mathbb{N}$ and*

$$0 = \eta_0 < \eta_1 < \cdots < \eta_n =: \eta^*$$

such that $[\eta_i, \eta_{i+1}] \ni \eta \mapsto \mathbf{x}^\eta$ is affine for every $i \in \{0, \ldots, n-1\}$, and moreover,

$$\mathbf{x}^\eta = \mathbf{x}^* \quad \text{for all } \eta \geq \eta^*.$$

Correspondingly, the suboptimality $\mathcal{E}(\eta) = \langle \mathbf{c}, \mathbf{x}^\eta - \mathbf{x}^* \rangle$ is also piecewise linear and converges stationarily to zero.

We can now state our main result for regime of small regularization: the threshold η^* beyond which $\mathbf{x}^\eta = \mathbf{x}^*$ and a bound for the slope of the suboptimality $\mathcal{E}(\eta)$ of (1) before the threshold. See Figures 1 and 2 for illustrations. We recall that $\mathcal{M}$ denotes the set of minimizers of (LP) and $\exp(\mathcal{P})$ denotes the extreme points of $\mathcal{P}$.

Theorem 2.5. *Let $\mathbf{x}^\eta$ be the unique minimizer of* (QLP) *and let $\mathbf{x}^*$ be the minimizer of* (LP) *with minimal norm, $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{M}} \|\mathbf{x}\|^2$. Let $0 = \eta_0 < \eta_1 < \cdots < \eta_n = \eta^*$ be the breakpoints of the curve $\eta \mapsto \mathbf{x}^\eta$ as in Lemma 2.4; in particular, η^* is the threshold such that $\mathbf{x}^\eta = \mathbf{x}^*$ for all $\eta \geq \eta^*$.*

(a) The threshold η^* is given by

$$\eta^* = 2 \max_{\mathbf{x}\in\exp(\mathcal{P})\setminus\mathcal{M}} \frac{\langle \mathbf{x}^*, \mathbf{x}^* - \mathbf{x} \rangle}{\langle \mathbf{c}, \mathbf{x} - \mathbf{x}^* \rangle}. \tag{2}$$

The right-hand side attains its maximum on the set $\mathcal{M}(\mathcal{P}, \mathbf{c}^*)$ of minimizers for the linear program (LP) *with the auxiliary cost $\mathbf{c}^* := \frac{\eta^*\mathbf{c}}{2} + \mathbf{x}^*$. Moreover, we have $\mathbf{x}^\eta \in \mathcal{M}(\mathcal{P}, \mathbf{c}^*)$ for all $\eta \in [\eta_{n-1}, \eta^*]$, so that $\eta^* = 2\frac{\langle \mathbf{x}^*, \mathbf{x}^* - \mathbf{x}^\eta \rangle}{\langle \mathbf{c}, \mathbf{x}^\eta - \mathbf{x}^* \rangle}$ for all $\eta \in [\eta_{n-1}, \eta^*]$.*

(b) The slope $\frac{\mathcal{E}(\eta)}{(\eta^*-\eta)}$ *of the last segment of the curve* $\eta \mapsto \mathcal{E}(\eta)$ *satisfies the bound*

$$\frac{\mathcal{E}(\eta)}{(\eta^*-\eta)} \le \frac{1}{2}\left\langle \mathbf{c}, \frac{\mathbf{x}^* - \mathbf{x}^{\eta_{n-1}}}{\|\mathbf{x}^* - \mathbf{x}^{\eta_{n-1}}\|}\right\rangle^2 \le \frac{\|\mathbf{c}\|^2}{2}, \qquad \eta \in [\eta_{n-1}, \eta^*). \tag{3}$$

It is worth noting that the first bound in (3) is in terms of the *angle* between $\mathbf{c}$ and $\mathbf{x}^* - \mathbf{x}^{\eta_{n-1}}$. The formula (2) for η^* is somewhat implicit in that it refers to $\mathbf{x}^*$. The following corollary states a bound for η^* using similar quantities as [51] uses for entropic regularization. In particular, we define the *suboptimality gap* of $\mathcal{P}$ as

$$\Delta := \min_{\mathbf{x}\in \mathrm{exp}(\mathcal{P})\setminus\mathcal{M}} \langle \mathbf{c}, \mathbf{x}\rangle - \min_{\mathbf{x}\in\mathcal{P}} \langle \mathbf{c}, \mathbf{x}\rangle = \min_{\mathbf{x}\in \mathrm{exp}(\mathcal{P})\setminus\mathcal{M}} \langle \mathbf{c}, \mathbf{x} - \mathbf{x}^*\rangle;$$

it measures the cost difference between the suboptimal and the optimal vertices of $\mathcal{P}$.

Corollary 2.6. *Let* $B = \sup_{\mathbf{x}\in\mathcal{P}} \|\mathbf{x}\|$ *and* $D = \sup_{\mathbf{x},\mathbf{x}'\in\mathcal{P}} \|\mathbf{x}-\mathbf{x}'\|$ *be the bound and diameter of* $\mathcal{P}$*, respectively. Then*

$$\eta^* \le \frac{2BD}{\Delta}.$$

For integer programs, where $\mathbf{c}$ and the vertices of $\mathcal{P}$ have integer coordinates, it is clear that $\Delta \ge 1$. In general, the explicit computation of Δ is not obvious. In Section 3 below we shall find it more useful to directly use (2).

We conclude this section with a quantitative result for the regime $\eta \to 0$ of large regularization. After rescaling with η, the quadratically regularized linear program (QLP) formally tends to the quadratic program

$$\text{minimize} \quad \|\mathbf{x}\|^2 \quad \text{subject to} \quad \mathbf{x} \in \mathcal{P}. \tag{QP}$$

The unique solution $\mathbf{x}^0$ of (QP) is simply the projection of the origin onto $\mathcal{P}$. It is known in several contexts that $\mathbf{x}^\eta \to \mathbf{x}^0$ as $\eta \to 0$ (e.g., [18, Properties 2,7]). The following result quantifies this convergence by establishing that $\|\mathbf{x}^\eta - \mathbf{x}^0\|$ tends to zero at a linear rate.

Proposition 2.7. *Let* $\mathbf{x}^\eta$ *and* $\mathbf{x}^0$ *be the minimizers of* (QLP) *and* (QP)*, respectively. Then*

$$\|\mathbf{x}^\eta - \mathbf{x}^0\| \le \frac{1}{2}\|\mathbf{c}\|\eta \quad \textit{for all } \eta > 0.$$

Remark 2.8. The bound in Proposition 2.7 is sharp in the example $\mathcal{P} = [0, 1]$ and $\mathbf{c} = -1$.

Remark 2.9. Proposition 2.7 and its proof apply to an arbitrary closed, bounded convex set $\mathcal{P}$ in a Hilbert space, not necessarily a polytope. In particular, the bounds also hold for continuous optimal transport problems.

3 Application to Optimal Transport

Recall from the Introduction the optimal transport problem with cost function $c(\cdot,\cdot)$ between probability measures μ and ν,

$$\inf_{\gamma\in\Gamma(\mu,\nu)} \int c(x,y)d\gamma(x,y), \tag{OT}$$

where $\Gamma(\mu,\nu)$ denotes the set of couplings of (μ,ν), and its quadratically regularized version

$$\inf_{\gamma\in\Gamma(\mu,\nu)} \int c(x,y)d\gamma(x,y) + \frac{1}{\eta}\left\|\frac{d\gamma}{d(\mu\otimes\nu)}\right\|^2_{L^2(\mu\otimes\nu)}. \tag{QOT}$$

Throughout this section, we consider given points $\mathbf{X}_i, \mathbf{Y}_i$, $1 \le i \le N$ (in $\mathbb{R}^D$, say) with their associated empirical measures and cost matrix

$$\mu = \frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{X}_i}, \qquad \nu = \frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{Y}_i}, \qquad C_{ij} := c(\mathbf{X}_i, \mathbf{X}_j).$$

Any coupling γ gives rise to a matrix $\gamma_{ij} = \gamma(\mathbf{X}_i, \mathbf{Y}_j)$ through its probability mass function. Those matrices form the set

$$\Gamma_N = \{\gamma \in \mathbb{R}^{N\times N} : \gamma\,\mathbf{1} = N^{-1}\mathbf{1},\ \gamma^\top\mathbf{1} = N^{-1}\mathbf{1},\ \gamma_{i,j} \ge 0\}.$$

It is more standard to work instead with the *Birkhoff polytope* of doubly stochastic matrices,

$$\Pi_N = \{\pi \in \mathbb{R}^{N\times N} : \pi\,\mathbf{1} = \mathbf{1},\ \pi^\top\mathbf{1} = \mathbf{1},\ \pi_{i,j} \ge 0\},$$

that is obtained through the bijection $\pi_{ij} = N\gamma_{ij}$. By Birkhoff's theorem (e.g., [13]), the extreme points $\exp(\Pi_N)$ are precisely the permutation matrices; i.e., matrices with binary entries whose rows and columns sum to one. Let $\langle A, B\rangle := \mathrm{Trace}(A^\top B) = \sum_{i=1}^N\sum_{j=1}^N A_{i,j}B_{i,j}$ be the Frobenius inner product on $\mathbb{R}^{N\times N}$ and $\|\cdot\|$ the associated norm. Then (QOT) becomes a particular case of (QLP), namely

$$\min_{\gamma\in\Gamma_N} \langle C,\gamma\rangle + \frac{N^2}{\eta}\|\gamma\|^2 \qquad \text{or equivalently} \qquad \min_{\pi\in\Pi_N} \frac{1}{N}\langle C,\pi\rangle + \frac{1}{\eta}\|\pi\|^2, \tag{4}$$

where the factor N^2 is due to $\mu\otimes\nu$ being the uniform measure on N^2 points. To have the same form as in (QLP) and Section 2, we write (4) as

$$\min_{\pi\in\Pi_N} \langle \mathbf{c},\pi\rangle + \frac{1}{\eta}\|\pi\|^2 \qquad \text{where } \mathbf{c}_{ij} := C_{ij}/N. \tag{5}$$

We can now apply the general results of Theorem 2.5 to (5) and infer the following for the regularized optimal transport problem (QOT); a detailed proof can be found in Section 4.

Proposition 3.1. *(a) The optimal coupling γ^η of* (QOT) *is optimal for* (OT) *if and only if*

$$\eta \ge \eta^* := 2\,N \cdot \max_{\pi\in\exp(\Pi_N)\setminus\mathcal{M}} \frac{\langle \pi^*, \pi^* - \pi\rangle}{\langle C, \pi - \pi^*\rangle}, \tag{6}$$

in which case γ^η is the minimum-norm solution γ^* of (OT).

(b) We have the following bound for the slope of the suboptimality,

$$\limsup_{\eta\to\eta^*} \frac{\int c(x,y)d\gamma^\eta(x,y) - \int c(x,y)d\gamma^*(x,y)}{\eta^* - \eta} \le \frac{1}{2}\left(\int c(x,y)^2 d(\mu\otimes\nu)(x,y) - \left(\int c(x,y)d(\mu\otimes\nu)(x,y)\right)^2\right). \tag{7}$$

The following example shows that Proposition 3.1 is sharp.

Example 3.1. Let $c(\mathbf{X}_i, \mathbf{Y}_j) = -\delta_{ij}$, so that $\pi^* = \mathrm{Id}$ is the identity matrix and $C = -\mathrm{Id}$. Note also that π^0 has entries $\pi^0_{i,j} = 1/N$. It follows from (6) that $\eta^* = 2N$, and the right-hand side of (7) evaluates to $\frac{N-1}{2N^2}$. We show below that $[0, \eta^*] \ni \eta \mapsto \mathbf{x}^\eta$ is affine, or more explicitly, that $\pi^\eta = \frac{2N-\eta}{2N}\pi^0 + \frac{\eta}{2N}\pi^* =: \tilde{\pi}^\eta$. As a consequence, we have for every $\eta \in [0, \eta^*)$ that

$$\begin{aligned}\frac{\int c(x,y)d\gamma^\eta(x,y) - \int c(x,y)d\gamma^*(x,y)}{\eta_* - \eta} &= \frac{\langle C, \pi^\eta - \pi^*\rangle}{N(\eta_* - \eta)} = -\frac{(2N-\eta) + (\eta - 2N)N}{2N^2(\eta_* - \eta)} \\ &= -\frac{(\eta^* - \eta) + (\eta - \eta^*)N}{2N^2(\eta_* - \eta)} = \frac{N-1}{2N^2},\end{aligned}$$

matching the right-hand side of (7).

It remains to show that $\pi^\eta = \tilde{\pi}^\eta$. Using $\mathbf{c} = \mathrm{Id}/N$, the definition of $\tilde{\pi}^\eta$ and $\pi^* = \mathrm{Id}$, we see that $\frac{\eta\mathbf{c}}{2} + \tilde{\pi}^\eta = \frac{2N-\eta}{2N}\pi^0$. The form of π^0 also implies that $\langle \pi^0, \pi' - \pi\rangle = 0$ for any $\pi, \pi' \in \Pi_N$. Together, it follows that $\langle -\frac{\eta\mathbf{c}}{2} - \tilde{\pi}^\eta, \tilde{\pi}^\eta - \pi\rangle = 0$ for all $\pi \in \Pi_N$. By Lemma 2.1, this implies $\tilde{\pi}^\eta = \pi^\eta$.

Next, we focus on a more representative class of transport problems. Our main interest is to see how our key quantities scale with N, the number of data points.

Corollary 3.2. *Assume that there exist $[\epsilon_m, \epsilon_M] \subset [0, \infty)$ and a permutation $\sigma^* : \{1, \dots, N\} \to \{1, \dots, N\}$ such that*

$$\kappa := \min_{i \in \{1,\dots,N\}, j \neq \sigma^*(i)} c(\mathbf{X}_i, \mathbf{Y}_j) > \epsilon_M \quad \text{and} \quad c(\mathbf{X}_i, \mathbf{Y}_{\sigma^*(i)}) \in [\epsilon_m, \epsilon_M] \quad \text{for all } i \in \{1, \dots, n\}.$$

Then

$$\frac{4\,N}{\kappa' - 2\epsilon_m} \leq \eta^* \leq \frac{2\,N}{\kappa - \epsilon_M}, \tag{8}$$

where $\kappa' := \min_{i \neq j} c(\mathbf{X}_i, \mathbf{Y}_{\sigma^*(j)}) + c(\mathbf{X}_j, \mathbf{Y}_{\sigma^*(i)})$. If the cost is symmetric around σ^* in the sense that $c(\mathbf{X}_i, \mathbf{Y}_{\sigma^*(j)}) = c(\mathbf{X}_j, \mathbf{Y}_{\sigma^*(i)})$ for all $i, j \in \{1, \dots, N\}$, then

$$\frac{2\,N}{\kappa - \epsilon_m} \leq \eta^* \leq \frac{2\,N}{\kappa - \epsilon_M}, \qquad \text{and in particular} \quad \eta^* = \frac{2\,N}{\kappa} \quad \text{if} \quad \epsilon_m = \epsilon_M = 0. \tag{9}$$

The proof is detailed in Section 4. We illustrate Proposition 3.1 and Corollary 3.2 with a representative example for scalar data.

Example 3.2. Consider the quadratic cost $c(x, y) = \|x - y\|^2$ and $\mathbf{X}_i = \mathbf{Y}_i = \frac{i}{N}$, $1 \leq i \leq N$ with $N \geq 2$, leading to the cost matrix

$$C_{ij} = \frac{|i-j|^2}{N^2}.$$

Then

$$\eta^* = 2N^3$$

and we have the following bound for the slope of the suboptimality,

$$\limsup_{\eta \to \eta^*} \frac{\int c(x,y)d\gamma^\eta(x,y) - \int c(x,y)d\gamma^*(x,y)}{\eta^* - \eta} \leq \frac{N-1}{N^6}. \tag{10}$$

Indeed, the value of η^* follows directly from the last part of (9) with $\kappa = 1/N^2$ and σ^* being the identity. The proof of (10) is longer and relegated to Section 4.

To study the accuracy of the bound (10), we compute numerically the limit

$$L_N = \lim_{\eta \to \eta^*} \frac{\int c(x,y) d\gamma^\eta(x,y) - \int c(x,y) d\gamma^*(x,y)}{\eta^* - \eta}$$

for $N = j * 30$ with $j = 2, \dots, 16$. Figure 3 shows $N \mapsto L_N$ in blue and the upper bound $N \mapsto \frac{N-1}{N^6}$ in red (in double logarithmic scale). We observe that both have the same order as a function of N.

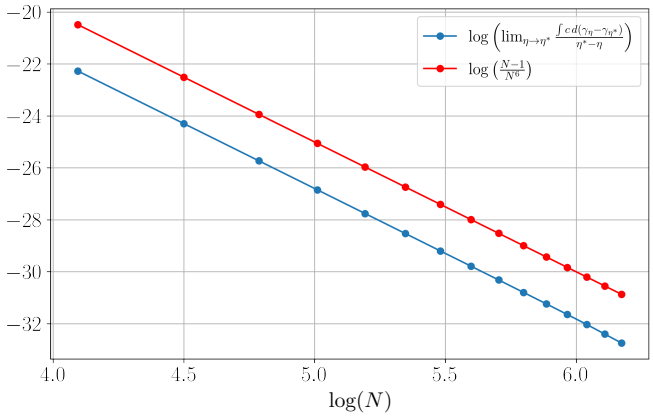

Figure 3: Accuracy of the bound (10). Plot of $N \mapsto \lim_{\eta \to \eta^*} \frac{\int c(x,y) d\gamma^\eta(x,y) - \int c(x,y) d\gamma^*(x,y)}{\eta^* - \eta}$ (blue) and the upper bound $N \mapsto \frac{N-1}{N^6}$ (red) in double logarithmic scale.

4 Proofs

Proof of Lemma 2.2. Let $\mathbf{x} \in \mathcal{P}$. Inserting the definition of Φ_η, expanding $\|\mathbf{x} - \mathbf{x}^\eta\|^2$, and applying Lemma 2.1 yield

$$\Phi_\eta(\mathbf{x}) = \Phi_\eta(\mathbf{x}^\eta) + \left\langle \mathbf{c} + \frac{2\mathbf{x}^\eta}{\eta}, \mathbf{x} - \mathbf{x}^\eta \right\rangle + \frac{\|\mathbf{x} - \mathbf{x}^\eta\|^2}{\eta} \geq \Phi_\eta(\mathbf{x}^\eta) + \frac{\|\mathbf{x} - \mathbf{x}^\eta\|^2}{\eta}.$$

Therefore,

$$0 \geq \Phi_\eta(\mathbf{x}^\eta) - \Phi_\eta(\mathbf{x}) + \frac{\|\mathbf{x} - \mathbf{x}^\eta\|^2}{\eta} = \langle \mathbf{c}, \mathbf{x}^\eta - \mathbf{x} \rangle + \frac{\|\mathbf{x}^\eta\|^2 - \|\mathbf{x}\|^2 + \|\mathbf{x} - \mathbf{x}^\eta\|^2}{\eta}$$

and in particular choosing $\mathbf{x} = \mathbf{x}^*$ gives

$$\mathcal{E}(\eta) = \langle \mathbf{c}, \mathbf{x}^\eta - \mathbf{x}^* \rangle \leq \frac{\|\mathbf{x}^*\|^2 - \|\mathbf{x}^\eta\|^2 - \|\mathbf{x}^* - \mathbf{x}^\eta\|^2}{\eta}$$

as claimed. □

Proof of Lemma 2.4 and Theorem 2.5. Step 1. Let $\eta_{(1)} < \eta_{(2)}$. We claim that if $\mathbf{x}^{\eta_{(1)}}, \mathbf{x}^{\eta_{(2)}} \in \mathrm{ri}(\mathcal{F})$ for some face[1] $\mathcal{F}$ of $\mathcal{P}$, then $[\eta_{(1)}, \eta_{(2)}] \ni \eta \mapsto \mathbf{x}^\eta$ is affine. Indeed, $\mathbf{x}^{\eta_{(i)}} = \mathrm{proj}_{\mathcal{P}}(-\eta_{(i)}\mathbf{c}/2)$ is the projection of $-\eta_{(i)}\mathbf{c}/2$ onto $\mathcal{P}$. As $\mathbf{x}^{\eta_{(i)}} \in \mathrm{ri}(\mathcal{F})$, it follows that $\mathbf{x}^{\eta_{(i)}} = \mathrm{proj}_A(-\eta_{(i)}\mathbf{c}/2)$ is also the projection onto the affine hull A of $\mathcal{F}$. Since A is an affine space, the map $\eta \mapsto \mathrm{proj}_A(-\eta\mathbf{c}/2)$ is affine. For $\eta_{(1)} \le \eta \le \eta_{(2)}$, convexity of $\mathrm{ri}(\mathcal{F})$ then implies $\mathrm{proj}_A(-\eta\mathbf{c}/2) \in \mathrm{ri}(\mathcal{F})$, which in turn implies $\mathrm{proj}_A(-\eta\mathbf{c}/2) = \mathrm{proj}_{\mathcal{F}}(-\eta\mathbf{c}/2) = \mathrm{proj}_{\mathcal{P}}(-\eta\mathbf{c}/2) = \mathbf{x}^\eta$.

Step 2. We can now define $\eta_1, \ldots, \eta_n$ recursively as follows. Recall first that each $\mathbf{x} \in \mathcal{P}$ is in the relative interior of exactly one face of $\mathcal{P}$ (possibly $\mathcal{P}$ itself), namely the smallest face containing $\mathbf{x}$ [12, Theorem 5.6]. Let $\mathcal{F}_0$ be the unique face such that $\mathbf{x}^0 := \arg\min_{\mathbf{x}\in\mathcal{P}} \|\mathbf{x}\| \in \mathrm{ri}(\mathcal{F}_0)$ and define

$$\eta_1 := \inf\{\eta > 0 : \ \mathbf{x}^\eta \notin \mathrm{ri}(\mathcal{F}_0)\},$$

where we use the convention that $\inf\emptyset = +\infty$. Then $(0, \eta_1) \ni \eta \mapsto \mathbf{x}^\eta$ is affine by Step 1. For $i > 1$, if $\eta_{i-1} < \infty$, let $\mathcal{F}_{i-1}$ be the face such that $\mathbf{x}^{\eta_{i-1}} \in \mathrm{ri}(\mathcal{F}_{i-1})$ and define

$$\eta_i := \inf\{\eta > \eta_{i-1} : \ \mathbf{x}^\eta \notin \mathrm{ri}(\mathcal{F}_{i-1})\}.$$

Again, $(\eta_{i-1}, \eta_i) \ni \eta \mapsto \mathbf{x}^\eta$ is affine by Step 1. Moreover, by continuity, $[\eta_{i-1}, \eta_i] \ni \eta \mapsto \mathbf{x}^\eta$ is also affine.

Step 3. Next, we establish the value (2) of η^*. Let us first observe that (2) is strictly positive. Indeed, the denominator is clearly positive. Suppose that the numerator $\langle \mathbf{x}^*, \mathbf{x} - \mathbf{x}^*\rangle \le 0$ for all $\mathbf{x} \in \exp(\mathcal{P}) \setminus \mathcal{M}$. Note that by the definition of $\mathbf{x}^*$, we also have $\langle \mathbf{x}^*, \mathbf{x} - \mathbf{x}^*\rangle \le 0$ for all $\mathbf{x} \in \mathcal{M}$. Thus $\langle \mathbf{x}^*, \mathbf{x} - \mathbf{x}^*\rangle \le 0$ for all $\mathbf{x} \in \exp(\mathcal{P})$, meaning that $\mathbf{x}^*$ is the projection of the origin onto $\mathcal{P}$, the degenerate situation we had excluded in our setup.

Let $\eta > 0$ and suppose that $\mathbf{x}^* = \mathbf{x}^\eta$. Then by Lemma 2.1,

$$-\langle \mathbf{x}^*, \mathbf{x} - \mathbf{x}^*\rangle \le \left\langle \frac{\eta\mathbf{c}}{2}, \mathbf{x} - \mathbf{x}^* \right\rangle \quad \text{for all } \mathbf{x} \in \mathcal{P}.$$

Using also that $\langle \mathbf{c}, \mathbf{x} - \mathbf{x}^*\rangle > 0$ for $\mathbf{x} \in \mathcal{P} \setminus \mathcal{M}$, we deduce

$$\eta \ge 2\frac{\langle \mathbf{x}^*, \mathbf{x}^* - \mathbf{x}\rangle}{\langle \mathbf{c}, \mathbf{x} - \mathbf{x}^*\rangle} \quad \text{for all } \mathbf{x} \in \exp(\mathcal{P}) \setminus \mathcal{M}. \tag{11}$$

Conversely, assume that (11) holds; we show that $\mathbf{x}^* = \mathbf{x}^\eta$. Recall that $\exp(\mathcal{P}) = \{\mathbf{v}_1, \ldots, \mathbf{v}_K\}$ denotes the set of extreme points of $\mathcal{P}$. Let $\mathbf{x} \in \mathcal{P}$, then there exist $\{\lambda_i\}_{i=1}^K \subset [0,1]$ with $1 = \sum_{i=1}^K \lambda_i$ such that $\mathbf{x} = \sum_{i=1}^K \lambda_i \mathbf{v}_i$. We note that (11) yields

$$\left\langle \frac{\eta\mathbf{c}}{2}, \mathbf{x} - \mathbf{x}^* \right\rangle = \sum_{i:\, \mathbf{v}_i \in \exp(\mathcal{P})\setminus\exp(\mathcal{M})} \lambda_i \left\langle \frac{\eta\mathbf{c}}{2}, \mathbf{v}_i - \mathbf{x}^* \right\rangle \ge - \sum_{i:\, \mathbf{v}_i \in \exp(\mathcal{P})\setminus\exp(\mathcal{M})} \lambda_i \left\langle \mathbf{x}^*, \mathbf{v}_i - \mathbf{x}^* \right\rangle.$$

On the other hand, the fact that $\mathbf{x}^*$ is the projection of the origin onto $\mathcal{M}$ yields

$$\sum_{i:\, \mathbf{v}_i \in \exp(\mathcal{M})} \lambda_i \left\langle \mathbf{x}^*, \mathbf{v}_i - \mathbf{x}^* \right\rangle \ge 0.$$

[1] A nonempty face $\mathcal{F}$ of the polytope $\mathcal{P}$ can be defined as a subset $\mathcal{F} \subset \mathcal{P}$ such that there exists an affine hyperplane $H = \{\mathbf{x} \in \mathbb{R}^d : \langle \mathbf{x}, \mathbf{a}\rangle = m\}$ with $H \cap \mathcal{P} = \mathcal{F}$ and $\mathcal{P} \subset \{\mathbf{x} \in \mathbb{R}^d : \langle \mathbf{x}, \mathbf{a}\rangle \le m\}$. See [12].

Together,

$$\left\langle \frac{\eta \mathbf{c}}{2}, \mathbf{x} - \mathbf{x}^* \right\rangle \geq - \sum_{i\colon \mathbf{v}_i \in \exp(\mathcal{P}) \setminus \exp(\mathcal{M})} \lambda_i \left\langle \mathbf{x}^*, \mathbf{v}_i - \mathbf{x}^* \right\rangle \geq - \sum_{i\colon \mathbf{v}_i \in \exp(\mathcal{P})} \lambda_i \left\langle \mathbf{x}^*, \mathbf{v}_i - \mathbf{x}^* \right\rangle = - \left\langle \mathbf{x}^*, \mathbf{x} - \mathbf{x}^* \right\rangle .$$

As $\mathbf{x} \in \mathcal{P}$ was arbitrary, Lemma 2.1 now shows that $\mathbf{x}^* = \mathbf{x}^\eta$. This completes the proof of Lemma 2.4 and (2).

Finally, note that $\mathbf{x}$ attains the maximum in (2) if and only if $\langle \mathbf{c}^*, \mathbf{x} - \mathbf{x}^* \rangle = 0$. Moreover, $\langle \mathbf{c}^*, \mathbf{x} - \mathbf{x}^* \rangle \geq 0$ for all $\mathbf{x} \in \mathcal{P}$ by Lemma 2.1. Hence the set of maximizers of (2) over $\exp(\mathcal{P}) \setminus \mathcal{M}$ equals the set of minimizers of $\langle \mathbf{c}^*, \cdot \rangle$ over $\exp(\mathcal{P})$.

Step 4. We prove the remaining claim in (a), namely that $\mathbf{x}^\eta \in \mathcal{M}(\mathcal{P}, \mathbf{c}^*)$ for all $\eta \in [\eta_{n-1}, \eta^*]$. By Lemma 2.1,

$$\left\langle -\frac{\eta \mathbf{c}}{2} - \mathbf{x}^\eta, \mathbf{x} - \mathbf{x}^\eta \right\rangle \leq 0 \quad \text{for all } \mathbf{x} \in \mathcal{P},\ \eta \in [\eta_{n-1}, \eta^*].$$

As $\mathbf{x}^\eta \in \operatorname{ri}([\mathbf{x}^{\eta_{n-1}}, \mathbf{x}^*])$ for $\eta \in (\eta_{n-1}, \eta^*)$, Lemma 2.1 moreover yields

$$\left\langle -\frac{\eta \mathbf{c}}{2} - \mathbf{x}^\eta, \mathbf{x}^{\eta'} - \mathbf{x}^\eta \right\rangle = 0 \quad \text{for all } \eta' \in [\eta_{n-1}, \eta^*],\ \eta \in (\eta_{n-1}, \eta^*),$$

and by continuity, the previous display also holds for $\eta \in [\eta_{n-1}, \eta^*]$. In summary, we have

$$\left\langle -\frac{\eta^* \mathbf{c}}{2} - \mathbf{x}^*, \mathbf{x} - \mathbf{x}^* \right\rangle \leq 0 \quad \text{for all } \mathbf{x} \in \mathcal{P} \tag{12}$$

and

$$\left\langle -\frac{\eta^* \mathbf{c}}{2} - \mathbf{x}^*, \mathbf{x}^{\eta_{n-1}} - \mathbf{x}^* \right\rangle = 0.$$

Therefore, $\mathbf{x}^{\eta_{n-1}} \in \mathcal{M}(\mathcal{P}, \mathbf{c}^*)$. On the other hand, (12) also states that $\mathbf{x}^{\eta^*} = \mathbf{x}^* \in \mathcal{M}(\mathcal{P}, \mathbf{c}^*)$, and then convexity implies the claim.

Step 5. It remains to prove (b). Let $\eta \in (\eta_{n-1}, \eta^*)$. Then Lemma 2.4 implies that $\mathbf{x}^\eta = \lambda \mathbf{x}^{\eta_{n-1}} + (1 - \lambda)\mathbf{x}^*$ for some $\lambda \in (0, 1)$ and thus

$$\langle \mathbf{c}, \mathbf{x}^\eta \rangle = \langle \mathbf{c}, \mathbf{x}^* \rangle + \lambda \langle \mathbf{c}, \mathbf{x}^{\eta_{n-1}} - \mathbf{x}^* \rangle.$$

Lemma 2.2 then yields

$$\lambda = \frac{\langle \mathbf{c}, \mathbf{x}^\eta - \mathbf{x}^* \rangle}{\langle \mathbf{c}, \mathbf{x}^{\eta_{n-1}} - \mathbf{x}^* \rangle} \leq \frac{\|\mathbf{x}^*\|^2 - \|\mathbf{x}^\eta\|^2 - \|\mathbf{x}^* - \mathbf{x}^\eta\|^2}{\eta \langle \mathbf{c}, \mathbf{x}^{\eta_{n-1}} - \mathbf{x}^* \rangle}.$$

Using

$$\|\mathbf{x}^\eta\|^2 = \|\mathbf{x}^*\|^2 + \lambda^2 \|\mathbf{x}^* - \mathbf{x}^{\eta_{n-1}}\|^2 + 2\lambda \langle \mathbf{x}^*, \mathbf{x}^{\eta_{n-1}} - \mathbf{x}^* \rangle$$

and $\|\mathbf{x}^\eta - \mathbf{x}^*\|^2 = \lambda^2 \|\mathbf{x}^* - \mathbf{x}^{\eta_{n-1}}\|^2$, it follows that

$$\lambda \leq \frac{2\lambda \langle \mathbf{x}^*, \mathbf{x}^* - \mathbf{x}^{\eta_{n-1}} \rangle - 2\lambda^2 \|\mathbf{x}^* - \mathbf{x}^{\eta_{n-1}}\|^2}{\eta \langle \mathbf{c}, \mathbf{x}^{\eta_{n-1}} - \mathbf{x}^* \rangle}.$$

and hence

$$\lambda \leq \frac{2\langle \mathbf{x}^*, \mathbf{x}^* - \mathbf{x}^{\eta_{n-1}} \rangle - \eta \langle \mathbf{c}, \mathbf{x}^{\eta_{n-1}} - \mathbf{x}^* \rangle}{2\|\mathbf{x}^* - \mathbf{x}^{\eta_{n-1}}\|^2}. \tag{13}$$

By the last part of (a) we have

$$\eta^* = \frac{2\,\langle \mathbf{x}^*, \mathbf{x}^* - \mathbf{x}^{\eta_{n-1}}\rangle}{\langle \mathbf{c}, \mathbf{x}^{\eta_{n-1}} - \mathbf{x}^*\rangle}.$$

Inserting this in (13) yields

$$\lambda \leq \frac{(\eta^* - \eta)\langle \mathbf{c}, \mathbf{x}^{\eta_{n-1}} - \mathbf{x}^*\rangle}{2\|\mathbf{x}^* - \mathbf{x}^{\eta_{n-1}}\|^2}$$

and now it follows that

$$\mathcal{E}(\eta) = \lambda\langle \mathbf{c}, \mathbf{x}^{\eta_{n-1}} - \mathbf{x}^*\rangle \leq \frac{(\eta^* - \eta)\langle \mathbf{c}, \mathbf{x}^{\eta_{n-1}} - \mathbf{x}^*\rangle^2}{2\|\mathbf{x}^* - \mathbf{x}^{\eta_{n-1}}\|^2}$$

as claimed. □

Proof of Proposition 2.7. Recall that $\mathbf{x}^\eta$ is the projection of $-\eta\mathbf{c}/2$ onto $\mathcal{P}$ whereas $\mathbf{x}^0$ is the projection of the origin onto $\mathcal{P}$. As the projection operator onto a convex set is non-expanding (i.e., Lipschitz continuous with constant one), this implies $\|\mathbf{x}^\eta - \mathbf{x}^0\| \leq \| - \eta\mathbf{c}/2\| = \frac{1}{2}\|\mathbf{c}\|\eta$. □

Proof of Proposition 3.1. Theorem 2.5(a) directly yields (6). Whereas for (7), a direct application of Theorem 2.5(b) only yields

$$\limsup_{\eta\to\eta^*} \frac{\int c(x,y)d\gamma^\eta(x,y) - \int c(x,y)d\gamma^*(x,y)}{\eta^* - \eta} \leq \frac{1}{2}\int c(x,y)^2 d(\mu\otimes\nu)(x,y).$$

To improve this bound, note that the optimizer of (QOT) does not change if the cost $c(x,y)$ is changed by an additive constant. Moreover, for any $m \in \mathbb{R}$,

$$\int c(x,y)d\gamma^\eta(x,y) - \int c(x,y)d\gamma^*(x,y) \;=\; \int (c(x,y)-m)d\gamma^\eta(x,y) - \int (c(x,y)-m)d\gamma^*(x,y).$$

Applying Theorem 2.5 with the modified cost $c(x,y)-m$ for the choice $m := \int c(x,y)d(\mu\otimes\nu)(x,y)$ yields (7). □

Proof of Corollary 3.2. Assume without loss of generality that σ^* is the identity, so that $\pi^* = \mathrm{Id}$ is the identity matrix. Let P_σ be the permutation matrix associated with a permutation $\sigma : \{1,\dots,N\} \to \{1,\dots,N\}$. We define $\mathcal{N}(\sigma) = \{i \in \{1,\dots,N\} :\ \sigma(i) = i\}$. Then

$$\frac{\langle \pi^*, \pi^* - P_\sigma\rangle}{\langle C, P_\sigma - \pi^*\rangle} = \frac{N - |\mathcal{N}(\sigma)|}{\sum_{i\notin\mathcal{N}(\sigma)} c(\mathbf{X}_i, \mathbf{Y}_{\sigma(i)}) - c(\mathbf{X}_i, \mathbf{Y}_i)}, \tag{14}$$

where $|\mathcal{N}(\sigma)|$ denotes the cardinality of $\mathcal{N}(\sigma)$.

For the upper bound in (8), we recall that $c(\mathbf{X}_i, \mathbf{Y}_i) \leq \epsilon_M$ and $c(\mathbf{X}_i, \mathbf{Y}_{\sigma(i)}) \geq \kappa$ for $i \notin \mathcal{N}(\sigma)$, so that (14) yields

$$\frac{\langle \pi^*, \pi^* - P_\sigma\rangle}{\langle C, P_\sigma - \pi^*\rangle} \leq \frac{1}{\kappa - \epsilon_M}\frac{N - |\mathcal{N}(\sigma)|}{N - |\mathcal{N}(\sigma)|} = \frac{1}{\kappa - \epsilon_M}.$$

Now Proposition 3.1 yields the claim. For the lower bound in (8), let $i^*, j^* \neq \sigma^*(i^*)$ be such that $\kappa' = c(\mathbf{X}_{i^*}, \mathbf{Y}_{j^*}) + c(\mathbf{X}_{j^*}, \mathbf{Y}_{i^*})$ and let σ be the permutation such that $\sigma(i) = i$ for all $i \notin \{i^*, j^*\}$, $\sigma(i^*) = j^*$ and $\sigma(j^*) = i^*$. Then

$$\frac{\langle \pi^*, \pi^* - P_\sigma \rangle}{\langle C, P_\sigma - \pi^* \rangle} = \frac{2}{c(\mathbf{X}_{i^*}, \mathbf{Y}_{j^*}) + c(\mathbf{X}_{j^*}, \mathbf{Y}_{i^*}) - (c(\mathbf{X}_{i^*}, \mathbf{Y}_{i^*}) + c(\mathbf{X}_{j^*}, \mathbf{Y}_{j^*}))} \geq \frac{2}{\kappa' - 2\epsilon_m}$$

and Proposition 3.1 again yields the claim. It remains to observe that $\kappa' = 2\kappa$ when the cost is symmetric. □

Proof for Example 3.2. Corollary 3.2 applies with σ^* being the identity and $\kappa = 1/N^2$. As a consequence, the critical value η^* is $2N^3$.

To prove (10), write $\pi^{\eta_{n-1}} = \sum_{i=1}^k \lambda_i P_{\sigma_i}$ with $\lambda_i \in (0, 1]$ and $\sum_{i=1}^k \lambda_i = 1$. Recall from Theorem 2.5(a) that $0 = \langle \mathbf{c}^*, \pi^{\eta_{n-1}} - \pi^* \rangle$. With the optimality of $\pi^* = \pi^{\eta^*}$ for $\langle \mathbf{c}^*, \cdot \rangle$, this implies

$$0 = \langle \mathbf{c}^*, P_{\sigma_i} - \pi^* \rangle = \left\langle \frac{\eta^* C}{2\,N} + \pi^*, P_{\sigma_i} - \pi^* \right\rangle = \left\langle N^2 C + \pi^*, P_{\sigma_i} - \pi^* \right\rangle \quad \text{for all } i = 1, \ldots, k.$$

As $\langle N^2 C + \pi^*, \pi^* \rangle = \langle N^2 C + \mathrm{Id}, \mathrm{Id} \rangle = N$, it follows that $\left\langle N^2 C + \mathrm{Id}, P_{\sigma_i} \right\rangle = N$. Using that P_{σ_i} has N entries equal to one and that the entries of $N^2 C + \mathrm{Id}$ are strictly larger than one outside the three principal diagonals, this implies that $|\sigma_i(j) - j| \leq 1$ for all $j \in \{1, \ldots, N\}$. As a consequence, $\pi^{\eta_{n-1}} = \sum_{i=1}^k \lambda_i P_{\sigma_i}$ vanishes outside the three principal diagonals; i.e., it is entry-wise smaller or equal to the tridiagonal matrix

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 1 & 1 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 1 \\ 0 & 0 & 0 & 0 & \cdots & 1 & 1 \end{pmatrix}.$$

Let $\bar{\mathbf{c}} := A \odot \mathbf{c}$ be the entry-wise product, meaning that entries of $\mathbf{c}$ outside the three principal diagonals are set to zero. As $\pi^{\eta_{n-1}} - \mathrm{Id}$ vanishes outside those diagonals, we have

$$\langle \pi^{\eta_{n-1}} - \mathrm{Id}, \mathbf{c} \rangle = \langle \pi^{\eta_{n-1}} - \mathrm{Id}, \bar{\mathbf{c}} \rangle.$$

We can now use Theorem 2.5(b) and the Cauchy–Schwarz inequality to find

$$\limsup_{\eta \to \eta^*} \frac{\langle \pi^\eta - \mathrm{Id}, \mathbf{c} \rangle}{(\eta^* - \eta)} \leq \frac{\langle \pi^{\eta_{n-1}} - \mathrm{Id}, \mathbf{c} \rangle^2}{2\|\pi^{\eta_{n-1}} - \mathrm{Id}\|^2} = \frac{\langle \pi^{\eta_{n-1}} - \mathrm{Id}, \bar{\mathbf{c}} \rangle^2}{2\|\pi^{\eta_{n-1}} - \mathrm{Id}\|^2} \leq \frac{\|\bar{\mathbf{c}}\|^2}{2} = \frac{1}{2N^2} \frac{2(N-1)}{N^4} = \frac{N-1}{N^6}$$

as claimed in (10). □

Statements and Declarations M. Nutz was partially supported by NSF Grants DMS-1812661, DMS-2106056, DMS-2407074. The authors have no relevant financial or non-financial interests to disclose. All authors have contributed to all parts of the paper.

References

[1] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM J. Math. Anal.*, 43(2):904–924, 2011.

[2] J. M. Altschuler, J. Niles-Weed, and A. J. Stromme. Asymptotics for semidiscrete entropic optimal transport. *SIAM J. Math. Anal.*, 54(2):1718–1741, 2022.

[3] D. Alvarez-Melis and T. Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, 2018.

[4] J. Backhoff-Veraguas, D. Bartl, M. Beiglböck, and M. Eder. All adapted topologies are equal. *Probab. Theory Related Fields*, 178(3-4):1125–1172, 2020.

[5] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proc. Natl. Acad. Sci. USA*, 117(48):30063–30070, 2020.

[6] E. Bayraktar, S. Eckstein, and X. Zhang. Stability and sample complexity of divergence regularized optimal transport. *Bernoulli*, 31(1):213–239, 2025.

[7] M. Beiglböck, P. Henry-Labordère, and F. Penkner. Model-independent bounds for option prices: a mass transport approach. *Finance Stoch.*, 17(3):477–501, 2013.

[8] M. Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

[9] E. Bernton, P. Ghosal, and M. Nutz. Entropic optimal transport: Geometry and large deviations. *Duke Math. J.*, 171(16):3363–3400, 2022.

[10] M. Blondel, V. Seguy, and A. Rolet. Smooth and sparse optimal transport. volume 84 of *Proceedings of Machine Learning Research*, pages 880–889, 2018.

[11] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.

[12] A. Brøndsted. *An introduction to convex polytopes*, volume 90 of *Graduate Texts in Mathematics*. Springer-Verlag, New York-Berlin, 1983.

[13] R. A. Brualdi. *Combinatorial matrix classes*, volume 108 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 2006.

[14] G. Carlier, V. Duval, G. Peyré, and B. Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM J. Math. Anal.*, 49(2):1385–1418, 2017.

[15] R. Cominetti and J. San Martín. Asymptotic analysis of the exponential penalty trajectory in linear programming. *Math. Programming*, 67(2, Ser. A):169–187, 1994.

[16] G. Conforti and L. Tamanini. A formula for the time derivative of the entropic cost and applications. *J. Funct. Anal.*, 280(11):108964, 2021.

[17] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pages 2292–2300. 2013.

[18] A. Dessein, N. Papadakis, and J.-L. Rouas. Regularized optimal transport and the rot mover's distance. *J. Mach. Learn. Res.*, 19(15):1–53, 2018.

[19] S. Di Marino and A. Gerolin. Optimal transport losses and Sinkhorn algorithm with general convex regularization. *Preprint arXiv:2007.00976v1*, 2020.

[20] S. Eckstein and M. Kupper. Computation of optimal transport and related hedging problems via penalization and neural networks. *Appl. Math. Optim.*, 83(2):639–667, 2021.

[21] S. Eckstein and M. Nutz. Convergence rates for regularized optimal transport via quantization. *Math. Oper. Res.*, 49(2):1223–1240, 2024.

[22] M. Essid and J. Solomon. Quadratically regularized optimal transport on graphs. *SIAM J. Sci. Comput.*, 40(4):A1961–A1986, 2018.

[23] M. Finzel and W. Li. Piecewise affine selections for piecewise polyhedral multifunctions and metric

projections. *J. Convex Anal.*, 7(1):73–94, 2000.

[24] A. Galichon. *Optimal transport methods in economics*. Princeton University Press, Princeton, NJ, 2016.

[25] A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems 29*, pages 3440–3448, 2016.

[26] A. González-Sanz and M. Nutz. Sparsity of quadratically regularized optimal transport: Scalar case. *Preprint arXiv:2410.03353v1*, 2024.

[27] A. González-Sanz, M. Nutz, and A. Riveros Valdevenito. Monotonicity in quadratically regularized linear programs. *Preprint arXiv:2408.07871v1*, 2024.

[28] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of Wasserstein GANs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5769–5779, 2017.

[29] W. W. Hager and H. Zhang. Projection onto a polyhedron that exploits sparsity. *SIAM J. Optim.*, 26(3):1773–1798, 2016.

[30] M. Hallin, E. del Barrio, J. Cuesta-Albertos, and C. Matrán. Distribution and quantile functions, ranks and signs in dimension d: a measure transportation approach. *Ann. Statist.*, 49(2):1139–1165, 2021.

[31] M. Hallin, D. Paindaveine, and M. Siman. Multivariate quantiles and multiple-output regression quantiles: from L_1 optimization to halfspace depth. *Ann. Statist.*, 38(2):635–669, 2010.

[32] R. Koenker and G. Bassett, Jr. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

[33] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017.

[34] C. Léonard. From the Schrödinger problem to the Monge-Kantorovich problem. *J. Funct. Anal.*, 262(4):1879–1920, 2012.

[35] L. Li, A. Genevay, M. Yurochkin, and J. Solomon. Continuous regularized Wasserstein barycenters. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17755–17765. Curran Associates, Inc., 2020.

[36] Q. Li, R. Xi, and N. Lin. Bayesian regularized quantile regression. *Bayesian Anal.*, 5(3):533–556, 2010.

[37] D. Lorenz and H. Mahler. Orlicz space regularization of continuous optimal transport problems. *Appl. Math. Optim.*, 85(2):Paper No. 14, 33, 2022.

[38] D. Lorenz, P. Manns, and C. Meyer. Quadratically regularized optimal transport. *Appl. Math. Optim.*, 83(3):1919–1949, 2021.

[39] O. L. Mangasarian. Normal solutions of linear programs. *Math. Programming Stud.*, 22:206–216, 1984. Mathematical programming at Oberwolfach, II (Oberwolfach, 1983).

[40] O. L. Mangasarian and R. R. Meyer. Nonlinear perturbation of linear programs. *SIAM J. Control Optim.*, 17(6):745–752, 1979.

[41] G. Mordant. Regularised optimal self-transport is approximate Gaussian mixture maximum likelihood. *Preprint arXiv:2310.14851v1*, 2023.

[42] M. Nutz. Quadratically regularized optimal transport: Existence and multiplicity of potentials. *Preprint arXiv:2404.06847v1*, 2024.

[43] M. Nutz and J. Wiesel. Entropic optimal transport: convergence of potentials. *Probab. Theory Related Fields*, 184(1-2):401–424, 2022.

[44] S. Pal. On the difference between entropic cost and the optimal transport cost. *Ann. Appl. Probab.*, 34(1B):1003–1028, 2024.

[45] V. M. Panaretos and Y. Zemel. Statistical aspects of Wasserstein distances. *Annu. Rev. Stat. Appl.*, 6:405–431, 2019.

[46] G. Peyré and M. Cuturi. Computational optimal transport: With applications to data science. *Foun-*

dations and Trends in Machine Learning, 11(5-6):355–607, 2019.

[47] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.*, 40:99–121, 2000.

[48] V. Seguy, B. B. Damodaran, R. Flamary, N. Courty, A. Rolet, and M. Blondel. Large scale optimal transport and mapping estimation. In *International Conference on Learning Representations*, 2018.

[49] C. Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

[50] C. Villani. *Optimal transport, old and new*, volume 338 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, Berlin, 2009.

[51] J. Weed. An explicit analysis of the entropic penalty in linear programming. volume 75 of *Proceedings of Machine Learning Research*, pages 1841–1855, 2018.

[52] J. Wiesel and X. Xu. Sparsity of quadratically regularized optimal transport: Bounds on concentration and bias. *Preprint arXiv:2410.03425v1*, 2024.

[53] S. Zhang, G. Mordant, T. Matsumoto, and G. Schiebinger. Manifold learning with sparse regularised optimal transport. *Preprint arXiv:2307.09816v1*, 2023.

[54] L. Zhou, F. Koehler, D. J. Sutherland, and N. Srebro. Optimistic rates: A unifying theory for interpolation learning and regularization in linear regression. *ACM / IMS Journal of Data Science*, 1(2):1–51, 2024.