entropy

MDPI

Article

Generalizations of Talagrand Inequality for Sinkhorn Distance Using Entropy Power Inequality †

Shuchan Wang [1], Photios A. Stavrou [1] and Mikael Skoglund [2,*]

[1] Communication Systems Department, EURECOM, 06904 Sophia Antipolis, France; shuchan.wang@eurecom.fr (S.W.); fotios.stavrou@eurecom.fr (P.A.S.)
[2] Division of Information Science and Engineering, KTH Royal Institute of Technology, 114 28 Stockholm, Sweden
* Correspondence: skoglund@kth.se
† This paper is an extended version of our paper published in 2021 IEEE Information Theory Workshop (ITW), Kanazawa, Japan.

Abstract: The distance that compares the difference between two probability distributions plays a fundamental role in statistics and machine learning. Optimal transport (OT) theory provides a theoretical framework to study such distances. Recent advances in OT theory include a generalization of classical OT with an extra entropic constraint or regularization, called entropic OT. Despite its convenience in computation, entropic OT still lacks sufficient theoretical support. In this paper, we show that the quadratic cost in entropic OT can be upper-bounded using entropy power inequality (EPI)-type bounds. First, we prove an HWI-type inequality by making use of the infinitesimal displacement convexity of the OT map. Second, we derive two Talagrand-type inequalities using the saturation of EPI that corresponds to a numerical term in our expressions. These two new inequalities are shown to generalize two previous results obtained by Bolley et al. and Bai et al. Using the new Talagrand-type inequalities, we also show that the geometry observed by Sinkhorn distance is smoothed in the sense of measure concentration. Finally, we corroborate our results with various simulation studies.

Keywords: entropic optimal transport; Schrödinger problem; Talagrand inequality; entropy power inequality; log-concave measures

Citation: Wang, S.; Stavrou, P.A.; Skoglund, M. Generalizations of Talagrand Inequality for Sinkhorn Distance Using Entropy Power Inequality. *Entropy* **2022**, *24*, 306. https://doi.org/10.3390/e24020306

Academic Editor: Takuya Yamano

Received: 4 January 2022
Accepted: 17 February 2022
Published: 21 February 2022

1. Introduction

OT theory studies how to transport one measure to another in the path with minimal cost. The Wasserstein distance is the cost given by the optimal path and closely connected with information measures; see, e.g., [1–5].

During the last decade, OT has been studied and applied extensively, especially in the machine learning community; see, e.g., [6–9]. Entropic OT, a technique to approximate the solution of the original OT, was given for computational efficiency in [10]. A key concept in the entropic OT is the Sinkhorn distance, which is a generalization of the Wasserstein distance with an extra entropic constraint. Due to the extra entropic constraint in the domain of the optimization problem, randomness is added to the original deterministic system, and the total cost increases from the original Wasserstein distance to a larger value. Therefore, a natural question is how to quantify the extra cost caused by the entropic constraint.

In this paper, we derive upper bounds for the quadratic cost of entropic OT, which are shown to include a term of entropy power responsible for quantifying the amount of uncertainty caused by the entropic constraint. This work is an extended version of [11].

1.1. Literature Review

The dynamical formulation of OT, also known as the Benamou–Brenier formula [12], generalizes the original Monge–Kantorovich formulation into a time-dependent problem.

It changes the original distance problem (i.e., find the distance between two prescribed measures) into a geodesic problem (i.e., find the optimal path between two prescribed measures). Using the displacement convexity of relative entropy along the geodesic, functional inequalities such as HWI inequality and Talagrand inequality can be obtained (see, e.g., ([13] Chapter 20)).

Talagrand inequality, first given in [1], upper bounds the Wasserstein distance by relative entropy. Recent results in [2,4] obtain several refined Talagrand inequalities with dimensional improvements on the multidimensional Euclidean space. These inequalities bound Wasserstein distance with entropy power, which is sharper compared to the original one with relative entropy.

An analogue of the dynamical OT problem is the SP [14]. The SP aims to find the most likely evolution of a system of particles with respect to a reference process. The most likely evolution is called a Schrödinger bridge. SP and OT intersect on many occasions; see, e.g., [15–17]. The problem we study in this paper is in this intersection and mostly related to [15]. In particular, Léonard in [15] showed that the entropic OT with quadratic cost is equivalent to the SP with a Brownian motion as the reference process. He further derived that the Schrödinger bridge also admits a Benamou–Brenier formula with an additional diffusion term. Conforti in [18,19] claimed that the process can also be formulated as a continuity equation and proved that the acceleration of particles is the gradient of the Fisher information. The result therein leads to a generalized Talagrand inequality for relative entropy. Later, Bai et al. in [20] upper-bounded the extra cost from the Brownian motion by separating one Gaussian marginal into two independent random vectors. Using this approach, they showed that the dimensional improvement can be generalized to entropic OT and gave a Gaussian Talagrand inequality for the Sinkhorn distance. Additional results in [20] include a strong data processing inequality derived from their new Talagrand inequality and a bound on the capacity of the relay channel.

Entropic OT has other interesting properties. For example, Rigollet and Weed studied the case with one side of empirical measure in [21]. Their result shows that entropic OT performs maximum-likelihood estimation for Gaussian deconvolution of the empirical measure. This result can be further applied in uncoupled isotonic regression (see [9]). The dimensionality is also observed in the applications of entropic OT. For example, sample complexity bounds in [22,23] appear to be dimensional-dependent. In the GAN model, Reshetova et al. in [24] showed that the entropic regularization of OT promotes sparsity in the generated distribution.

Another element in our paper is EPI (for details on EPI, see, e.g., [25–27]). This inequality provides a clear expression to bound the differential entropy of two distributions' convolution. We refer the interested reader to [28–32] for the connections between EPI and functional inequalities, and [33] for the connections between EPI and SP.

1.2. Contributions

In this paper, we upper-bound the quadratic cost of entropic OT by deconvolution of one side measure and EPIs. Using this approach, we avoid any discussion related to the dynamics of SP and instead we capture the uncertainty caused by the Brownian motion quantitatively. Our contributions can be articulated as follows:

(1) We derive an HWI-type inequality for Sinkhorn distance using a modification of Bolley's proof in [4] (see Theorem 2).
(2) We prove two new Talagrand-type inequalities (see Theorems 3 and 4). These inequalities are obtained via a numerical term C related to the saturation, or the tightness, of EPI. We claim that this term can be computed with arbitrary deconvolution of one side marginal, while the optimal deconvolution is shown to be unknown beyond the Gaussian case. Nevertheless, we simulate suboptimally this term for a variety of distributions in Figure 1.
(3) We show that the geometry observed by Sinkhorn distance is smoothed in the sense of measure concentration. In other words, Sinkhorn distance implies a dimensional

measure concentration inequality following Marton's method (see Corollary 2). This inequality has a simple form of normal concentration that is related to the term C and is weaker than the one implied by Wasserstein distance.

(4) Our theoretical results are validated via numerical simulations (see Section 4). These simulations reveal several reasons for which our bounds can be either tight or loose.

Connections to Prior Art

The novelty of our work is that it comprises naturally ideas from Bolley et al. in [4] and from Bai et al. in [20] to develop new entropic OT inequalities. The dimensional improvement of Bolley et al. in [4] separates an independent term of entropy power from the original Talagrand inequality. This allows us to utilize an approach to study the entropic OT problem, which is the OT with randomness, based on the convolutional property of entropy power. On the other hand, we generalize the constructive proof of Bai et al. in [20], where they separate one Gaussian random vector into two independent Gaussian random vectors. We further claim that, for any distribution, we can always find similar independent pairs satisfying several assumptions, to upper-bound the Sinkhorn distance. As a consequence of the above, our results generalize the Talagrand inequalities of Bolley et al. in ([4] Theorem 2.1) from classical OT to entropic OT and the results of Bai et al. in ([20] Theorem 2.2) from the Gaussian case to the strongly log-concave case. In particular, we show that Theorem 3 recovers ([4] Theorem 2.1) (see Corollary 1 and the discussion in Remark 6) and that Theorem 4 recovers ([20] Theorem 2.2) (see Remark 9). It should be noted that in our analysis, we focus on the primal problem defined in [10], as opposed to the studies of its Lagrangian dual in [18,19].

1.3. Notation

$\mathbb{N}$ is the set of positive integers $\{1, 2, 3, ...\}$. $\mathbb{R}$ is the set of real numbers. $\mathbb{R}^n$ is the n-dimension Euclidean space. $\mathbb{R}_+$ denotes the set $\{x \in \mathbb{R} : x \geq 0\}$.

Let $\mathcal{X}, \mathcal{Y}$ be two Polish spaces, i.e., separable complete metric spaces. We write an element $x \in \mathcal{X}$ in lower-case letters and a random vector X on $\mathcal{X}$ in capital letters. We denote $\mathcal{P}(\mathcal{X})$ as the set of all probability measures on $\mathcal{X}$. Let μ be a Borel measure on $\mathcal{X}$. For a measurable map $T : \mathcal{X} \to \mathcal{Y}$, $T_{\#}\mu$ denotes the pushing forward of μ to $\mathcal{Y}$, i.e., for all $A \subset \mathcal{Y}$, $T_{\#}\mu[A] = \mu[T^{-1}(A)]$. For $p \geq 1$, $L^p(\mathcal{X})$ or $L^p(d\mu)$ denotes the Lebesgue space of p-th order for the reference measure μ.

∇ is the gradient operator, $\nabla \cdot$ is the divergence operator, Δ is the Laplacian operator, D^2 is the Hessian operator, I_n is the n-dimension identity matrix, Id is the identity map, $\| \cdot \|$ is the Euclidean norm, C^k is the set of functions that is k-times continuously differentiable, Ric is the Ricci curvature.

$h(\cdot)$, $I(\cdot;\cdot)$, $D(\cdot\|\cdot)$, $J(\cdot)$, $I(\cdot|\cdot)$ denote differential entropy, mutual information, relative entropy, Fisher information and relative Fisher information, respectively. All the logarithms are natural logarithms. $\exists!$ is unique existence. $*$ is the convolution operator.

1.4. Organization of the Paper

The rest of the paper is organized as follows: in Section 2, we give the technical preliminaries of the theories and tools that we use; in Section 3, we state our main theoretical results; in Section 4, we give numerical simulations for our theorems, and in Section 5, we give the conclusions and future directions. Long proofs and background material are included in the Appendix.

2. Preliminaries

In this section, we give an overview of the theories and tools that we use.

2.1. Synopsis of Optimal Transport

We first give a brief introduction of OT theory. The OT problem was initialized by Gaspard Monge. The original formulation can be described as follows.

Definition 1 (Monge Problem [34]). *Let P_X and P_Y be two probability measures supported on two Polish spaces $\mathcal{X}$, $\mathcal{Y}$. Given a lower semi-continuous (see Definition A1) cost function $c(x,y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+ \cup \{+\infty\}$, the Monge problem wants to find a transport map $T : \mathcal{X} \to \mathcal{Y}$ minimizing the total cost:*

$$\inf_{T:T_\# P_X = P_Y} \int_{\mathcal{X}} c(x, T(x))\, dP_X(x). \tag{1}$$

Then, Kantorovich gave a probabilistic interpretation to the OT. This is stated next.

Definition 2 (Kantorovich Problem [35]). *Let X and Y be two random vectors on two Polish spaces $\mathcal{X}$, $\mathcal{Y}$. X and Y have probability measures $P_X \in \mathcal{P}(\mathcal{X})$, $P_Y \in \mathcal{P}(\mathcal{Y})$. We denote $\Pi(P_X, P_Y)$ as the set of all joint probability measures on $\mathcal{X} \times \mathcal{Y}$ with marginal measures P_X, P_Y. Given a lower semi-continuous cost function $c(x,y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+ \cup \{+\infty\}$, the Kantorovich problem can be written as:*

$$\inf_{P \in \Pi(P_X, P_Y)} \int_{\mathcal{X} \times \mathcal{Y}} c(x,y)\, dP. \tag{2}$$

It can be further proven that (2) gives the same optimizer as (1) (see, e.g., [36]). One can define the Wasserstein distance ([13] Definition 6.1) from (2). Let $\mathcal{X} = \mathcal{Y}$ and let d be a metric on $\mathcal{X}$. Then, the Wasserstein distance of order p, $p \geq 1$, is defined as follows:

$$\mathcal{W}_p(P_X, P_Y) := \inf_{P \in \Pi(P_X, P_Y)} \left[\int_{\mathcal{X} \times \mathcal{Y}} d^p(x,y)\, dP \right]^{\frac{1}{p}}. \tag{3}$$

We note that the Wasserstein distance is a metric between two measures.

Cuturi in [10] gave the concept of entropic OT. In this definition, he adds an information theoretic constraint to (2), i.e.,

$$\inf_{P \in \Pi(P_X, P_Y; R)} \int_{\mathcal{X} \times \mathcal{Y}} c(x,y)\, dP, \tag{4}$$

where

$$\Pi(P_X, P_Y; R) := \{P \in \Pi(P_X, P_Y) : I(X;Y) \leq R\},$$

with $I(X;Y) := D(P \| P_X \times P_Y)$ denoting the mutual information [32] between X and Y, and $R \in \mathbb{R}_+$. It is well known that the constraint set is convex and compact with respect to the topology of weak convergence (for details, see, e.g., ([13] Lemma 4.4), ([37] Section 1.4)). Using the lower semi-continuity of $c(x,y)$ and ([13] Lemma 4.3), we know that the objective function $f : P \to \int c\, dP$ is also lower semi-continuous. Using the compactness of the constraint set and the lower semi-continuity of f, then, from Weierstrass' extreme value theorem, the minimum in (4) is attained. Moreover, the solution is always located on its boundary, i.e., $I(X;Y) = R$, because the objective function of (4) is linear.

Entropic OT is an efficient way to approximate solutions of the Kantorovich problem. The Lagrangian dual of (4), which was introduced by Cuturi in [10], can be solved iteratively. The dual problem of (4) can be reformulated as follows:

$$\max_{\epsilon \geq 0} \inf_{P \in \Pi(P_X, P_Y)} \left\{ \int_{\mathcal{X} \times \mathcal{Y}} c(x,y)\, dP + \epsilon \big(I(X;Y) - R \big) \right\}, \tag{5}$$

where ϵ is a Lagrange multiplier. Using the Lagrange duality theorem ([38] Theorem 1, pp. 224–225), it can be shown that (4) and (5) give the same optimizer P^*.

The uncertainty of entropic OT can be understood as follows. We can write $I(X;Y) = h(Y) - h(Y|X)$, where $h(Y)$ is fixed. The conditional entropy encapsulates the randomness of the conditional distribution. The randomness decreases when $I(X;Y)$ increases. Thus, unlike (1) and (2), there is no deterministic map anymore for (4) and (5), because a one-to-one mapping leads to infinite mutual information. Note that ϵ in (5) also has an explicit physical meaning. In particular, entropic OT with quadratic cost coincides with SP with a

reference measure of Brownian motion (see [15]). Then, ϵ is a diffusion coefficient of the Fokker–Planck equation associated with the Schrödinger bridge.

In our main results, we study (4) instead of (5) for two reasons. First, the mutual information in (4) gives a global description of the amount of uncertainty, while the coefficient ϵ in (5) and its associated Fokker–Planck equation are more related to local properties, from the definitions of the Lagrangian dual and Fokker–Planck equation. Further on this point, there is no explicit expression for the correspondence between R and ϵ in the duality. Second, the expectation of cost function in (2) is comparable to the Wasserstein distance. As we demonstrate in the following, it gives a smooth version of the Wasserstein distance.

Similar to the Wasserstein distance, the Sinkhorn distance of order p is defined as follows:

$$\mathcal{W}_p(P_X, P_Y; R) := \inf_{P \in \Pi(P_X, P_Y; R)} \left[\int_{\mathcal{X} \times \mathcal{Y}} d^p(x, y)\, dP \right]^{\frac{1}{p}}. \tag{6}$$

Clearly, $\Pi(P_X, P_Y; R)$ is a subset of $\Pi(P_X, P_Y)$. Because of the minimization problem, it is easy to see that $\mathcal{W}_p(P_X, P_Y; R) > \mathcal{W}_p(P_X, P_Y)$. For this reason, we say that entropic OT is a smoothed version of classical OT. We note that the Sinkhorn distance is not a metric because it does not fulfill the axiom of identity of indiscernibles.

Since entropic OT is concerned with mutual information, it may be of interest to introduce a conditional Sinkhorn distance. This is defined as follows:

$$\mathcal{W}_p(P_{X|Z}, P_{Y|Z}|P_Z; R) := \inf_{P \in \Pi(P_{X|Z}, P_{Y|Z}|P_Z); I(X;Y|Z) \le R} \{\mathbb{E}_P[d^p(X, Y)]\}^{1/p}, \tag{7}$$

where the conditional mutual information $I(X;Y|Z) := \int I(P_{X|Z=z}; P_{Y|Z=z})\, dP_Z(z)$ and $\Pi(P_{X|Z}, P_{Y|Z}|P_Z) := \{P_{X,Y|Z} \cdot P_Z : P_{X,Y|Z=z} \in \Pi(P_{X|Z=z}, P_{Y|Z=z}) \text{ for } z \text{ a.e.}\}$. Conditional Sinkhorn distance is utilized in [20] and leads to a data processing inequality. Since the constraint of conditional mutual information is a linear form of $I(P_{X|Z=z}; P_{Y|Z=z})$, the constraint set is still convex. The objective function is also a linear form of P. Therefore, the functional and topological properties of the conditional Sinkhorn distance are similar to the unconditional one.

Next, we state some known results of Talagrand inequality [1].

Definition 3 (Talagrand Inequality)**.** *Let P_X be a reference probability measure with density $e^{-V(x)}$, where $V : \mathcal{X} \to \mathbb{R}$. We say that P_X satisfies $\boldsymbol{T}(\lambda) > 0$, i.e., Talagrand inequality with parameter $\lambda > 0$, if, for any $P_Y \in \mathcal{P}(\mathcal{Y})$,*

$$\mathcal{W}_2(P_X, P_Y) \le \sqrt{\frac{2}{\lambda} D(P_Y \| P_X)}. \tag{8}$$

Remark 1. (8) *was originally introduced by Talagrand in [1] when P_X is Gaussian. Blower in [39] gave a refinement and proved that*

$$D^2 V \ge \lambda I_n \quad \Rightarrow \quad \boldsymbol{T}(\lambda).$$

When going beyond the Euclidean space to a manifold, Otto and Villani in [40] showed that the Bakry–Emery condition $D^2 V + \mathrm{Ric}$ also implies $\boldsymbol{T}(\lambda)$.

Recently, refined inequalities with dimensional improvements were obtained in multidimensional Euclidean space. These dimensional improvements were first observed in the Gaussian case of logarithmic Sobolev inequality, Brascamp–Lieb (or Poincaré) inequality [41] and Talagrand inequality [2]. For a standard Gaussian measure P_X, the dimensional Talagrand inequality has the form:

$$\mathcal{W}_2^2(P_X, P_Y) \le \mathbb{E}[\|Y\|^2] + n - 2ne^{\frac{1}{2n}(\mathbb{E}[\|Y\|^2] - n - 2D(P_Y \| P_X))}. \tag{9}$$

Bolley et al. in [4] generalized the results in [2,41] from Gaussian to strongly log-concave or log-concave. Next, we state their result. Let $dP_X = e^{-V}$, where $V : \mathbb{R}^n \to \mathbb{R}$ is C^2 continuous, $D^2V \geq \lambda I_n$. Bolley's dimensional Talagrand inequality is given as follows:

$$\frac{\lambda}{2}\mathcal{W}_2^2(P_X, P_Y) \leq \mathbb{E}[V(Y)] - \mathbb{E}[V(X)] + n - ne^{\frac{1}{n}(\mathbb{E}[V(Y)]-\mathbb{E}[V(X)]-D(P_Y\|P_X))}. \tag{10}$$

The dimensional Talagrand inequalities (9) and (10) are tighter than (8). To see this result, one may refer to our Remark 6 below.

Bai et al. in [20] gave a generalization of (9) to Sinkhorn distance. When P_X is standard Gaussian,

$$\mathcal{W}_2^2(P_X, P_Y; R) \leq \mathbb{E}[\|Y\|^2] + n - 2n\sqrt{\frac{1}{2\pi e}(1 - e^{-\frac{2}{n}R})}e^{\frac{1}{n}h(Y)}. \tag{11}$$

When $R \to +\infty$, this inequality coincides with (9).

2.2. Measure Concentration

The measure concentration phenomenon describes how the probability of a random variable X changes with the deviation from a given value such as its mean or median. Marton introduced an approach of concentration directly on the level of probability measures using OT (see, e.g., ([13] Chapter 22)).

To give the notation of concentration of measure, we first introduce the probability metric space. Let $\mathcal{X}$ be a Polish space. Let d be a metric on $\mathcal{X}$. Let μ be a probability measure defined on the Borel set of $\mathcal{X}$. Then, we say that the triple $(\mathcal{X}, d, \mu)$ is a probability metric space.

For an arbitrary set $A \subset \mathcal{X}$ and any $r \geq 0$, we define A_r as

$$A_r := \{x \in \mathcal{X} : d(x, A) > r\},$$

where $d(x, A) := \inf_{a\in A} d(x, a)$. Then, we say that a probability measure μ has normal (or Gaussian) concentration on $(\mathcal{X}, d)$ if there exists positive K and κ such that

$$\mu(A) \geq \frac{1}{2} \quad \Rightarrow \quad \mu(A_r) \leq Ke^{-\kappa r^2}, \forall r > 0. \tag{12}$$

There is another weaker statement of normal concentration, such that

$$\mu(A) \geq \frac{1}{2} \quad \Rightarrow \quad \mu(A_r) \leq Ke^{-\kappa(r-r_0)^2}, \forall r > r_0. \tag{13}$$

It is not difficult to see that (12) can be obtained from (13), possibly with degraded constants, i.e., larger K and/or smaller κ.

The next theorem gives the connection between normal concentration and Talagrand inequality.

Theorem 1 (Theorem 3.4.7 [5])**.** *Let $(\mathcal{X}, d, \mu)$ be a probability metric space. Then, the following two statements are equivalent:*

- *μ satisfies $\boldsymbol{T}(\lambda)$.*
- *μ has a dimension-free normal concentration with $\kappa = \frac{1}{2\lambda}$.*

The intuition behind Marton's method is that OT theory can give a metric between two probability measures by the metric structure of the supporting Polish space. The metric can be further connected with probability divergence using Talagrand inequality.

2.3. Entropy Power Inequality and Deconvolution

EPI [25] states that, for all independent continuous random vectors X and Y,

$$N(X+Y) \geq N(X) + N(Y), \tag{14}$$

where $N(X) := \frac{1}{2\pi e} e^{\frac{2}{n}h(X)}$ denotes the entropy power of X. The equality is achieved when X and Y are Gaussian random vectors with proportional covariance matrices.

Deconvolution is a problem of estimating the distribution $f(x)$ by the observations Y_1,...,Y_k corrupted by additive noise Z_1,...,Z_k, written as

$$Y_i = X_i + Z_i,$$

where $k, i \in \mathbb{N}$ and $1 \leq i \leq k$. X_i's are i.i.d. in $f(x)$, Z_i's are i.i.d in $h(z)$. X_i's and Z_i's are mutually independent. Let $g(y)$ be the probability density function of Y that is given by the convolution $g = f * h$. Then, their entropies can be bounded by EPI directly.

In our problem, we slightly abuse the concept by simply separating a random vector Y into two independent random vectors X and Z. We use this approach to introduce the uncertainty to entropic OT and consequently bound the Sinkhorn distance by EPI. Deconvolution is generally a more challenging problem than convolution. For instance, the log-concave family is convolution stable, i.e., convolution of two log-concave distributions is still log-concave, but we cannot guarantee that the deconvolution of two log-concave distributions is still log-concave. A trivial case is that wherein the deconvolution of a log-concave distribution by itself is a Dirac function. Moreover, f may not in general be positive or integrable for arbitrary given g and h, as shown in [42]. However, it should be noted that there are many numerical methods to compute deconvolution; see, e.g., [42–44].

3. Main Theoretical Results

In this section, we derive our main theoretical results. First, we give a new HWI-type inequality.

Theorem 2 (HWI-Type Inequality)**.** *Let $\mathcal{X} = \mathcal{Y} = \mathbb{R}^n$. Let μ be a probability measure with density $e^{-V(x)}$ with $\lambda > 0$, where $V : \mathbb{R}^n \to \mathbb{R}$ is C^2 continuous, $D^2V \geq \lambda I_n$. Let P_X, P_Y be two probability measures on $\mathbb{R}^n$, $P_X, P_Y \ll \mu$. For any independent Y_1, Y_2 satisfying $Y_1 + Y_2 = Y$, $\mathbb{E}[Y_2] = 0$ and $h(Y) - h(Y_2) \leq R$, the following bound holds:*

$$\frac{\lambda}{2}\mathcal{W}_2^2(P_X, P_Y; R) \leq \mathbb{E}[V(Y)] - \mathbb{E}[V(X)] + n - n\, e^{\frac{1}{n}(h(Y_1)-h(X))} + \mathcal{W}_2(P_X, P_{Y_1})\sqrt{I(P_X|\mu)}, \tag{15}$$

where the relative Fisher information $I(P_X|\mu) := \int \frac{\|\nabla f\|^2}{f} d\mu, f = \frac{dP_X}{d\mu}$.

Proof. See Appendix A. □

Remark 2. *In Theorem 2, we construct Y_1 and Y_2, where Y_1 and X have a deterministic relationship. We note that the uncertainty in our construction, i.e., the independent Y_2, is located at one marginal, whereas the uncertainty of the true dynamics of the entropic OT is all along the path. The simplicity of our construction allows for the specific bound. We further note that there always exist such Y_1, Y_2 satisfying the assumptions given in Theorem 2. A trivial proof is that $Y_1 = \mathbb{E}[Y]$ and $Y_2 = Y - \mathbb{E}[Y]$ fulfill the above assumptions.*

The next result gives a new Talagrand-type inequality.

Theorem 3 (Talagrand-Type Inequality). *Let $\mathcal{X} = \mathcal{Y} = \mathbb{R}^n$. Let $dP_X = e^{-V(x)}dx$, where $V : \mathbb{R}^n \to \mathbb{R}$ is C^2 continuous, $D^2V \geq \lambda I_n$ with $\lambda > 0$, $P_Y \ll P_X$. Then, the following bound holds:*

$$\frac{\lambda}{2}\mathcal{W}_2^2(P_X, P_Y; R) \leq \mathbb{E}[V(Y)] - \mathbb{E}[V(X)] + n - nC(P_Y, R)e^{\frac{1}{n}(h(Y)-h(X))}, \tag{16}$$

where $C(P_Y, R) \in [0, 1]$ is a numerical term for the given P_Y and $R \geq 0$.

Proof. Let $dP_X = e^{-V}$ in (15). In such a case, we have $I(P_X|\mu) = 0$ from the definition of relative Fisher information. Take $C(P_Y, R) = e^{\frac{1}{n}(h(Y_1)-h(Y))}$; then, (16) is proven from (15). □

Next, we state some technical remarks on Theorem 3.

Remark 3 (On Theorem 3). *In Theorem 3, we show that the Sinkhorn distance of two random vectors can be upper-bounded by a difference of a functional on two marginals, i.e., $\mathbb{E}[V(Y)] - \mathbb{E}[V(X)]$, and a term related to entropy power, i.e., $nC(P_Y, R)e^{\frac{1}{n}(h(Y)-h(X))}$. Interestingly, only the latter term is related to the constraint R. This means that the effect of information constraint is only associated with the randomness of the two random vectors instead of their positions. Recalling the physical meaning of entropic OT with quadratic cost, we can see that this expression is very natural, because the information constraint R is directly related to the randomness of the Schrödinger bridge.*

Remark 4 (On the numerical term $C(\cdot, \cdot)$). *The numerical term $C = e^{\frac{1}{n}(h(Y_1)-h(Y))}$ can be computed by arbitrary Y_1 satisfying the assumptions that we gave in Theorem 2, i.e., Y_1, Y_2 are independent, $Y_1 + Y_2 = Y$, $\mathbb{E}[Y_2] = 0$ and $h(Y) - h(Y_2) \leq R$. We observe that $e^{\frac{1}{n}h(Y_1)}$ has the form of a square root of entropy power. Using EPI and the fact that $N(\cdot) \geq 0$, we have*

$$N(Y) \geq N(Y_1) + N(Y_2) \geq N(Y_1).$$

Therefore, $C = e^{\frac{1}{n}(h(Y_1)-h(Y))} = \sqrt{N(Y_1)/N(Y)} \in [0, 1]$. When $R = 0$, then $Y_2 = Y - \mathbb{E}[Y]$ and the density of Y_1 is $\delta(x - \mathbb{E}[Y])$. This means that $e^{\frac{1}{n}h(Y_1)} = 0$, hence $C = 0$. When $R = +\infty$, then $Y = Y_1$, $e^{\frac{1}{n}h(Y_1)} = e^{\frac{1}{n}h(Y)}$, and consequently $C = 1$. Therefore, $C(\cdot, 0) = 0$, $C(\cdot, +\infty) = 1$ for all P_Y.

Moreover, we can show that there always exists such a sequence C non-decreasing with respect to R. We know that $C = e^{\frac{h(Y_1)}{n} - \frac{h(Y)}{n}}$ subject to $Y_1 + Y_2 = Y$, $\mathbb{E}[Y_2] = 0$ and $h(Y) - h(Y_2) \leq R$. Thus, for a larger value of R, there exists at least a $h(Y_2)$ non-increasing. This further leads to a non-decreasing $h(Y_1)$. Therefore, there exists at least a $C(\cdot, R + \Delta R)$ not smaller than $C(\cdot, R)$, $\forall \Delta R > 0$, i.e., $C(\cdot, R)$ is monotonic non-decreasing with respect to R.

We note that, for particular distributions, we may have an explicit expression of $C(\cdot, \cdot)$. For instance, when P_Y is Gaussian, we can always take the linear combination $Y = Y_1 + Y_2$, where Y_1 and Y_2 are independent Gaussian and have proportional covariance matrices. In such a case, EPI is saturated as follows:

$$e^{\frac{2}{n}h(Y_1)} = e^{\frac{2}{n}h(Y)} - e^{\frac{2}{n}h(Y_2)} = (1 - e^{-\frac{2}{n}R})e^{\frac{2}{n}h(Y)}.$$

As a result, we have $C(P_Y, R) = e^{\frac{1}{n}(h(Y_1)-h(Y))} = \sqrt{1 - R^{\frac{2}{n}}}$. For Cauchy distribution $Cauchy(x_0, \gamma)$, its differential entropy is $\log(4\pi\gamma)$. The summation of independent Cauchy random variables $\sum_i^n Cauchy(x_i, \gamma_i) \sim Cauchy(\sum_i^n x_i, \sum_i^n \gamma_i)$. When Y is i.i.d. Cauchy, i.e., $(Y)_i \sim Cauchy(x_0, \gamma)$, we take $(Y_1)_i \sim Cauchy(x_0, \frac{1}{4\pi}e^{\frac{1}{n}h(Y)} \cdot (1 - e^{-\frac{R}{n}}))$ and $(Y_2)_i \sim Cauchy(0, \frac{1}{4\pi}e^{\frac{1}{n}h(Y)} \cdot e^{-\frac{R}{n}})$. We can see that this linear combination satisfies our assumption $h(Y) - h(Y_2) \leq R$ and $C(P_Y, R) = e^{\frac{1}{n}(h(Y_1)-h(Y))} = 1 - e^{-\frac{R}{n}}$.

Note that the linear combination $Y = Y_1 + Y_2$ is not unique, according to the assumption of Theorem 2. Consequently, this implies the non-uniqueness of $C(\cdot,\cdot)$. In order to obtain the tightest bound in (16), we need to solve the following optimization problem

$$C^*(P_Y, R) = \sup e^{\frac{1}{n}(h(Y_1)-h(Y))}, \tag{17}$$

subject to $Y_1 + Y_2 = Y$ and $h(Y) - h(Y_2) \leq R$. To look into this optimization problem, we recall Courtade's reverse EPI ([31] Corollary 1) as follows. If we have independent X and Y with finite second moments and choose θ to satisfy $\theta/(1-\theta) = N(Y)/N(X)$, then

$$N(X+Y) \leq (N(X) + N(Y))(\theta p(X) + (1-\theta)p(Y)), \tag{18}$$

where $p(X) := \frac{1}{n}N(X)J(X) \geq 1$ is the Stam defect and $J(X) := I(P_X|\mu)$, where $d\mu = dx$ is the Fisher information. We note that $p(X)$ is affine invariant, i.e., $p(X) = p(tX)$, $t > 0$ because $t^2N(X) = N(tX)$ and $t^2J(tX) = J(X)$. We note that the equality $p(X) = 1$ holds only if X is Gaussian. In our case, $\theta = N(Y_1)/(N(Y_1) + N(Y_2))$. When $\theta \to 1$, (18) becomes

$$N(Y) \lesssim (N(Y_1) + N(Y_2)) \cdot p(Y_2).$$

This means that the saturation of EPI is controlled by $p(Y_2)$ when the noise Y_2 is small, i.e., when R is large. In such a case, $C^*(P_Y, R) \approx \sqrt{1 - R^{\frac{2}{n}}}$ if we let Y_2 be close to Gaussian, i.e., $p(Y_2) = 1$. On the other hand, when $\theta \to 0$, EPI can also be saturated if we let Y_1 be close to Gaussian.

In Figure 1, we illustrate numerical simulations of $C(\cdot,\cdot)$ for the one-dimensional case. For general distributions beyond Gaussian and i.i.d. Cauchy, one can approximate $C(\cdot,\cdot)$ using kernel methods of deconvolution; see, e.g., [43,44]. Our strategy of deconvolution in Figure 1 is to let $Y_2 = tY'$, where Y' is a copy of Y and $t \in [0,1]$. Gaussian mixture is an exception for this strategy because its spectrum would not be integrable. Instead, we let Y_2 be Gaussian for a Gaussian mixture. We note that this strategy is mostly not optimal and the optimal way to maximize the entropy power in (17) remains an open question.

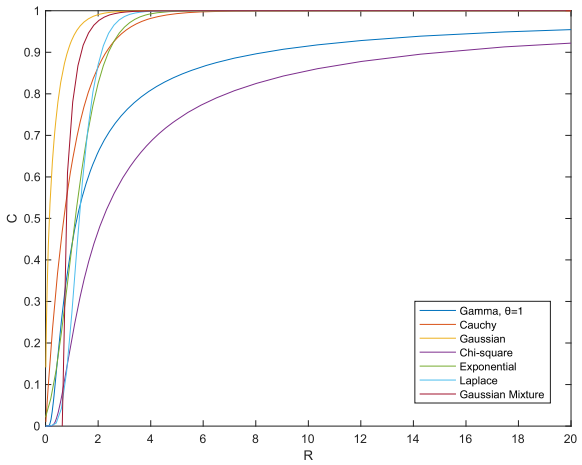


Figure 1. Plot of the numerical term C subject to the information constraint R evaluated with respect to different distributions for the one-dimensional case.

Remark 5 (On the condition of identity of Theorem 3)**.** *To show the condition of identity of (16), we need the inequalities in (A1) and in Lemma A1 to be equalities. The equality of (A1) holds when P_X is isotropic Gaussian, i.e., $P_X \sim \mathcal{N}(\mu, \sigma^2 I_n)$ for some $\mu \in \mathbb{R}^n$ and $\sigma > 0$. The equality in Lemma A1 holds when $\nabla\varphi$ is affine and $D^2\varphi$ has identical eigenvalues, i.e., $\nabla\varphi = k \cdot Id$, $k \in \mathbb{R}$, see ([4] Lemma 2.6). From ([45] Theorem 1), we know that the linear combination $Y = Y_1 + Y_2$ in Theorem 2 is the optimizer for entropic OT when X and Y are isotropic Gaussian. In such a case, the equality of (16) holds and $C(\cdot, R) = \sqrt{1 - R^{\frac{2}{n}}}$.*

The following corollary is immediate from Theorem 3.

Corollary 1. *Wasserstein distance is bounded by*

$$\frac{\lambda}{2}\mathcal{W}_2^2(P_X, P_Y) \leq \mathbb{E}[V(Y)] - \mathbb{E}[V(X)] + n - ne^{\frac{1}{n}(h(Y)-h(X))}. \tag{19}$$

Proof. This is immediate from Theorem 3 when $R \to \infty$. In this case, $C(\cdot, +\infty) = 1$. □

Remark 6 (On Corollary 1)**.** *We note that (19) is equivalent to Bolley's dimensional Talagrand inequality (10) and it is tighter than the classical Talagrand inequality (8). To make this point clear, note that, under our assumptions, $h(X) = \mathbb{E}[V(X)]$ and $D(P_Y\|P_X) = \mathbb{E}[V(Y)] - h(Y)$ because $dP_X = e^{-V(x)}dx$. Clearly, by substituting these expressions to the last term of (19), we obtain (10). Since $e^{\mu} \geq 1 + \mu$, (10) is, in general, tighter than the classical Talagrand inequality (8), i.e., RHS of (10) $\leq$ RHS of (8). The equality holds if and only if $h(Y) = h(X)$.*

We notice that C is the only difference between (10) and (16), from Remark 6. Therefore, we can immediately obtain a result related to measure concentration following ([4] Corollary 2.4). Next, we state the result on measure concentration obtained from (16).

Corollary 2. *Let $d\mu = e^{-V}$, where $V : \mathbb{R}^n \to \mathbb{R}$ is C^2 continuous, $D^2V \geq \lambda I_n$ with $\lambda > 0$. Let $A \subset \mathbb{R}^n$, $A_r := \{x \in \mathbb{R}^n : \forall y \in A, \|x - y\| > r\}$ for $r \geq 0$ and $c_A := \sqrt{2\lambda^{-1}\log(1/\mu(A))}$. Then, for $r \geq c_A$, we obtain*

$$\mu(A_r) \leq C^{-n} \cdot e^{-\frac{\lambda}{2}(r-c_A)^2}. \tag{20}$$

Proof. See Appendix B. □

Next, we state some technical comments on Corollary 2.

Remark 7 (On Corollary 2)**.** *We note that in the derivation of Corollary 2, we follow the method of Marton in [46], which utilizes the geometrical properties of Wasserstein distance. From our discussion above, the information constraint leads to the uncertainty in entropic OT. In this result, we further show that the uncertainty smooths the geometrical properties of Wasserstein distance, i.e., Sinkhorn distance implies a looser measure concentration inequality. We begin with two extreme cases. When $C = 0$, the two random vectors are independent and entropic OT has the most uncertainty. It is natural that the quadratic difference of two independent random vectors does not imply any concentration. When $C = 1$, the inequality is the same as the one in Theorem 1. Between these two extremes, i.e., when $0 < C < 1$, Sinkhorn distance leads to a weaker normal concentration, compared to Theorem 1. Furthermore, we include in Appendix C the proof that demonstrates that the Sinkhorn distance gives a weaker measure concentration inequality in high dimensions.*

The next theorem is another Talagrand-type inequality. Compared to Theorem 3, the following result is a bound obtained using a term related to the saturation of P_X, instead of the saturation of P_Y that was used in Theorem 3.

Theorem 4. *Let $\mathcal{X} = \mathcal{Y} = \mathbb{R}^n$. Without loss of generality, let X be a zero-mean random vector with density $e^{-V(x)}$, where $V : \mathbb{R}^n \to \mathbb{R}$ is C^2 continuous, $D^2V \geq \lambda I_n$ with $\lambda > 0$, $P_Y \ll P_X$. Then, the following bound holds:*

$$\frac{\lambda}{2}\mathcal{W}_2^2(P_X, P_Y; R) \leq \mathbb{E}[V(Y)] - \mathbb{E}[V(X)] + n - nC_x(P_X, R)e^{\frac{1}{n}(h(Y)-h(X))} + \epsilon', \tag{21}$$

where ϵ' is a term related to the linearity of V.

Proof. See Appendix D. □

We offer the following technical comments on Theorem 4.

Remark 8 (On Theorem 4). *Similar to $C(P_Y, R)$, $C_x(P_X, R) \in [0,1]$ can be computed by the equation $X = C_x X' + X_2$ with arbitrary independent X', X_2 under the assumptions that X' is a copy of X, $\mathbb{E}[X_2] = 0$, $h(X) - h(X_2) \leq R$, as shown in the proof. However, (21) is less natural than (16) because of the extra term ϵ'. When ∇V is nearly linear, ϵ' should be small. When ∇V is far from linear, ϵ' is unknown.*

Theorem 4 can also give a measure concentration inequality, namely

$$\mu(A_r) \leq C_x^{-n} \cdot e^{-\frac{\lambda}{2}(r-c_A)^2+\epsilon'}, \tag{22}$$

where A_r and c_A are the same as those defined in Corollary 2. We omit the proof of (22) because it follows using similar steps to the ones used to prove Corollary 2.

Remark 9. *When ∇V is linear, ϵ' is zero and $C_x(\cdot, R) = \sqrt{1 - R^{\frac{2}{n}}}$, as simply taking $t = \sqrt{1 - R^{\frac{2}{n}}}$ in the proof. In such a case, (21) recovers (11) by taking X as a standard Gaussian, i.e., $V(x) = \|x\|^2/2 + k$, where k is a normalization factor. Substitute V and times 2 on both sides of (21), we have*

$$\begin{aligned} \mathcal{W}_2^2(X, Y; R) \leq & \mathbb{E}[\|Y\|^2] - \mathbb{E}[\|X\|^2] + 2n - 2n\sqrt{1 - e^{-\frac{2}{n}R}} e^{\frac{1}{n}(h(Y)-h(X))} \\ = & \mathbb{E}[\|Y\|^2] + n - 2n\sqrt{\frac{1}{2\pi e}(1 - e^{-\frac{2}{n}R})} e^{\frac{1}{n}h(Y)}, \end{aligned} \tag{23}$$

which is exactly the same as (11).

The next theorem gives a Talagrand-type bound for the conditional Sinkhorn distance.

Theorem 5 (Talagrand-type bound for conditional Sinkhorn distance). *Let $\mathcal{X} = \mathcal{Y} = \mathbb{R}^n$. Given a probability measure P_Z and two conditional probability measures $P_{X|Z}$ and $P_{Y|Z}$, where the probability density $dP_{X|Z=z_0} = dP_X = e^{-V(x)}dx, \forall z_0 \in \mathcal{Z}$, let $V : \mathbb{R}^n \to \mathbb{R}$ be C^2 continuous, $D^2V \geq \lambda I_n$ with $\lambda > 0$, $P_{Y|Z=z_0} \ll P_X$. Then, the following bound holds:*

$$\frac{\lambda}{2}\mathcal{W}_2^2(P_X, P_Y|P_Z; R) \leq \mathbb{E}[V(Y|Z)] - \mathbb{E}[V(X)] + n - nC'(P_{Y|Z}, R)e^{\frac{1}{n}(h(Y|Z)-h(X))}, \tag{24}$$

where $C'(P_{Y|Z}, R)$ is a numerical term. A direct observation is that $C'(P_{Y|Z}, R)$ can take $\inf_{z_0 \in \mathcal{Z}} C(P_{Y|Z=z_0}, R)$.

Proof. See Appendix E. □

4. Numerical Simulations

In this section, we describe several numerical simulations to illustrate the validity of our theoretical findings. To check the tightness of our bounds, we use as a reference bound the numerical solution obtained via the Sinkhorn algorithm, which can be found in the POT library [47]. As an iterative method, the Sinkhorn algorithm has computational error, since the iteration stops when it converges to a certain rate. For example, in Figure 2a, we plot the result for Theorem 3 with two Gaussian marginals, which is the scenario when the identity of (16) holds. From the figure, we can see that the simulation is slightly greater than the bound. Nevertheless, we note that the precision is reasonably small.

The simulations for Theorems 3 and 4 are given in Figures 2 and 3, respectively. Since the optimal value of C in Theorem 3 and the error term ϵ' in Theorem 4 beyond the linear case are unknown, we mainly simulate the case with one side Gaussian, i.e., with $C = \sqrt{1 - R^{\frac{2}{n}}}$ and $\epsilon' = 0$. In this way, we avoid the unknown factors and deduce several observations related to the tightness of the bounds derived in these two theorems.

The first observation is about absolute continuity. We observe that the original Talagrand inequality (8) is not tight when P_Y is not absolutely continuous with respect to P_X,

because $D(P_Y||P_X) = +\infty$ in this case. In Figure 4, we illustrate one such case with almost discontinuity between two strongly log-concave distributions, i.e., the Radon–Nikodym derivative $dP_Y/dP_X = \exp[(x/5)^4] + k'$, where $k' \in \mathbb{R}$ is a normalizing factor, goes to ∞ when $x \to \infty$. Consequently, the bound (16) from Theorem 3 is loose, as illustrated in Figure 5. The bound can be much looser if we increase the discontinuity, i.e., we let $dP_Y/dP_X = \exp[(x/5)^8] + k'$, as shown in Figure 6. By simply changing the sides of distributions P_X and P_Y, we preserve the absolute continuity and the bound becomes tight, as we can see in Figures 5b and 6b.

The second observation is related to the numerical term C. By comparing Figures 2b and 3a, we observe that $C = \sqrt{1 - R^{\frac{2}{n}}}$ gives a better description than $C = 1 - R^{\frac{1}{n}}$, i.e., the former one gives a tighter bound. This is reasonable according to our previous discussion, i.e., the independent linear combination of Cauchy random variables is not the optimal deconvolution. Actually, even if P_Y is not Gaussian in (16), $C = \sqrt{1 - R^{\frac{2}{n}}}$ seems to be true for all the simulated distributions.

Furthermore, we observe that the tightness of the bounds in Theorems 3 and 4 is related to the linearity of the transport map, which can be seen as a similarity between the two marginals. For example, Cauchy and Laplace distributions are similar to the Gaussian distribution. Thus, they show a tight bound in Figure 3a,d. On the other hand, Gaussian mixture and exponential distribution are relatively far from the Gaussian distribution. Hence, Figure 3c,f give looser bounds.

In Figure 7, we plot the dimensionality of Sinkhorn distance between isotropic Gaussians. Different curves correspond to a pair of Gaussian distributions in different dimensions and these pairs have the same Wasserstein distance. It can be seen that the information constraint causes more smoothing in higher dimensions, which is consistent with Corollary 2.

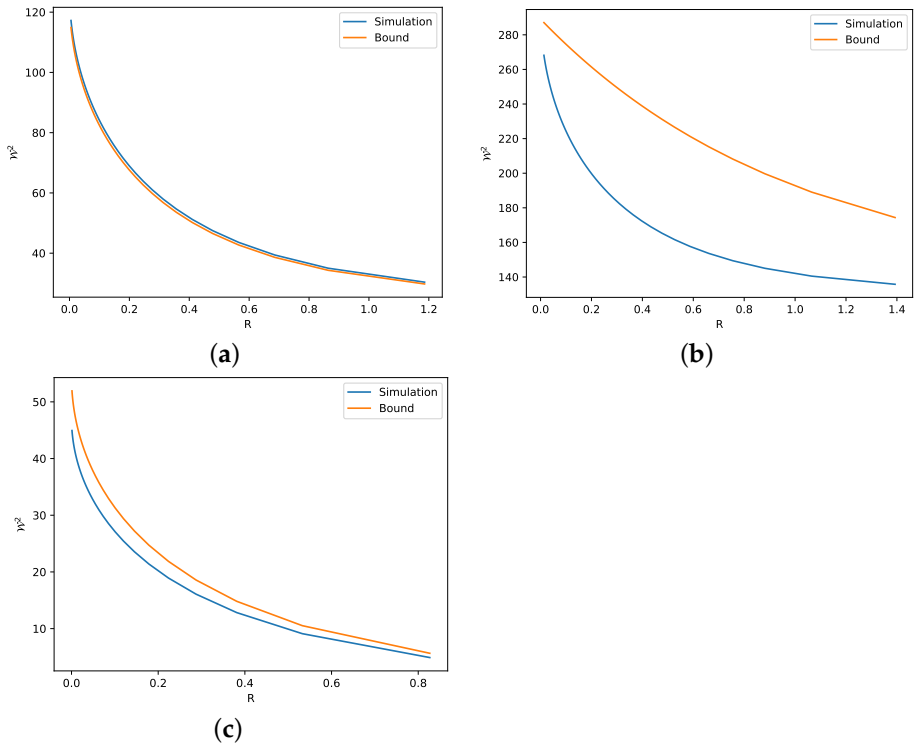


Figure 2. Numerical simulations and bounds via (16) for different R. (**a**) $dP_X \sim \mathcal{N}(0, \frac{1}{25})$ and $dP_Y \sim \mathcal{N}(0, \frac{1}{100})$. (**b**) $dP_X \sim \mathcal{N}(0, \frac{1}{25})$ and $dP_Y \sim Cauchy(0, 10)$. (**c**) $dP_X = e^{-V}, V = (x/5)^2/2 + |x/10| + e^{-|x/10|} + k, k \in \mathbb{R}$ and $dP_Y \sim \mathcal{N}(0, \frac{1}{25})$.

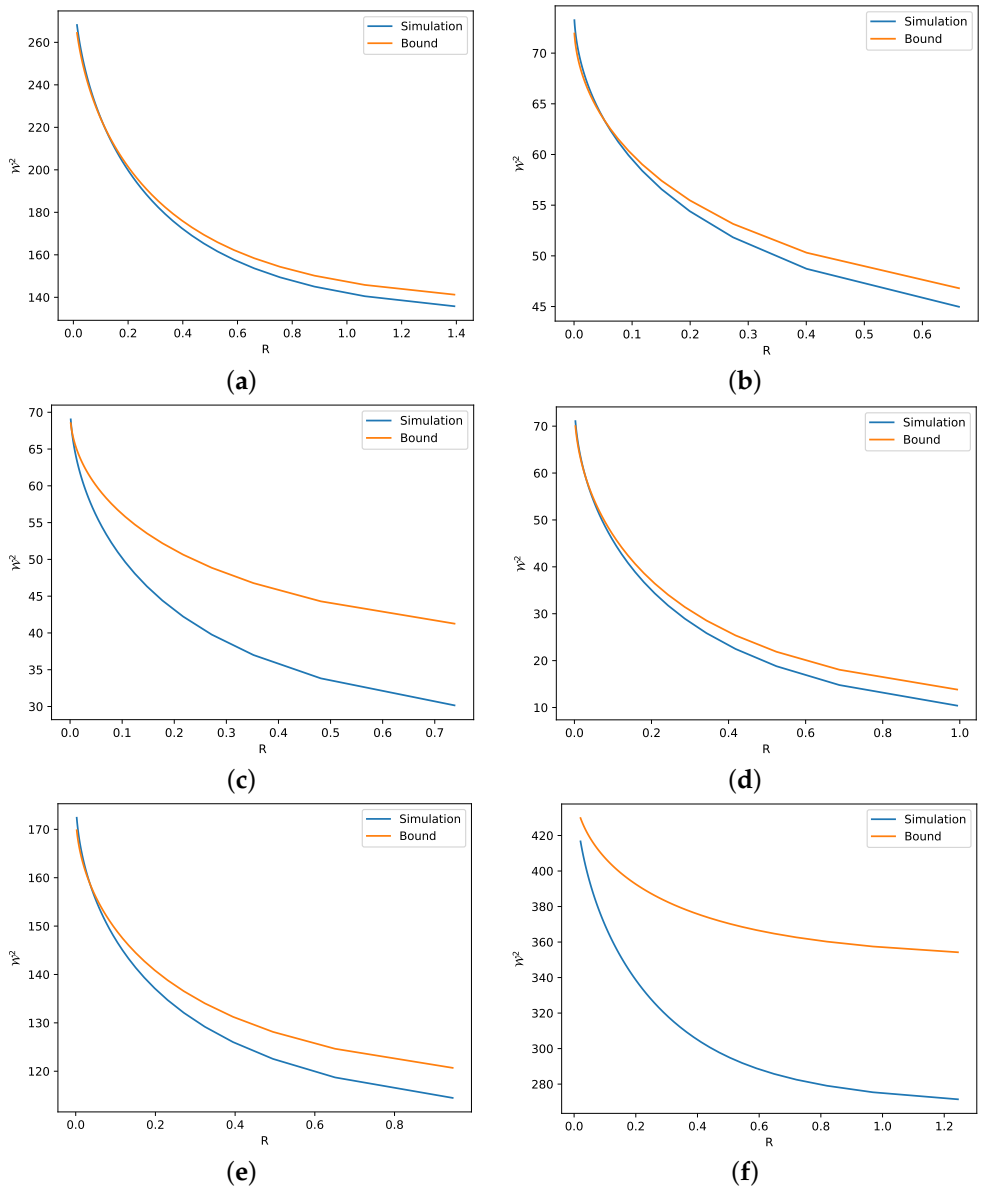


Figure 3. Numerical simulations and bounds via (21) for different R. (**a**) $dP_X \sim \mathcal{N}(0, \frac{1}{25})$ and $dP_Y \sim Cauchy(0, 10)$. (**b**) $dP_X \sim \mathcal{N}(0, \frac{1}{25})$ and $dP_Y \sim \chi^2(6)$. (**c**) $dP_X \sim \mathcal{N}(0, \frac{1}{25})$ and $dP_Y \sim Exp(0.2)$. (**d**) $dP_X \sim \mathcal{N}(0, \frac{1}{25})$ and $dP_Y \sim Laplace(0, 5)$. (**e**) $dP_X \sim \mathcal{N}(0, \frac{1}{25})$ and dP_Y is Gamma distribution with $\alpha = 2$ and $\beta = 0.2$. (**f**) $dP_X \sim \mathcal{N}(0, \frac{1}{25})$ and $dP_Y = \frac{1}{2}\mathcal{N}(-20, \frac{1}{25}) + \frac{1}{2}\mathcal{N}(20, \frac{1}{25})$.

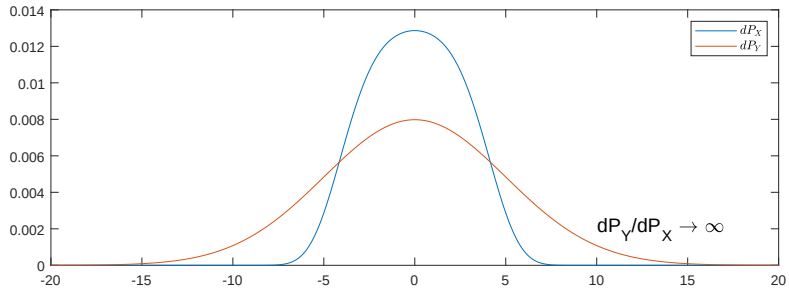
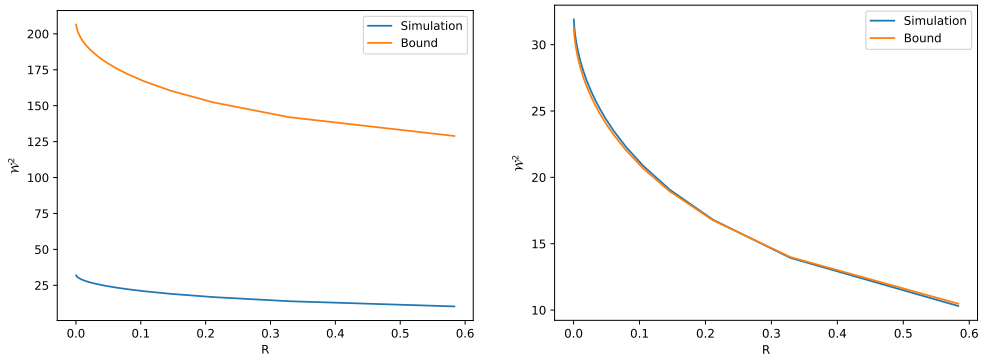
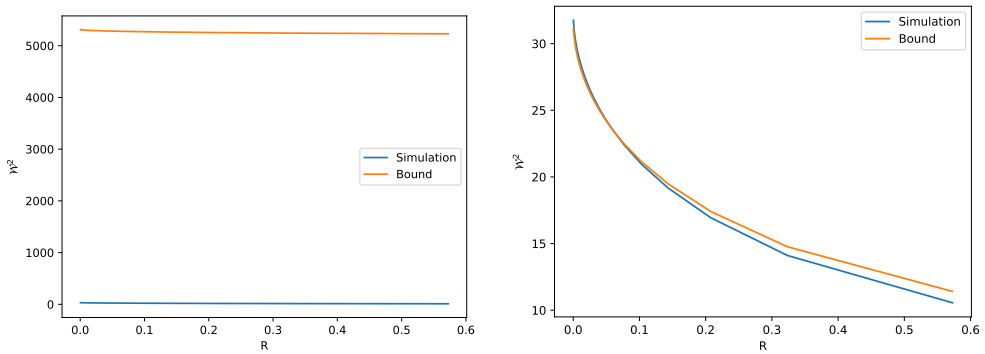


Figure 4. Probability densities of $dP_X = e^{-V}, V = (x/5)^2/2 + (x/5)^4 + k, k \in \mathbb{R}$ and $dP_Y \sim \mathcal{N}(0, \frac{1}{25})$.



(**a**) (**b**)

Figure 5. Numerical simulations and bounds for different R, with $d\mu = e^{-V}, V = (x/5)^2/2 + (x/5)^4 + k, k \in \mathbb{R}$ and $d\mu \sim \mathcal{N}(0, \frac{1}{25})$. (**a**) Bound via (16) with $dP_X = d\mu, dP_Y = d\nu$. (**b**) Bound via (21) with $dP_X = d\nu, dP_Y = d\mu$.



(**a**) (**b**)

Figure 6. Numerical simulations and bounds for different R, with $d\mu = e^{-V}, V = (x/5)^2/2 + (x/5)^8 + k, k \in \mathbb{R}$ and $d\mu \sim \mathcal{N}(0, \frac{1}{25})$. (**a**) Bound via (16) with $dP_X = d\mu, dP_Y = d\nu$. (**b**) Bound via (21) with $dP_X = d\nu, dP_Y = d\mu$.

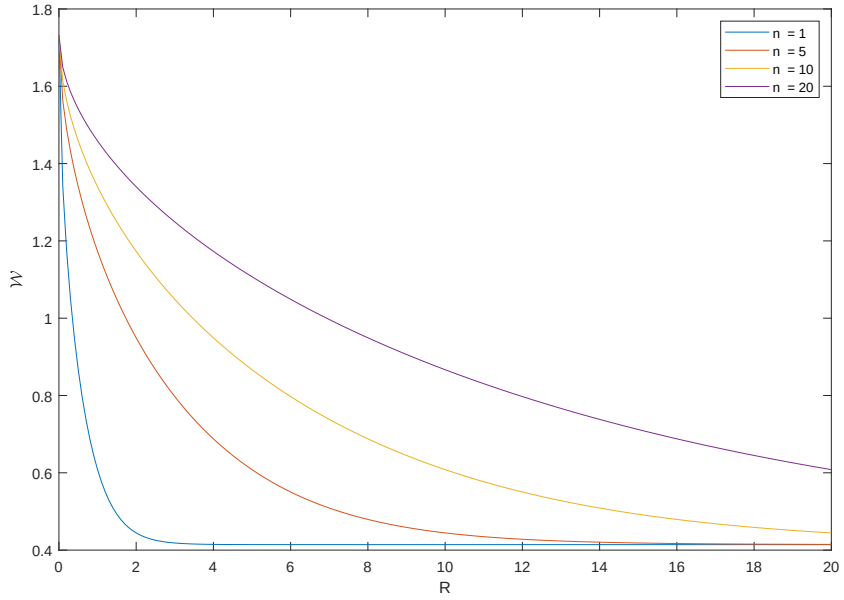


Figure 7. Sinkhorn distances between isotropic Gaussians in different dimensions.

5. Conclusions and Future Directions

In this paper, we considered a generalization of OT with an entropic constraint. We showed that the constraint leads to uncertainty and the uncertainty can be captured by EPI. We first derived an HWI-type inequality for the Sinkhorn distance. Then, we derived two Talagrand-type inequalities. Because of the strong geometric implication of Talagrand inequality, these two Talagrand-type inequalities can also give a weaker measure concentration inequality, respectively. From this result, we claimed that the geometry implied by the Sinkhorn distance can be smoothed out by the entropic constraint. We also showed that our results can be generalized into a conditional version of entropic OT inequality.

However, there are two factors unknown in the inequalities we derived, i.e., the optimal value of the term C in Theorem 3 and the error term ϵ' in Theorem 4 when one goes beyond the linear case. Although we showed that we can compute a suboptimal C using the arbitrary linear combination of two random vectors, the optimal value C^* is an intriguing open question to answer. We believe that the improvement of the term C may be related to the Fisher information. Without the assumption of strong log-concavity, it requires an extra term of relative Fisher information to upper-bound the Wasserstein distance in Theorem 2. The reversing of EPI in [31] is also concerned with Fisher information. If we consider the changing of Fisher information along the Schrödinger bridge, a better estimate of term C may be feasible.

Author Contributions: Conceptualization, P.A.S. and S.W.; methodology, S.W. and P.A.S.; software, S.W.; validation, P.A.S. and S.W.; formal analysis, S.W.; investigation, S.W.; resources, M.S.; writing—original draft preparation, S.W. and P.A.S.; writing—review and editing, P.A.S.; supervision, P.A.S.; project administration, P.A.S.; funding acquisition, M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Swedish Research Council (VR), grant number 2019-03606.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

OT	optimal transport
EPI	entropy power inequality
SP	Schrödinger problem
GAN	generative adversarial network
RHS	right-hand side
a.e.	almost everywhere
POT	Python Optimal Transport

Appendix A. Proof of Theorem 2

For a C^2 continuous function $V : \mathbb{R}^n \to \mathbb{R}$, $D^2V \geq \lambda I_n$, by Taylor formula ([4] Lemma 2.5), there exists a $t \in [0,1]$ satisfying

$$\begin{aligned} V(y) - V(x) =& \nabla V(x) \cdot (y-x) + (y-x) \cdot D^2V(tx + (1-t)y)(y-x)/2 \\ \geq& \nabla V(x) \cdot (y-x) + \frac{\lambda}{2}\|y-x\|^2. \end{aligned} \tag{A1}$$

Hence, we can bound the second-order cost by

$$\frac{\lambda}{2}\int_{\mathcal{X}\times\mathcal{Y}} \|y-x\|^2 \, dP \leq \int_{\mathcal{X}\times\mathcal{Y}} V(y) - V(x) - \nabla V(x) \cdot (y-x) \, dP. \tag{A2}$$

Because entropic OT is a minimization problem, we can take any case in $\Pi(P_X, P_Y; R)$ to bound $\mathcal{W}_2(P_X, P_Y; R)$. We take a linear combination $Y = Y_1 + Y_2$, where Y_1 and Y_2 are independent, $h(Y) - h(Y_2) \leq R$ and $\mathbb{E}[Y_2] = 0$. Assume that there is a Brenier map between Y_1 and X, i.e., $Y_1 = \nabla\varphi(X)$, which always exists, according to Theorem A1 (see Appendix F). Then, we can see that this special case is in $\Pi(P_X, P_Y, R)$, namely,

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) \\ &= h(Y) - h(Y_1 + Y_2|X) \\ &= h(Y) - h(Y_2) \\ &\leq R. \end{aligned}$$

Let $d\mu = e^{-V}$, where $V : \mathbb{R}^n \to \mathbb{R}$ is C^2 continuous, $D^2V \geq \lambda I_n$. In order to bound the Sinkhorn distance, we simply need to bound $\int_{\mathcal{X}\times\mathcal{Y}} \nabla V(x) \cdot (y - x)\, dP$, according to (A2). This term can be bounded as follows:

$$\begin{aligned} \int_{\mathcal{X}\times\mathcal{Y}} \nabla V(x) \cdot (y - x)\, dP &= \iint \nabla V(x) \cdot (\nabla\varphi(x) + y_2 - x)\, dP_{Y_2}\, dP_X \\ &= \int \nabla V(x) \cdot (\nabla\varphi(x) - x)\, dP_X \\ &= \int \nabla V(x) \cdot (\nabla\varphi(x) - x) \frac{dP_X}{d\mu}\, d\mu \\ &\geq \int \Delta\varphi(x) f\, d\mu - n + \int (\nabla\varphi(x) - x) \cdot \nabla f\, d\mu \qquad \text{(A3)} \\ &\geq n e^{\frac{1}{n}(h(Y_1) - h(X))} - n - \mathcal{W}_2(P_X, P_{Y_1}) \cdot \sqrt{I(P_X|\mu)}, \end{aligned}$$

where we take the Radon–Nikodym derivative $f = \frac{dP_X}{d\mu}$ in (A3) and apply Lemma A1 in Appendix F. This completes the derivation.

Appendix B. Proof of Corollary 2

Let $\mu_A = \frac{1_A}{\mu(A)}\mu$ and $\mu_{A_r} = \frac{1_{A_r}}{\mu(A_r)}\mu$ be the conditional probability measure μ restricted to A and A_r. Using the triangle inequality of $\mathcal{W}_2$, we have

$$r \leq \mathcal{W}_2(\mu_A, \mu_{A_r}; R) \leq \mathcal{W}_2(\mu_A, \mu) + \mathcal{W}_2(\mu, \mu_{A_r}; R). \qquad \text{(A4)}$$

From (8), we know that

$$\mathcal{W}_2(\mu_A, \mu) \leq \sqrt{2\lambda^{-1} D(\mu_A \| \mu)} = \sqrt{2\lambda^{-1} \log(1/\mu(A))} := c_A.$$

Let μ and μ_A be the probability measures of two random vectors X and Y, respectively. By substituting c_A into (A4) and using (16), we have

$$\begin{aligned} (r - c_A)^2 \leq & \mathcal{W}_2^2(\mu, \mu_{A_r}; R) \\ \leq & \frac{2}{\lambda}\left(\mathbb{E}[V(Y)] - \mathbb{E}[V(X)] + n - nCe^{\frac{1}{n}(h(Y) - h(X))}\right) \\ = & \frac{2}{\lambda}\left(\mathbb{E}[V(Y)] - \mathbb{E}[V(X)] + n - nCe^{\frac{1}{n}(\mathbb{E}[V(Y)] - \mathbb{E}[V(X)] - D(\mu_{A_r}\|\mu))}\right). \end{aligned}$$

Substituting $\mu_{A_r} = \frac{1_{A_r}}{\mu(A_r)}\mu$ into the definition of relative entropy, we have

$$\begin{aligned} D(\mu_{A_r}\|\mu) &= \int_{\mathbb{R}^n} \log \frac{1_{A_r}}{\mu(A_r)} \cdot \frac{1_{A_r}}{\mu(A_r)}\, d\mu \\ &= -\frac{\log \mu(A_r)}{\mu(A_r)} \int_{A_r} d\mu \\ &= -\log \mu(A_r). \end{aligned}$$

Then, we obtain, for $r \geq c_A$,

$$\mu(A_r) \leq C^{-n} \cdot e^{\mathbb{E}[V(X)]-\mathbb{E}[V(Y)]} \left[1 + \frac{1}{n}\left(\mathbb{E}[V(Y)] - \mathbb{E}[V(X)] - \frac{\lambda}{2}(r-c_A)^2\right)\right]^n. \tag{A5}$$

(A5) already indicates a concentration of measure. Using the inequality $(1+u/n)^n \leq e^u$, it can be further shown that (A5) also implies normal concentration, as follows:

$$\mu(A_r) \leq C^{-n} \cdot e^{-\frac{\lambda}{2}(r-c_A)^2}.$$

Appendix C. Proof of the Dimensionality of (20)

To prove that (20) is weaker when the dimension increases, we only need to show the decreasing of C^n with respect to n. When we increase the dimension, for given random vectors $Y_1 + Y_2 = Y$, we assume that this convolution relation still holds and their entropies increase proportionally to n, i.e., the shapes of their distributions do not change.

Since now n is a variable of C, we let $n \in \mathbb{R}$ and let C be a function with respect to $\frac{R}{n}$, in order to compute the partial derivative. We first show that C is a non-decreasing function with respect to $\frac{R}{n}$. We know that $C = e^{\frac{h(Y_1)}{n} - \frac{h(Y)}{n}}$ subject to $Y_1 + Y_2 = Y$ and $\frac{h(Y)}{n} - \frac{h(Y_2)}{n} \leq \frac{R}{n}$. Thus, for a larger value of $\frac{R}{n}$, there at least exists a $\frac{h(Y_2)}{n}$ that is non-increasing. It further leads to a non-decreasing $\frac{h(Y_1)}{n}$. Therefore, there at least exists a $C(\cdot, R+\Delta R)$ that is not smaller than $C(\cdot, R)$, $\forall \Delta R > 0$, i.e., C is non-decreasing with respect to $\frac{R}{n}$.

Then, we take the logarithm of C^n and compute the partial derivative with respect to n:

$$\begin{aligned} \frac{\partial}{\partial n} n \cdot \log C(\frac{R}{n}) &= \log C(\frac{R}{n}) + \frac{n}{C}\frac{\partial}{\partial n} C(\frac{R}{n}) \\ &= \log C(\frac{R}{n}) + \frac{nC'}{C} \cdot (-\frac{R}{n^2}). \end{aligned} \tag{A6}$$

We know that $C \in [0,1]$ and non-decreasing. Thus, (A6) is less than 0 when $C \in (0,1)$. This completes the proof, where we show that C^n is decreasing with respect to n.

Appendix D. Proof of Theorem 4

Let X' be a copy of X. X can be written as a linear combination $X = tX' + X_2$, where X_2 is zero-mean and independent of X', $h(X) - h(X_2) \leq R$, $t \in [0,1]$. Assume that there

exists a Brenier map $Y = \nabla\varphi(X')$. Similar to the proof of Theorem 3, this case is also in $\Pi(P_X, P_Y, R)$. Then, we have

$$\begin{aligned}
&\int_{\mathcal{X}\times\mathcal{Y}} \nabla V(x)\cdot(y-x)\,dP \\
=&\iint \nabla V(tx'+x_2)\cdot\nabla\varphi(x')\,dP_{X'}\,dP_{X_2} - n && \text{(A7)}\\
=&\iint (\nabla V(tx'+x_2) - t\cdot\nabla V(x'))\cdot\nabla\varphi(x')\,dP_{X'}\,dP_{X_2} + t\int \nabla V(x')\cdot\nabla\varphi(x')\,dP_{X'} - n \\
=&\iint (\nabla V(tx'+x_2) - t\cdot\nabla V(x'))\cdot\nabla\varphi(x')\,dP_{X'}\,dP_{X_2} + t\int \Delta\varphi\,dP_{X'} - n && \text{(A8)}\\
\geq& t\cdot n e^{\frac{1}{n}(h(Y)-h(X))} - \epsilon' - n, && \text{(A9)}
\end{aligned}$$

where we use Lemma A2 in (A7) and (A8). In (A9), we let $\epsilon' = -\iint(\nabla V(tx'+x_2) - t\cdot\nabla V(x'))\cdot\nabla\varphi(x')\,dP_{X'}\,dP_{X_2}$ and apply (A15). After changing the order of integral, we can see that $\int \nabla V(tx'+x_2)\,dP_{X_2}$ is a smoothed version of $\nabla V(tx')$. When ∇V is a linear function perturbed by zero-mean noise, i.e., $\nabla V(tx) = t\cdot\nabla V(x) + W$, the integral of x_2 is cancelled out and $\epsilon' = 0$. By taking $C_x(P_X, R) = t$, we complete the proof.

Appendix E. Proof of Theorem 5

According to Theorem 3, $\forall z_0 \in \mathcal{Z}$, there exists such $P_{X,Y|Z=z_0} \in \Pi(P_{X|Z=z_0}, P_{Y|Z=z_0}; R)$ satisfying $I(X;Y|Z=z_0) \leq R$ and following

$$\begin{aligned}
&\frac{\lambda}{2}\mathbb{E}[\|X-Y\|^2|Z=z_0] \\
\leq&\mathbb{E}[V(Y|Z=z_0)] - \mathbb{E}[V(X)] + n - nC(P_Y|Z=z_0, R)e^{\frac{1}{n}(h(Y|Z=z_0)-h(X))} \\
\leq&\mathbb{E}[V(Y|Z=z_0)] - \mathbb{E}[V(X)] + n - nC'(P_{Y|Z}, R)e^{\frac{1}{n}(h(Y|Z=z_0)-h(X))}. && \text{(A10)}
\end{aligned}$$

The conditional mutual information of the specific distribution $P_{X,Y|Z}$ can be bounded in this case as follows:

$$I(X;Y|Z) = \mathbb{E}_{P_Z}[I(X;Y|Z=z_0)] \leq \mathbb{E}_{P_Z}[R] = R.$$

Therefore, this $P_{X,Y|Z=z_0}$ yields the following estimate:

$$\begin{aligned}
&\frac{\lambda}{2}\mathcal{W}_2^2(P_{X|Z}, P_{Y|Z}|P_Z; R) \\
\leq&\frac{\lambda}{2}\mathbb{E}_{P_Z}\{\mathbb{E}[\|X-Y\|^2|Z=z_0]\} && \text{(A11)}\\
\leq&\mathbb{E}[V(Y|Z)] - \mathbb{E}[V(X)] + n - nC'(P_{Y|Z}, R)\mathbb{E}_{P_Z}[e^{\frac{1}{n}(h(Y|Z=z_0)-h(X))}] && \text{(A12)}\\
\leq&\mathbb{E}[V(Y|Z)] - \mathbb{E}[V(X)] + n - nC'(P_{Y|Z}, R)e^{\frac{1}{n}(\mathbb{E}_{P_Z}[h(Y|Z=z_0)]-h(X))} && \text{(A13)}\\
=&\mathbb{E}[V(Y|Z)] - \mathbb{E}[V(X)] + n - nC'(P_{Y|Z}, R)e^{\frac{1}{n}(h(Y|Z)-h(X))},
\end{aligned}$$

where (A11) follows from definition (7), (A12) follows from (A10), and (A13) follows from Jensen's inequality. This completes the proof.

Appendix F. Background Material

Theorem A1 (Brenier's, Theorem 2.12 [48]). *Let $P_X \in \mathcal{P}(\mathcal{X})$, $P_Y \in \mathcal{P}(\mathcal{Y})$ with $\mathcal{X} \subset \mathbb{R}^n$, $\mathcal{Y} \subset \mathbb{R}^n$ and assume that dP_X, dP_Y both have finite second moments. If P_X does not give mass to small sets, then, for the Kantorovich problem with cost $c(x,y) = \frac{1}{2}\|x-y\|^2$, $\exists!\varphi : \mathcal{X} \to \mathbb{R}$ gives the optimal coupling*

$$P^* = (Id \times \nabla\varphi)_{\#}P_X,$$

where φ is convex. $\nabla\varphi$ is called the Brenier map.

Lemma A1 (Theorem 9.17 [48]). *Let $d\mu = e^{-V}$. Let $f = \frac{dP_X}{d\mu}$ being a Radon–Nikodym derivative between two measures P_X and μ. Let $\nabla\varphi$ be a Brenier map. We have*

$$\begin{aligned}\int \nabla V(x)\cdot(\nabla\varphi(x)-x)f(x)d\mu(x) &\geq \int[(\Delta\varphi - n)f + (\nabla\varphi - x)\cdot\nabla f]\,d\mu \\ &= \int \Delta\varphi f\,d\mu - n + \int(\nabla\varphi - x)\cdot\nabla f\,d\mu, \end{aligned} \tag{A14}$$

where $\nabla\varphi(x) - x$ is called displacement. For the first term of (A14), because φ is convex, from ([4] Lemma 2.6), we have

$$\int \Delta\varphi\,dP_X \geq n e^{\frac{1}{n}(h(\nabla\varphi(X))-h(X))}. \tag{A15}$$

Moreover, the last term of (A14) can be bounded using Cauchy–Schwarz inequality as follows:

$$\begin{aligned}\int(\nabla\varphi - x)\cdot\nabla f\,d\mu &\geq -\left[\int\|\nabla\varphi - x\|^2 f\,d\mu\right]^{1/2}\left[\int\frac{\|\nabla f\|^2}{f}d\mu\right]^{1/2} \\ &= -\mathcal{W}_2(P_X, \nabla\varphi_{\#}P_X)\cdot\sqrt{I(P_X|\mu)}.\end{aligned}$$

Lemma A2 (Fact 7 [3]). *For any $\nabla\varphi \in L^1(\mathcal{X})\cap L^2(\mathcal{X})$ on a Polish space $(\mathcal{X},\mu)$ and $d\mu = e^{-V}$, we have*

$$\int \Delta\varphi\,d\mu = \int \nabla\varphi\cdot\nabla V\,d\mu.$$

Definition A1 (Lower Semi-Continuity, Appendix B, p. 602 [49]). *Given a metric space $\mathcal{X}$, $f:\mathcal{X}\to\mathbb{R}\cup\{\infty\}$ is lower semi-continuous if there exists a convergent sequence $\{x_n\}, x_n\to x$, that $f(x)\leq\liminf_{n\to\infty} f(x_n)$, where $\liminf_{n\to\infty} x_n := \lim_{n\to\infty}(\inf_{m\geq n} x_m)$.*

Definition A2 (Log-Concavity, Definition 2.1 [50]). *A density function f with respect to the Lebesgue measure on $(\mathbb{R}^n,\mathcal{B}^n)$ is log-concave if $f = e^{-V}$, where V is a convex function.*

Definition A3 (Strong Log-Concavity, Definition 2.8 [50]). *A density function f is called strongly log-concave if it has the form*

$$f(x) = g(x)\varphi(x),$$

with some log-concave g and some $\varphi\sim N(\mu,\Sigma)$.

References

1. Talagrand, M. Transportation cost for Gaussian and other product measures. *Geom. Funct. Anal.* **1996**, *6*, 587–600. [CrossRef]
2. Bakry, D.; Bolley, F.; Gentil, I. Dimension dependent hypercontractivity for Gaussian kernels. *Probab. Theory Relat. Fields* **2012**, *154*, 845–874. [CrossRef]
3. Cordero-Erausquin, D. Transport inequalities for log-concave measures, quantitative forms, and applications. *Can. J. Math.* **2017**, *69*, 481–501. [CrossRef]
4. Bolley, F.; Gentil, I.; Guillin, A. Dimensional improvements of the logarithmic Sobolev, Talagrand and Brascamp–Lieb inequalities. *Ann. Probab.* **2018**, *46*, 261–301. [CrossRef]
5. Raginsky, M.; Sason, I. Concentration of Measure Inequalities in Information Theory, Communications and Coding. In *Foundations and Trends in Communications and Information Theory*; NOW Publishers: Boston, MA, USA, 2018.
6. Zhang, R.; Chen, C.; Li, C.; Carin, L. Policy Optimization as Wasserstein Gradient Flows. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 5737–5746.
7. Montavon, G.; Müller, K.R.; Cuturi, M. Wasserstein Training of Restricted Boltzmann Machines. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 3718–3726.
8. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein Generative Adversarial Networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 214–223.

9. Rigollet, P.; Weed, J. Uncoupled isotonic regression via minimum Wasserstein deconvolution. *Inf. Inference* **2019**, *8*, 691–717. [CrossRef]
10. Cuturi, M. Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 2292–2300.
11. Wang, S.; Stavrou, P.A.; Skoglund, M. Generalized Talagrand Inequality for Sinkhorn Distance using Entropy Power Inequality; In Proceedings of the 2021 IEEE Information Theory Workshop (ITW), Kanazawa, Japan, 17–21 October 2021; pp. 1–6.
12. Benamou, J.D.; Brenier, Y. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numer. Math.* **2000**, *84*, 375–393. [CrossRef]
13. Villani, C. *Optimal Transport: Old and New*; Springer: Norwell, MA, USA, 2008, Volume 338.
14. Schrödinger, E. *Über die Umkehrung der Naturgesetze*; Verlag der Akademie der Wissenschaften in Kommission bei Walter De Gruyter: Berlin, Germany, 1931.
15. Léonard, C. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discret. Contin. Dyn. Syst.* **2014**, *34*, 1533–1574. [CrossRef]
16. Chen, Y.; Georgiou, T.T.; Pavon, M. On the relation between optimal transport and Schrödinger bridges: A stochastic control viewpoint. *J. Optim.Theory Appl.* **2016**, *169*, 671–691. [CrossRef]
17. Chen, Y.; Georgiou, T.T.; Pavon, M. Optimal transport over a linear dynamical system. *IEEE Trans. Autom. Control* **2016**, *62*, 2137–2152. [CrossRef]
18. Conforti, G. A second order equation for Schrödinger bridges with applications to the hot gas experiment and entropic transportation cost. *Probab. Theory Relat. Fields* **2019**, *174*, 1–47. [CrossRef]
19. Conforti, G.; Ripani, L. Around the entropic Talagrand inequality. *Bernoulli* **2020**, *26*, 1431–1452. [CrossRef]
20. Bai, Y.; Wu, X.; Özgür, A. Information Constrained Optimal Transport: From Talagrand, to Marton, to Cover. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Los Angeles, CA, USA, 21–26 June 2020; pp. 2210–2215.
21. Rigollet, P.; Weed, J. Entropic optimal transport is maximum-likelihood deconvolution. *C. R. Mathem.* **2018**, *356*, 1228–1235. [CrossRef]
22. Mena, G.; Niles-Weed, J. Statistical bounds for entropic optimal transport: Sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: New York, NY, USA, 2019; Volume 32.
23. Genevay, A.; Chizat, L.; Bach, F.; Cuturi, M.; Peyré, G. Sample Complexity of Sinkhorn Divergences. In Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, Naha, Japan, 16–18 April 2019; pp. 1574–1583.
24. Reshetova, D.; Bai, Y.; Wu, X.; Özgür, A. Understanding Entropic Regularization in GANs. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Victoria, Australia, 11–16 July 2021; pp. 825–830.
25. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
26. Stam, A.J. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Inf. Control* **1959**, *2*, 101–112. [CrossRef]
27. Rioul, O. Information theoretic proofs of entropy power inequalities. *IEEE Trans. Inf. Theory* **2010**, *57*, 33–55. [CrossRef]
28. Courtade, T.A.; Fathi, M.; Pananjady, A. Quantitative stability of the entropy power inequality. *IEEE Trans. Inf. Theory* **2018**, *64*, 5691–5703. [CrossRef]
29. Bobkov, S.; Madiman, M. Reverse Brunn—Minkowski and reverse entropy power inequalities for convex measures. *J. Funct. Anal.* **2012**, *262*, 3309–3339. [CrossRef]
30. Bobkov, S.G.; Madiman, M.M. On the problem of reversibility of the entropy power inequality. In *Limit Theorems in Probability, Statistics and Number Theory*; Springer: Norwell, MA, USA, 2013; pp. 61–74.
31. Courtade, T.A. A strong entropy power inequality. *IEEE Trans. Inf. Theory* **2017**, *64*, 2173–2192. [CrossRef]
32. Cover, T.M. *Elements of Information Theory*; John Wiley & Sons: New York, NY, USA, 1999.
33. Tamanini, L. A generalization of Costa's Entropy Power Inequality. *arXiv* **2020**, arxiv:2012.12230.
34. Monge, G. Mémoire sur la théorie des déblais et des remblais. In *Histoire de l'Académie Royale des Sciences de Paris*; De l'Imprimerie Royale: Paris, France, 1781.
35. Kantorovich, L.V. On the translocation of masses. *J. Math. Sci.* **2006**, *133*, 1381–1382. [CrossRef]
36. Kantorovich, L.V. On a Problem of Monge. *J. Math. Sci.* **2006**, *133*, 1383–1383. [CrossRef]
37. Dupuis, P.; Ellis, R.S. *A Weak Convergence Approach to the Theory of Large Deviations*; Wiley Series in Probability and Statistics; Wiley-Interscience: Hoboken, NJ, USA, 2011.
38. Luenberger, D.G. *Optimization by Vector Space Methods*; John Wiley & Sons: New York, NY, USA, 1997.
39. Blower, G. The Gaussian isoperimetric inequality and transportation. *Positivity* **2003**, *7*, 203–224. [CrossRef]
40. Otto, F.; Villani, C. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *J. Funct. Anal.* **2000**, *173*, 361–400. [CrossRef]
41. Bakry, D.; Ledoux, M. A logarithmic Sobolev form of the Li-Yau parabolic inequality. *Rev. Matemática Iberoam.* **2006**, *22*, 683–702. [CrossRef]
42. Masry, E. Multivariate probability density deconvolution for stationary random processes. *IEEE Trans. Inf. Theory* **1991**, *37*, 1105–1115. [CrossRef]
43. Stefanski, L.A.; Carroll, R.J. Deconvolving kernel density estimators. *Statistics* **1990**, *21*, 169–184. [CrossRef]
44. Fan, J. On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Stat.* **1991**, *19*, 1257–1272. [CrossRef]

45. Janati, H.; Muzellec, B.; Peyré, G.; Cuturi, M. Entropic optimal transport between unbalanced Gaussian measures has a closed form. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 10468–10479.
46. Marton, K. A measure concentration inequality for contracting Markov chains. *Geom. Funct. Anal.* **1996**, *6*, 556–571. [CrossRef]
47. Flamary, R.; Courty, N.; Gramfort, A.; Alaya, M.Z.; Boisbunon, A.; Chambon, S.; Chapel, L.; Corenflos, A.; Fatras, K.; Fournier, N.; et al. POT: Python Optimal Transport. *J. Mach. Learn. Res.* **2021**, *22*, 1–8.
48. Villani, C. *Topics in Optimal Transportation*; Number 58; American Mathematical Soc.: Providence, RI, USA, 2003.
49. Puterman, M.L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*; John Wiley & Sons: New York, NY, USA, 2014.
50. Saumard, A.; Wellner, J.A. Log-concavity and strong log-concavity: A review. *Stat. Surv.* **2014**, *8*, 45–114. [CrossRef]