Sparse Regularized Optimal Transport with Deformed $q$-Entropy

Han Bao $^1$* and Shinsaku Sakaue $^2$

1 Graduate School of Informatics and The Hakubi Center for Advanced Research, Kyoto University, Kyoto 604-8103, Japan
2 Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 153-8505, Japan
* Correspondence: bao@i.kyoto-u.ac.jp

Abstract: Optimal transport is a mathematical tool that has been widely used to measure the distance between two probability distributions. To mitigate the cubic computational complexity of the vanilla formulation of the optimal transport problem, regularized optimal transport has received attention in recent years, which is a convex program to minimize the linear transport cost with an added convex regularizer. Sinkhorn optimal transport is the most prominent one regularized with negative Shannon entropy, leading to densely supported solutions, which are often undesirable in light of the interpretability of transport plans. In this paper, we report that a deformed entropy designed by $q$-algebra, a popular generalization of the standard algebra studied in Tsallis statistical mechanics, makes optimal transport solutions supported sparsely. This entropy with a deformation parameter $q$ interpolates the negative Shannon entropy ($q = 1$) and the squared 2-norm ($q = 0$), and the solution becomes more sparse as $q$ tends to zero. Our theoretical analysis reveals that a larger $q$ leads to a faster convergence when optimized with the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm. In summary, the deformation induces a trade-off between the sparsity and convergence speed.

Keywords: optimal transport; Sinkhorn algorithm; convex analysis; entropy; quasi-Newton method

1. Introduction

Optimal transport (OT) is a classic problem in operations research, and it is used to compute a transport plan between suppliers and demanders with a minimum transportation cost. The minimum transportation cost can be interpreted as the closeness between the distributions when considering suppliers and demanders as two probability distributions. The OT problem has been extensively studied (also as the Wasserstein distance) [1] and used in robust machine learning [2], domain adaptation [3], generative modeling [4], and natural language processing [5], attributed to its many useful properties, such as the distance between two probability distributions. Recently, the OT problem has been employed for various modern applications, such as interpretable word alignment [6] and the locality-aware evaluation of object detection [7], because it can capture the geometry of data and provide a measurement method for closeness and alignment among different objects. From a computational perspective, a naïve approach is to use a network simplex algorithm or interior point method to solve the OT problem as a usual linear program; this approach requires supercubic time complexity [8] and is not scalable. A number of approaches have been suggested to accelerate the computation of the OT problem: entropic regularization [9,10], accelerated gradient descent [11], and approximation with tree [12] and graph metrics [13]. We focused our attention on entropic-regularized OT because it allows a unique solution attributed to strong convexity and transforms the original constrained optimization into an unconstrained problem with a clear primal–dual relationship. The
celebrated Sinkhorn algorithm solves entropic-regularized OT with square-time complexity [9]. Furthermore, the Sinkhorn algorithm is amenable to differentiable programming, and it is easily incorporated into end-to-end learning pipelines [14,15].

Despite the popularity of the Sinkhorn algorithm, one of the main drawbacks is that Shannon entropy blurs the OT solution, i.e., solutions of entropic-regularized OT are always densely supported. The Shannon entropy induces a probability distribution that has strictly positive values everywhere on its support owing to the nature of the Shannon entropy [16] whereas the vanilla (unregularized) OT produces extremely sparse transport plans located on the boundaries of a polytope [17,18]. If we are interested in alignment and matching between different objects (such as in the several applications of natural language processing [6,19]), dense transport plans are not so interpretable that matching information between objects may be obfuscated by unimportant small densities contained in the transport plans. One attempt toward realizing sparse OT is to use the squared two-norm as an alternative regularizer. Blondel et al. [20] showed that the dual of this optimization problem can be solved via the L-BFGS method [21]; the primal solution corresponds to a transport plan recovered from the dual solution in a closed form, which is sparse. Although they successfully obtained a sparse OT formulation with a numerically stable algorithm, the degree of the sparsity cannot be easily modulated when we prefer to control the sparsity given a final application. Furthermore, the theoretical convergence rates of solving regularized OT are yet to be known.

In this study, we aimed to examine the relationship between the sparsity of transport plans and the convergence guarantee of regularized OT. Specifically, we propose yet another entropic regularizer called deformed q-entropy with a deformation parameter $q$ that allows us to control the solution sparsity. We start with a dual solution of the entropic-regularized OT given by the Gibbs kernel to introduce a new regularizer; the Gibbs kernel associated with Shannon entropy induces nonsparsity, and, therefore, we replace the Gibbs kernel with another sparse kernel based on $q$-exponential distribution [22], following the idea of Tsallis statistics [23]. The deformed $q$ entropy is derived from the dual solution characterized by the sparse kernel. Interestingly, the deformed $q$ entropy recovers the Shannon entropy at the limit of $q \to 1$ and matches the (negative) squared two-norm at $q = 0$; this means that the deformed $q$ entropy interpolates between the two regularizers. We confirm that the solution becomes increasingly sparse as $q$ approaches zero. We call the regularized OT with the deformed $q$ entropy deformed $q$-optimal transport ($q$-DOT). The $q$-DOT reveals an interesting connection between the OT solution and the $q$-exponential distribution, which is an independent interest. From the optimization perspective, we can solve the unconstrained dual of $q$-DOT with many standard solvers, as reported in Blondel et al. [20]. We can see that the convergence becomes faster with the BFGS method [24] as the deformation parameter $q$ approaches one, as a result of our analysis of the convergence rate of the dual optimization. Therefore, the weaker deformation (larger $q$) leads to faster convergence while sacrificing sparsity. Finally, we demonstrate the trade-off between sparsity and convergence in the numerical experiments.

Our contributions can be summarized as: (i) showing a clear connection between the regularized OT problem and the $q$-exponential distribution; (ii) demonstrating the trade-off of the $q$-DOT between sparsity and convergence; (iii) providing a formal convergence guarantee of the $q$-DOT when solved with the BFGS method. The rest of this paper is organized as follows: Section 2 introduces the necessary background to the OT problem and entropic regularization. In Section 3, the Lagrange dual of the entropic-regularized OT problem is first shown; then, the dual optimal formula and the $q$-exponential distribution is connected to sparsify the transport matrix. Section 4 specifically focuses on the optimization perspective of the regularized OT problem, and a convergence guarantee with the BFGS method is provided, which shows the theoretical trade-off between sparsity and convergence. Finally, the empirical behavior and the trade-off of the regularized OT are numerically confirmed in Section 5.
2. Background

2.1. Preliminaries

For $x \in \mathbb{R}$, let $[x]_+ = x$ if $x > 0$ and 0 otherwise, and let $[x]_+^p$ represent $([x]_+)^p$ hereafter. For a convex function $f, \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ represents a Euclidean vector space equipped with an inner product $\langle \cdot, \cdot \rangle$, the Fenchel–Legendre conjugate $f^* : \mathcal{X} \to \mathbb{R}$ is defined as $f^*(y) := \sup_{x \in \mathcal{X}} \langle x, y \rangle - f(x)$. The relative interior of a set $S$ is denoted by $\text{ri} S$, and the effective domain of a function $f$ is denoted by $\text{dom}(f)$. A differentiable function $f$ is said to be $M$-strongly convex over $S \subseteq \text{ri} \text{dom}(f)$ if, for all $x, y \in S$, we have $f(x) - f(y) \leq \langle \nabla f(x), x - y \rangle - \frac{M}{2} \|x - y\|^2_2$. If $f$ is twice differentiable, the strong convexity is equivalent to $\nabla^2 f(x) \succeq MI$ for all $x \in S$. Similarly, a differentiable function $f$ is said to be $M$-smooth over $S \subseteq \text{ri} \text{dom}(f)$ if for all $x, y \in S$, we have $\|\nabla f(x) - \nabla f(y)\|_2 \leq M\|x - y\|_2$, which is equivalent to $\nabla^2 f(x) \preceq MI$ for all $x \in S$ if $f$ is twice differentiable.

2.2. Optimal Transport

The OT is a mathematical problem to find a transport plan between two probability distributions with the minimum transport cost. The discussions in this paper are restricted to discrete distributions. Let $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m b_j \delta_{y_j}$ be histograms supported on the finite sets of points $(x_i)_n^{i=1} \subseteq \mathcal{X}$ and $(y_j)_m^{j=1} \subseteq \mathcal{X}$, respectively, where $a \in \Delta_n$ and $b \in \Delta_m$ are probability vectors. The OT between two discrete probability measures $\mu$ and $\nu$ is the optimization problem

$$ T(\mu, \nu) := \inf_{\Pi \in \mathcal{U}(\mu, \nu)} \sum_{i=1}^n \sum_{j=1}^m d(x_i, y_j) \Pi_{ij}, $$

(1)

where $\mathcal{U}$ represents the transport polytope, defined as

$$ \mathcal{U}(\mu, \nu) := \left\{ \Pi \in \mathbb{R}^{n \times m}_{\geq 0} \mid \Pi m = a, \Pi ^\top 1_n = b \right\}. $$

(2)

The transport polytope $\mathcal{U}$ defines the constraints on the row/column marginals of a transport matrix $\Pi$. These constraints are often referred to as coupling constraints. For notational simplicity, matrix $D_{ij} := d(x_i, y_j)$ and expectation $\langle D, \Pi \rangle := \sum_{i=1}^n \sum_{j=1}^m D_{ij} \Pi_{ij}$ are used hereafter. $T(\mu, \nu)$ is known as a 1-Wasserstein distance, which defines a metric space over histograms [1].

Equation (1) is a linear program and can be solved by well-studied algorithms such as the interior point and network simplex methods. However, its computational complexity is $O(n^3 \log n)$ (assuming $n = m$), so it is not scalable to large datasets [8].

2.3. Entropic Regularization and Sinkhorn Algorithm

The entropic-regularized formulation is commonly used to reduce the computational burden. Here, we introduce regularized OT with negative Shannon entropy [9] as

$$ T_{\lambda H}(\mu, \nu) := \inf_{\Pi \in \mathcal{U}(\mu, \nu)} \langle D, \Pi \rangle + \lambda \sum_{i=1}^n \sum_{j=1}^m (\Pi_{ij} \log \Pi_{ij} - \Pi_{ij}), $$

(3)

negative Shannon entropy
where $\lambda > 0$ represents the regularization strength. Let us review the derivation of the updates of the Sinkhorn algorithm. The Lagrangian of the optimization problem in Equation (3) is

$$L(\Pi, \alpha, \beta) := \sum_{i=1}^{n} \sum_{j=1}^{m} (D_{ij}\Pi_{ij} + \lambda(\Pi_{ij} \log \Pi_{ij} - \Pi_{ij}))$$

where $\alpha \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^m$ represent the Lagrangian multipliers. Equation (4) ignores the constraints $\Pi_{ij} \geq 0$ (for all $i \in [n]$ and $j \in [m]$); however, they will be automatically satisfied. By taking the derivative in $\Pi_{ij}$,

$$\nabla_{\Pi_{ij}} L = D_{ij} + \lambda \log \Pi_{ij} + \alpha_i + \beta_j,$$

and, hence, the stationary condition $\nabla_{\Pi_{ij}} L = 0$ induces the solution

$$\Pi_{ij} = \exp\left(-\frac{\alpha_i + \beta_j + D_{ij}}{\lambda}\right).$$

The decomposition $\Pi_{ij} = \exp\left(-\frac{D_{ij}}{\lambda}\right) / \exp\left(-\frac{\alpha_i + \beta_j}{\lambda}\right)$ suggests that the stationary point is the (normalized) Gibbs kernel $\exp\left(-\frac{D_{ij}}{\lambda}\right)$. One can easily infer that the Sinkhorn solution is dense because the Gibbs kernel is supported on the entire $\mathbb{R}_+^n$, i.e., $\exp\left(-\frac{z}{\lambda}\right) > 0$ for all $z \in \mathbb{R}_+^n$. We can write Equation (6) into a matrix form by applying the variable transforms $u_i := \exp\left(-\frac{\alpha_i}{\lambda}\right)$, $v_j := \exp\left(-\frac{\beta_j}{\lambda}\right)$, and $K_{ij} := \exp\left(-\frac{D_{ij}}{\lambda}\right)$ as

$$\Pi = \text{diag}(u) K \text{diag}(v).$$

The following Sinkhorn updates are used to make Equation (7) meet the marginal constraints:

$$\begin{cases} u' \leftarrow \frac{a}{(Kv)} \\ v' \leftarrow \frac{b}{(K^T u)} \end{cases}$$

where $z / \eta$ represents the element-wise division of the two vectors $z$ and $\eta$. The computational complexity is $O(Knm)$ because the Sinkhorn updates involve only matrix-vector multiplications and element-wise divisions; $K$ represents the number of the Sinkhorn updates. Finer analysis of the number of updates required to meet the error tolerance is provided in the literature [25].

3. Deformed $q$-Entropy and $q$-Regularized Optimal Transport

3.1. Regularized Optimal Transport and Its Dual

Let us consider the following primal problem with a general regularization function $\Omega$.

**Definition 1** (Primal of regularized OT).

$$\mathcal{T}_{\Omega}(\mu, \nu) = \inf_{\Pi \in \mathcal{U}(\mu, \nu)} \langle D, \Pi \rangle + \sum_{i,j} \Omega(\Pi_{ij}),$$

where $\Omega : \mathbb{R} \rightarrow \mathbb{R}$ represents a proper closed convex function.
Next, we derive its dual by Lagrange duality. The Lagrangian of Equation (9) is defined as
\[
\mathcal{L}(\Pi, \alpha, \beta) := \langle D, \Pi \rangle + \sum_{i,j} \Omega(\Pi_{ij}) + \langle \alpha, \Pi^\top_1 - \mathbf{a} \rangle + \langle \beta, \Pi^\top_0 - \mathbf{b} \rangle, \tag{10}
\]
with dual variables \(\alpha \in \mathbb{R}^n\) and \(\beta \in \mathbb{R}^m\). Then, the primal can be rewritten in terms of the Lagrangian
\[
\mathcal{T}_\Omega(\mu, v) = \inf_{\Pi \in \mathbb{R}^{n \times m}_{\geq 0}} \sup_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^m} \mathcal{L}(\Pi, \alpha, \beta). \tag{11}
\]
In this Lagrangian formulation, we let the constraints \(\Pi \in \mathbb{R}^{n \times m}_{\geq 0}\) remain for a technical reason. The constrained optimization problem in (11) can be reformulated into the following unconstrained one with an indicator function \(I_{\mathbb{R}^{n \times m}_{\geq 0}}\).
\[
\mathcal{T}_\Omega(\mu, v) = \inf_{\Pi \in \mathbb{R}^{n \times m}} \sup_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^m} \mathcal{L}(\Pi, \alpha, \beta) + I_{\mathbb{R}^{n \times m}_{\geq 0}}(\Pi), \tag{12}
\]
which corresponds to an optimization problem with the convex objective function \(\langle D, \Pi \rangle + \sum_{i,j} \Omega(\Pi_{ij}) + I_{\mathbb{R}^{n \times m}_{\geq 0}}(\Pi)\) with only the linear constraints \(\Pi^*_{1,n} = \mathbf{a}\) and \(\Pi^*_{0,n} = \mathbf{b}\). By invoking the Sinkhorn–Knopp theorem [26], the existence of a strictly feasible solution, namely, a solution satisfying \(\Pi^*_{1,n} = \mathbf{a}\) and \(\Pi^*_{0,n} = \mathbf{b}\), can be confirmed. Hence, we see that the Slater condition is satisfied, and the strong duality holds as follows:
\[
\mathcal{T}_\Omega(\mu, v) = \sup_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^m} \inf_{\Pi \in \mathbb{R}^{n \times m}} \mathcal{L}(\Pi, \alpha, \beta)
= \sup_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^m} -\langle \mathbf{a}, \alpha \rangle - \langle \mathbf{b}, \beta \rangle + \inf_{\Pi \in \mathbb{R}^{n \times m}} \sum_{i,j} \langle D_{ij} + \alpha_i + \beta_j \rangle \Pi_{ij} + \Omega(\Pi_{ij})
= \sup_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^m} -\langle \mathbf{a}, \alpha \rangle - \langle \mathbf{b}, \beta \rangle - \left( \sup_{\Pi \in \mathbb{R}^{n \times m}} \sum_{i,j} \langle -D_{ij} + \alpha_i + \beta_j \rangle \Pi_{ij} - \Omega(\Pi_{ij}) \right)
= \sup_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^m} -\langle \mathbf{a}, \alpha \rangle - \langle \mathbf{b}, \beta \rangle - \sum_{i,j} \Omega^*( -D_{ij} - \alpha_i - \beta_j), \tag{13}
\]
where \(\Omega^*\) represents the Fenchel–Legendre conjugate of \(\Omega : \mathbb{R} \rightarrow \mathbb{R}\)
\[
\Omega^*(\eta) := \sup_{\pi \geq 0} \eta \pi - \Omega(\pi). \tag{14}
\]
Although each element of the transport plans ranges over \([0, 1]\), it is sufficient to define the Fenchel–Legendre conjugate as the supremum over \(R_{\geq 0}\) because of how \(\Omega^*\) emerges in the strong duality (13). According to Danskin’s theorem [27], the supremum of the Fenchel–Legendre conjugate can be attained at
\[
\Pi^*_{ij} = \nabla \Omega^* (-D_{ij} - \alpha_i - \beta_j). \tag{15}
\]
Therefore, the dual of regularized OT is formulated as follows:

**Definition 2 (Dual of regularized OT).**
\[
\mathcal{T}_\Omega(\mu, v) = \sup_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^m} -\langle \mathbf{a}, \alpha \rangle - \langle \mathbf{b}, \beta \rangle - \sum_{i,j} \Omega^* (-D_{ij} - \alpha_i - \beta_j), \tag{16}
\]
where \(\Omega^*\) represents the Fenchel–Legendre conjugate \(\Omega^*(\eta) := \sup_{\pi \geq 0} \eta \pi - \Omega(\pi)\). The optimal solution of the primal is given by the dual map \(\nabla \Omega^*\) such that \(\Pi^*_{ij} = \nabla \Omega^* (-D_{ij} - \mathbf{a}_i^* - \mathbf{b}_j^*)\), where \((\mathbf{a}^*, \mathbf{b}^*)\) represents the dual optimal solution.
Next, we see several examples that are summarized in Table 1.

Example 1 (Negative Shannon entropy). Let \( \Omega(\pi) = -\lambda H(\pi) = \lambda (\pi \log \pi - \pi) \); then \( \Omega^*(\eta) = \lambda e^{\eta/\lambda} \) and \( \nabla \Omega^*(\eta) = e^{\eta/\lambda} \). The optimal solution represented with the optimal dual variables \( (\alpha^*, \beta^*) \) is \( \Pi_{ij}^* = \exp\left(-\frac{D_{ij} + \alpha_i^* + \beta_j^*}{\lambda}\right) \). This recovers the stationary point of the Sinkhorn OT in Equation (6). The solution is dense because the regularizer \( \Omega \) induces the Gibbs kernel \( \nabla \Omega^*(\eta) = e^{\eta/\lambda} > 0 \) for all \( \eta \in \mathbb{R} \).

Example 2 (Squared 2-norm). Let \( \Omega(\pi) = \frac{1}{2} \pi^2 \); then \( \Omega^*(\eta) = \frac{1}{2\pi} [\eta]^2 \) and \( \nabla \Omega^*(\eta) = \frac{1}{2\pi} [\eta]_+ \). The optimal solution represented with the optimal dual variables \( (\alpha^*, \beta^*) \) is \( \Pi_{ij}^* = \frac{1}{\lambda} \left[-D_{ij} - \alpha_i^* - \beta_j^*\right]_+ \). As mentioned by Blondel et al. [20], the squared 2-norm can sparsify the solution because \( \nabla \Omega^*(\eta) = \frac{1}{\lambda} [\eta]_+ \) may take the value 0.

Table 1. Summary of \( \Omega(\pi) \), \( \Omega^*(\eta) \), and \( \nabla \Omega^*(\eta) \) for several regularizers. The relationship between \( \Omega \), its conjugate, and the derivatives are summarized in Bao and Sugiyama [28].

| \( \Omega(\pi) \) | \( \Omega^*(\eta) \) | \( \nabla \Omega^*(\eta) \) |
|----------------|----------------|----------------|
| Negative entropy | \( \lambda (\pi \log \pi - \pi) \) | \( \lambda e^{\eta/\lambda} \) | \( e^{\eta/\lambda} \) |
| Squared 2-norm | \( \frac{1}{2} \pi^2 \) | \( \frac{1}{2\pi} [\eta]^2 \) | \( \frac{1}{2\pi} [\eta]_+ \) |
| Deformed \( q \) entropy | \( \frac{1}{2-q} (\pi \log_q (\pi) - \pi) \) | \( \frac{1}{2-q} \exp_q(\eta/\lambda)^{2-q} \) | \( \exp_q(\eta/\lambda) \) |

3.2. \( q \) Algebra and Deformed Entropy

As shown in the last few examples, the dual map \( \nabla \Omega^* \) plays an important role in the OT solution sparsity. In addition, the induced \( \nabla \Omega^* \) is the Gibbs kernel when the negative Shannon entropy is used as \( \Omega \). Therefore, one may think of designing a regularizer from \( \nabla \Omega^* \) by utilizing a kernel function that induces sparsity. One candidate is a \( q \)-exponential distribution. We begin with some basics required to formulate \( q \)-exponential distributions.

First, we introduce \( q \)-algebra, which has been well studied in the field of Tsallis statistical mechanics [23,29,30]. \( q \) algebra has been used in the machine-learning literature for regression [31], Bayesian inference [32], and robust learning [33]. For a deformation parameter \( q \in [0, 1] \), the \( q \)-logarithm and \( q \)-exponential functions are defined as

\[
\log_q(x) := \begin{cases} 
\frac{e^{1-q} - 1}{1-q} \log(x) & \text{if } q \in (0, 1) \\
1 & \text{if } q = 1
\end{cases}, \quad \exp_q(x) := \begin{cases} 
(1 + (1-q)x)^{1/(1-q)} & \text{if } q \in (0, 1) \\
\exp(x) & \text{if } q = 1
\end{cases}.
\]

The \( q \) logarithm is defined for only \( x > 0 \), as in the natural logarithm; they are inverse functions to each other (in an appropriate domain) and they recover the natural definition of the logarithm and exponential as \( q \to 1 \). Their derivatives are \( (\log_q(x))' = \frac{1}{x^q} \) and \( (\exp_q(x))' = \exp_q(x)^q \), respectively. The additive factorization property \( \exp(x + y) = \exp(x) \exp(y) \) satisfied by the natural exponential no longer holds for the \( q \) exponential, such that \( \exp_q(x+y) \neq \exp_q(x) \exp_q(y) = \exp_q(x+y+(1-q)xy) \). Instead, we can construct another algebraic structure by introducing the other operation called the \( q \) product \( \otimes_q \):

\[
x \otimes_q y = [x^{1-q} + y^{1-q} - 1]_+^{1/(1-q)}.
\]

With this product, the pseudoaddititive factorization \( \exp_q(x+y) = \exp_q(x) \otimes_q \exp_q(y) \) holds. Thus, the \( q \) algebra captures rich nonlinear structures, and it is often used to extend the Shannon entropy to the Tsallis entropy [23]

\[
T_q(\pi) = -\sum_{i=1}^n \pi_i^q \log_q(\pi_i).
\]
One can see that the Tsallis entropy has an equivalent power formulation
\[ T_q(\pi) = \sum_{i=1}^{n} \pi_i - \frac{\pi_i^q}{1-q}, \]
which means that it is often suitable for modeling heavy-tailed phenomena such as the power law. Although the introduced \( q \) logarithm and exponential can look arbitrary, they can be axiomatically derived by assuming the essential properties of the algebra (see Naudts \[29\]). For more physical insights, we recommend readers to refer to the literature \[30\].

Next, we introduce the \( q \)-exponential distribution. We introduce a simpler form for our purpose, whereas more general formulations of the \( q \)-exponential distribution have been introduced in the literature \[22\]. Given the form of the Gibbs kernel \( k(\xi) := \exp(-\xi/\lambda) \), we define the \( q \)-Gibbs kernel as follows:

**Definition 3** (\( q \)-Gibbs kernel). For \( \xi \geq 0 \), we define the \( q \)-Gibbs kernel as \( k_q(\xi) := \exp_q(-\xi/\lambda) \) for a deformation parameter \( q \in [0,1) \) and a temperature parameter \( \lambda \in \mathbb{R}_{>0} \).

If we take \( \xi \) as the (centered) squared distance, then \( k_q(\xi) \) represents the \( q \)-Gaussian distribution \[22\]. We illustrate the \( q \)-Gibbs kernel with different deformation parameters in Figure 1.

![Figure 1. Plots of the \( q \)-Gibbs kernels with different \( q \) (\( \lambda = 1 \)).](image)

By definition, the support of the \( q \)-Gibbs kernel is \( \text{supp}(k_q) = [0, \frac{\lambda}{1-q}] \) for \( q \in [0,1) \) and \( \text{supp}(k_q) = \mathbb{R}_{>0} \) for \( q = 1 \). This indicates that the \( q \)-Gibbs kernel ignores the effect of a too-large \( \xi \) (or too large a distance between two points); its threshold is smoothly controlled by the temperature parameter \( \lambda \) and deformation parameter \( q \).

Finally, we derive an entropic regularizer that induces sparsity by using the \( q \)-Gibbs kernel. Given the stationary condition in Equation (15), we impose the following functional form on the dual map:

\[ \pi = \nabla \Omega^*(\eta) = \exp_q\left(\frac{\eta}{\lambda}\right), \tag{20} \]

where \((\pi, \eta) = (\Pi_{ij}^*, -D_{ij} - \alpha_i - \beta_j)\). Equation (20) results in the factorization

\[ \Pi_{ij}^* = \exp_q\left(-\frac{\alpha_i}{\lambda}\right) \otimes_q \exp_q\left(-\frac{D_{ij}}{\lambda}\right) \otimes_q \exp_q\left(-\frac{\beta_j}{\lambda}\right), \tag{21} \]
and a sufficiently large input distance \( D_{ij} \) drives \( \Pi_{ij} \) to zero; though \( \exp_q(-D_{ij}/\lambda) = 0 \) does not immediately imply \( \Pi_{ij} = 0 \) because the \( q \)-product \( \otimes_q \) lacks an absorbing element.

By solving Equation (20),

\[
\nabla \Omega(\pi) = \lambda \log_q(\pi), \quad \Omega(\pi) = \frac{\lambda}{2-q} \left( \pi \log_q(\pi) - \pi \right).
\]

For the completeness, its derivation is shown in Appendix A. Hence, we define the \textit{deformed} \( q \) \textit{entropy} as follows:

**Definition 4** (Deformed \( q \)-entropy). For \( \pi \in \Delta^{n-1} \), the deformed \( q \) entropy is defined as

\[
H_q(\pi) = -\frac{1}{2-q} \sum_{i=1}^{n} (\pi_i \log_q(\pi_i) - \pi_i).
\]

The deformed \( q \)-entropic regularizer for an element \( \pi_i \) is

\[
\Omega(\pi_i) = \frac{\lambda}{2-q} (\pi_i \log_q(\pi_i) - \pi_i).
\]

The deformed \( q \) entropy recovers the Shannon entropy at the limit \( q \rightarrow 1 \): \( H_1(\pi) = -\sum_{i} (\pi_i \log(\pi_i) - \pi_i) \). In addition, the limit \( q \downarrow 0 \) recovers the negative of the squared 2-norm: \( H_0(\pi) = -\frac{1}{2} \sum_{i} (\pi_i^2 - 2\pi_i) = -\frac{1}{2} \| \pi \|_2^2 + 1 \). Therefore, the deformed \( q \) entropy is an interpolation between the Shannon entropy and squared 2-norm. Hereafter, we consider the regularized OT with the deformed \( q \) entropy

\[
T_{-\lambda H_q}(\mu, \nu) = \inf_{\Pi \in \mathcal{U}(\mu, \nu)} \langle D, \Pi \rangle - \lambda H_q(\Pi),
\]

by solving its dual counterpart. The deformed \( q \) entropy is different from the Tsallis entropy \( T_q \) (see Equation (19)) in that the Tsallis entropy and deformed \( q \) entropy are defined by the \( q \) expectation \( \langle \pi^q, \cdot \rangle \) \cite{34} and the usual expectation \( \langle \pi, \cdot \rangle \), respectively, while both are defined by the \( q \) logarithm.

**Remark 1.** The primary reason we picked the deformed \( q \) entropy \( H_q \) to design the regularizer is owing to its natural connection to the \( q \)-Gibbs kernel through the dual map, \( \nabla (-\lambda H_q)^*(\eta) = \exp_q(\eta/\lambda) \). When the Tsallis entropy \( T_q \) is used, the dual map is

\[
\nabla (-\lambda T_q)^*(\eta) = \frac{q^{1/(1-q)}}{\exp_q(-\eta/\lambda)},
\]

which is not naturally connected to the \( q \)-Gibbs kernel. Muzellec et al. \cite{35} proposed regularized OT with the Tsallis entropy, but they did not discuss its sparsity. As we show in Appendix D.1, the Tsallis entropy does not empirically induce sparsity.

In Figure 2, the deformed \( q \) entropy with a different deformation parameter is plotted for the one-dimensional simplex \( \Delta^1 \). One can easily confirm that \( H_q(\pi) \) is concave for \( \pi \in \mathbb{R}_{\geq 0}^n \), as illustrated in the figure.
4. Optimization and Convergence Analysis

4.1. Optimization Algorithm

We occasionally write $\Omega = -\lambda H_q$ to simplify the notation in this section. By simple algebra, we confirm

$$\Omega^*(\eta) = \frac{\lambda}{2 - q} \exp\left(\frac{\eta}{\lambda}\right)^{2-q},$$

which is convex because of the concavity of $H_q$. To solve Equation (24), we solve the dual

$$\mathcal{T}_{-\lambda H_q}(\mu, \nu) = \sup_{a \in \mathbb{R}^n, b \in \mathbb{R}^m} -\langle a, \alpha \rangle - \langle b, \beta \rangle - \frac{\lambda}{2 - q} \sum_{i,j} \exp\left(-\frac{D_{ij} + \alpha_i + \beta_j}{\lambda}\right)^{2-q},$$

where $z := (\alpha, \beta)$ denotes dual variables. As Equation (27) is an unconstrained optimization problem, many famous optimization solvers can be used to solve it; here, we use the BFGS method [24]. For the sake of convergence analysis (Section 4.2), we optimize the convex $\ell_2$-regularized dual objective

$$\min z \; \bar{F}(z) := \langle a, \alpha \rangle + \langle b, \beta \rangle + \sum_{i,j} \Omega^*\left(-D_{ij} - \alpha_i - \beta_j\right) + \frac{\kappa}{2} \|z\|_2^2,$$

where $\kappa > 0$ represents the $\ell_2$-regularization parameter. In practice, $\ell_2$ regularization hardly affects the performance when $\kappa$ is sufficiently small. We can characterize the convergence rate by introducing (small) $\ell_2$ regularization, which makes the objective strongly convex, whereas the convergence guarantee without its rate is still possible without $\ell_2$ regularization [36].

We briefly summarize the algorithm in Algorithm 1, where $d^{(k)}, \rho^{(k)},$ and $g^{(k)} := \nabla \bar{F}(z^{(k)})$ represent the $k$th update direction, $k$th step size, and gradient at the current variable $z^{(k)}$, respectively.

$$s^{(k)} := z^{(k+1)} - z^{(k)} \quad \text{and} \quad g^{(k)} := g^{(k+1)} - g^{(k)}$$

Figure 2. Plots of deformed $q$ entropy with different $q$ values. A constant term is ignored in the plots so that the end points are calibrated to zero.
are the differences of the dual variables and gradients between the next and current steps, respectively. Furthermore, let \( \langle \gamma, \gamma' \rangle \) be the tolerance parameter for the Wolfe conditions, i.e., update directions and step sizes satisfy the conditions

\[
\tilde{F}(z^{(k)} + \rho^{(k)}d^{(k)}) \leq \tilde{F}(z^{(k)}) + \gamma' \rho^{(k)}g^{(k)}\top d^{(k)},
\]

(Armijo condition) (30)

\[
g^{(k+1)}\top d^{(k)} \geq \gamma g^{(k)}\top d^{(k)}.
\]

(curvature condition) (31)

**Algorithm 1:** BFGS algorithm for dual regularized OT

\begin{verbatim}
Input : \( z^{(0)} \) initial point, \( 0 < \gamma' < \frac{1}{2} \) tolerance parameter for the Armijo condition, \( \gamma' < \gamma < 1 \) tolerance parameter for the curvature condition, and \( B^{(0)} = I \) initial Hessian estimate

1 for \( k = 0, \ldots, K - 1 \) do
  \triangleright Calculate update direction
  2 \( d^{(k)} \leftarrow -[B^{(k)}]^{-1}g^{(k)} \)
  \triangleright Determine step size by line search
  3 \( \rho^{(k)} \leftarrow \text{line_search}(d^{(k)}, g^{(k)}, \gamma, \gamma') \)
  \triangleright Update dual variables
  4 \( z^{(k+1)} \leftarrow z^{(k)} + \rho^{(k)}d^{(k)} \)
  \triangleright Update Hessian estimate
  5 \( B^{(k+1)} \leftarrow B^{(k)} - \frac{B^{(k)}s^{(k)}s^{(k)}\top}{s^{(k)}\top B^{(k)}s^{(k)}} + \frac{\xi^{(k)}\xi^{(k)}\top}{\xi^{(k)}\top s^{(k)}} \)

end
7 return \( (\hat{a}, \hat{b}) \leftarrow z^{(K)} \)
\end{verbatim}

After obtaining the dual solution \( (\hat{a}, \hat{b}) \), the primal solution can be recovered from Equation (15).

4.2. Convergence Analysis

We provide a convergence guarantee for Algorithm 1. A technical assumption is stated beforehand.

**Assumption 1.** Let \( z_* \) be the global optimum of \( \tilde{F} \). For \( \tau \in (0, 1) \), we define the set \( Z_{\tau} \subseteq \text{dom}(\tilde{F}) \) as

\[
Z_{\tau} := \{ z \mid \nabla \Omega^*(-D_{ij} - a_i - b_j) \leq \tau \text{ for all } i, j \}. \tag{32}
\]

Assume that \( z^{(K)} \) obtained by Algorithm 1 and \( z_* \) are contained in \( Z_{\tau} \).

The dual map \( \nabla \Omega^* \) translates dual variables into primal variables, as in Equation (15). It is easy to confirm that \( Z_{\tau} \) is a closed convex set attributed to the convexity of \( \nabla \Omega^* \). Assumption 1 essentially assumes that all elements of the primal matrix (of \( z^{(k)} \) and \( z_* \)) are strictly less than 1; this always holds for \( z_* \) (unless \( n = m = 1 \)) because of the strong duality. Moreover, this assumption is natural for \( z^{(K)} \) values sufficiently close to the optimum \( z_* \). The bound parameter \( \tau \) is a key element for characterizing the convergence speed.

**Theorem 1.** Let \( N := \max\{n, m\} \). Under Assumption 1, Algorithm 1 with the parameter choice \( \kappa = 2Nt\lambda^{-1} \) returns a point \( z^{(K)} \) satisfying

\[
\|g^{(K)}\|_2 < \sqrt{\frac{16(\tilde{F}(z^{(0)}) - \tilde{F}_*)Nt\tau^r\tau^K}{\lambda}} \tag{33}
\]

where \( \tilde{F}_* := \inf_{z \in \Omega} \tilde{F}(z) \) represents the optimal value of the \( \ell_2 \)-regularized dual objective and \( 0 < r < 1 \) is an absolute constant independent from \( (\lambda, \tau, q, N) \).
The proof is shown in Section 4.3. We conclude that a larger deformation parameter \( q \) yields better convergence because the coefficient in Equation (33) is \( O(\tau^{1/2}) \) with the base \( \sqrt{\tau} < 1 \). Therefore, the deformation parameter introduces a new trade-off: \( q \searrow 0 \) yields a more sparse solution but slows down the convergence, whereas \( q \nearrow 1 \) ameliorates the convergence while sacrificing sparsity. One may obtain the solution faster than the squared 2-norm regularizer used in Blondel et al. [20], which corresponds to the case \( q = 0 \), by modulating the deformation parameter \( q \).

In regularized OT, it is a common approach to use weaker regularization (i.e., a smaller \( \lambda \)) to obtain a solution sparser and closer to the unregularized solution; however, a smaller \( \lambda \) results in numerical instability and slow computation [37]. This can be observed from Equation (33) because a smaller \( \lambda \) drives its upper bound considerably large.

Subsequently, we compared the computational complexity of \( q \)-DOT with the BFGS method and Sinkhorn algorithm. Altschuler et al. [25] showed that the Sinkhorn algorithm satisfies coupling constraints within the \( \ell_1 \) error \( \varepsilon \) in \( O(N^2(\log N)\varepsilon^{-3}) \) time, which is the sublinear convergence rate. In contrast, our convergence rate in Equation (33) is translated into the iteration complexity \( K = O(\log (Ne^{-1})) \), where \( \|g^{(K)}\|_2 \leq \varepsilon \). The gradient of \( \tilde{F} \) is

\[
\nabla \tilde{F}(z) = \begin{bmatrix}
\alpha_i - \sum_{j=1}^m \nabla \Omega^*(-D_{ij} - \alpha_i - \beta_j) + \kappa \alpha_i \\
\beta_j - \sum_{i=1}^n \nabla \Omega^*(-D_{ij} - \alpha_i - \beta_j) + \kappa \beta_j
\end{bmatrix},
\]

and \( \nabla \Omega^*(\cdot) \) represents the mapping from the dual variables \( (\alpha_i, \beta_j) \) to the primal transport matrix \( \Pi_{ij} \) in Equation (15). Therefore, the gradient norm of \( \tilde{F} \) and coupling constraint error are comparable when the \( \ell_2 \)-regularization parameter \( \kappa \) is sufficiently small. The overall computational complexity is \( O(N^2 \log (Ne^{-1})) \) because the one step of Algorithm 1 runs in \( O(N^2) \) time; this is the linear convergence rate. To confirm the one step of Algorithm 1 runs in \( O(N^2) \) time, we note that the update direction can be computed with \( O(N^2) \) time using the Sherman–Morrison formula to invert \( B^{(k)} \). In addition, the Hessian estimate can be updated with \( O(N^2) \) time because \( B^{(k)} \) is the rank-1 update and the computation of its inverse only requires the matrix-vector products of size \( N \). Hence, Algorithm 1 exhibits better convergence in terms of the stopping criterion \( \varepsilon \). The comparison is summarized in Table 2.

### Table 2. Comparison of the computational complexity of the Sinkhorn algorithm and deformed \( q \)-optimal transport. \( N = \max\{n, m\} \).

|                | \( \text{Sinkhorn} \) | \( \text{\( q \)-DOT} \) |
|----------------|------------------------|-------------------------|
|                | \( O(N^2(\log N)\varepsilon^{-3}) \) | \( O(N^2 \log (Ne^{-1})) \) |

4.3. Proofs

To prove Theorem 1, we leveraged several lemmas shown below. Lemma 2 is based on Powell [24] and Byrd et al. [36]. The missing proofs are provided in Appendix C.

**Lemma 1.** For the initial point \( z^{(0)} \) and sequence \( z^{(1)}, z^{(2)}, \ldots, z^{(K)} \) obtained by Algorithm 1, we define the following set and its bound:

\[
Z := \text{conv} \left\{ \left\{ z^{(0)}, z^{(1)}, z^{(2)}, \ldots, z^{(K)} \right\} \right\}, \quad R := \sup_{x \in Z} \nabla \Omega^*(-D_{ij} - \alpha_i - \beta_j),
\]

where \( \text{conv}(S) \) represents the convex hull of the set \( S \). Then, \( \tilde{F} : \mathbb{R}^{n+m} \to \mathbb{R} \) is \( M_1 \) strongly convex and \( M_2 \)-smooth over \( \bar{Z} \), where \( M_1 = \kappa \) and \( M_2 \leq \kappa + 2N\varepsilon^{-1} \). Moreover, \( \tilde{F} \) is \( M_2' \)-smooth over \( \bar{Z}_{\varepsilon} \) (defined in Equation (32)), where \( M_2' \leq \kappa + 2N\tau^{1/2} \).
Lemma 2. Let $z^{(1)}, z^{(2)}, \ldots, z^{(k)}$ be a sequence generated by Algorithm 1 given an initial point $z^{(0)}$. In addition, let $c_1, c_2, c_3, c_4,$ and $c_5$ be the constants
\[
c_1 := \frac{1 - \gamma}{M_2}, \quad c_2 := \frac{n + m}{K} + M_2, \quad c_3 := \left(\frac{K}{n + m}\right)^{\alpha / 2}, \quad c_4 := \frac{c_3}{1 - \gamma}, \quad c_5 := \frac{2(1 - \gamma')}{M_1}.
\]

Then,
\[
\tilde{F}(z^{(k)}) - \tilde{F}_* \leq \left(1 - \frac{\gamma' c_1 M_1}{2 c_3 c_5^2}\right)^{K/2} (\tilde{F}(z^{(0)}) - \tilde{F}_*).
\]

Lemma 3. Let $c_1, c_2, c_3, c_4,$ and $c_5$ be the same constants defined in Lemma 2. Then,
\[
\frac{\gamma' c_1 M_1}{c_3 c_5^2} > \frac{(1 - \gamma)^3 \gamma' e^{-2(n + m)/e}}{4(1 - \gamma')^2} \left(\frac{M_1}{M_2}\right)^3.
\]

Proof of Theorem 1. Because $\tilde{F}$ is differentiable and strongly convex, there exists an optimum $z_*$ such that $\nabla \tilde{F}(z_*) = 0$; this implies $\|g^{(K)}\|_2 = \|g^{(K)} - g_*\|_2$.

By using Assumption 1 and Lemma 1, we obtain $\|g^{(k)} - g_*\|_2 = \|\tilde{g}(z^{(k)}) - \nabla \hat{F}(z^{(k)}) - \nabla \hat{F}(z_*)\|_2 \leq M_2\|z^{(k)} - z_*\|_2$. In addition, $\|z^{(k)} - z_*\|_2 \leq \frac{2}{M_1} (\tilde{F}(z^{(k)}) - \hat{F}_*)$ as $\tilde{F}$ is $M_1$ strongly convex over $Z$ and the stationary condition $\nabla \hat{F}(z_*) = 0$ holds. We obtain the convergence bound by using Lemmas 2 and 3 as
\[
\|g^{(k)}\|_2 = \|g^{(k)} - g_*\|_2 \\
\leq M_2\|z^{(k)} - z_*\|_2 \\
\leq M_2 \sqrt{\frac{2(\tilde{F}(z^{(k)}) - \hat{F}_*)}{M_1}} \\
\leq M_2 \sqrt{\frac{2(\tilde{F}(z^{(0)}) - \hat{F}_*)}{M_1} \left(1 - \frac{\gamma' c_1 M_1}{2 c_3 c_5^2}\right)^{K/2}} \\
< M_2 \sqrt{\frac{2(\tilde{F}(z^{(0)}) - \hat{F}_*)}{M_1} \left(1 - \frac{(1 - \gamma)^3 \gamma' e^{-2(n + m)/e}}{8(1 - \gamma')^2} \left(\frac{M_1}{M_2}\right)^3\right)^{K/2}} \\
\leq \left(\frac{\kappa + 2N\tau^q}{\varepsilon}\right) \sqrt{\frac{2(\tilde{F}(z^{(0)}) - \hat{F}_*)}{\kappa} \left(1 - \frac{C}{(1 + 2NR\tau^q\varepsilon^{-1})^3}\right)^{K/2}},
\]
where we define $\kappa := \frac{(1 - \gamma)^3 \gamma' e^{-2(n + m)/e}}{8(1 - \gamma')^2}$ and Lemma 1 is used at the last inequality to replace $M_1$, $M_2$, and $M_3$. We can immediately confirm $C \leq \frac{1}{16}$ from $0 < \gamma' < \gamma < 1, \gamma' < \frac{1}{2},$ and $e^{-2(n + m)/e} < 1$. Finally, by choosing $\kappa = \frac{2N\tau^q}{\varepsilon}$,
\[
\|g^{(k)}\|_2 \leq \sqrt{\frac{16(\tilde{F}(z^{(0)}) - \hat{F}_*)N\tau^q}{\kappa} \left(1 - \frac{C}{(1 + (R/\tau)^q)^3}\right)^{K/2}} \\
\leq \sqrt{\frac{16(\tilde{F}(z^{(0)}) - \hat{F}_*)N\tau^q}{\kappa} r^K},
\]
where we use $(R/\tau)^q \geq 1$ (owing to $R \geq \tau$ by definition) and let $r := (1 - C/8)^{1/4}$ and $\sqrt{\frac{127}{128}} \leq r < 1$. \qed
Remark 2. More precisely, Altschuler et al. [25] showed that the Sinkhorn algorithm converges in $O(N^2L^3(\log N)\epsilon^{-3})$ time, where $L := \|D\|_{\infty}$. For $q$-DOT, its computational complexity is not directly comparable to that of the Sinkhorn in $L$; instead, the following analysis provides us a qualitative comparison. First, the convergence rate of $q$-DOT in Equation (33) is translated into the iteration complexity $K = O((\log(N\epsilon^{-1})/\log(1/r)))$. The rate $r$ is introduced in the proof of Theorem 1 (see Equation (40)): $r \geq \left(1 - \frac{\epsilon}{1+(\log(1/r))^2}\right)^{1/4}$. Then, by the Taylor expansion, we have a rough estimate $K \approx O(N^2R^{-\delta}\log(N\epsilon^{-1}))$, where $R$ is a bound on the possible primal variables defined in Equation (35). We cannot directly compare $R^{-\delta}$ and $L$; nevertheless, $R^{-\delta}$ and $L$ can be considered in the same magnitude given a reasonably sized domain $Z$, noting that $\nabla \Omega(\pi) \approx O(\pi^{1-\delta})$. Hence, it is reasonable to suppose that both the Sinkhorn algorithm and $q$-DOT roughly converge in cubic time with respect to $L$.

5. Numerical Experiments

5.1. Sparsity

All the simulations described in this section were executed on a 2.7 GHz quad-core Intel® Core™ i7 processor. We used the following synthetic dataset: $(x_i)^m_{i=1} \sim N(\mathbf{1}_2, \mathbf{I}_2)$, $(y_i)^m_{i=1} \sim N(-\mathbf{1}_2, \mathbf{I}_2)$, and $n = m = 30$, where $N(\mu, \Sigma)$ represents the Gaussian distribution with mean $\mu$ and covariance $\Sigma$. For each of the unregularized OTs, $q$-DOT, and Sinkhorn algorithm, we computed the transport matrices. For $q$-DOT and the Sinkhorn algorithm, different regularization parameters $\lambda$ were compared: $\lambda \in \{1 \times 10^{-2}, 1 \times 10^{-1}, 1\}$; and $\epsilon = 1 \times 10^{-6}$ was used as the stopping criterion: $q$-DOT stopped after the gradient norm was less than $\epsilon$, and the Sinkhorn algorithm stopped after the $l_1$ error of the coupling constraints was less than $\epsilon$. We compared different deformation parameters $q \in \{0, 0.25, 0.5, 0.75\}$ and fixed the dual $l_2$-regularization parameter $\kappa = 1 \times 10^{-6}$ for $q$-DOT. The $q$-DOT with $q = 0$ corresponded to a regularized OT with the squared 2-norm proposed by Blondel et al. [20]. For the unregularized OT, we used the implementation of the Python optimal transport package [38]. For $q$-DOT, we used the L-BFGS-B method (instead of the vanilla BFGS) provided by the SciPy package [39]. To determine zero entries in the transport matrix, we did not impose any positive threshold to disregard small values (as in Swanson et al. [6]) but regarded entries smaller than machine epsilon as zero.

The simulation results are shown in Table 3 and Figure 3. First, we qualitatively evaluated each method by using Figure 3 such that $q$-DOT obtained a very similar transport matrix to the unregularized OT solution. The solution was slightly blurred with increases in $q$ and $\lambda$. In contrast, the Sinkhorn algorithm output considerably uncertain transport matrices. Furthermore, the Sinkhorn algorithm was numerically unstable with a very small regularization such as $\lambda = 0.01$.

From Table 3, we further quantitatively observed the behavior. The transport matrices obtained by $q$-DOT were very sparse in most cases, and the sparsity was close to that of the unregularized OT. Furthermore, we observed the tendency such that smaller $q$ and $\lambda$ yielded a sparser solution. Significantly, the Sinkhorn algorithm obtained completely dense matrices (sparsity = 0). Although the transport matrices of $q$-DOT with $(q, \lambda) = (0.5, 1), (0.75, 1)$ appear somewhat similar to the Sinkhorn solutions in Figure 3, the former is much sparser. This suggests that a deformation parameter $q$ slightly smaller than 1 is sufficient for $q$-DOT to output a sparse transport matrix.
Figure 3. Comparison of transport matrices. Wasserstein represents the result of the unregularized OT. Sinkhorn ($\lambda = 0.01$) does not work well because of numerical instability.

For the obtained cost values $\langle \mathbf{D}, \mathbf{\hat{\Pi}} \rangle$, we did not see a clear advantage of using a specific $q$ and $\lambda$ from the results of $q$-DOT. Nevertheless, it is evident that $q$-DOT more accurately estimated the Wasserstein cost than the Sinkhorn algorithm regardless of the $q$ and $\lambda$ used in this simulation.
Table 3. Comparison of the sparsity and cost with the synthetic dataset. Sparsity indicates the ratio of zero entries in each transport matrix. We counted the number of entries smaller than machine epsilon to measure the sparsity instead of imposing a small positive threshold for determining zero entries. Sinkhorn ($\lambda = 0.01$) does not work well because of numerical instability.

| Method                  | Sparsity | Cost ($D, \tilde{D}$) |
|-------------------------|----------|-----------------------|
| Wasserstein (unregularized) | 0.967    | 7.126                 |
| $q$-DOT ($q = 0.00, \lambda = 0.01$) | 0.962    | 7.129                 |
| $q$-DOT ($q = 0.00, \lambda = 0.10$) | 0.950    | 7.144                 |
| $q$-DOT ($q = 0.25, \lambda = 0.01$) | 0.963    | 7.129                 |
| $q$-DOT ($q = 0.25, \lambda = 0.10$) | 0.959    | 7.126                 |
| $q$-DOT ($q = 0.25, \lambda = 1.00$) | 0.912    | 7.133                 |
| $q$-DOT ($q = 0.50, \lambda = 0.01$) | 0.963    | 7.136                 |
| $q$-DOT ($q = 0.50, \lambda = 0.10$) | 0.946    | 7.127                 |
| $q$-DOT ($q = 0.50, \lambda = 1.00$) | 0.879    | 7.155                 |
| $q$-DOT ($q = 0.75, \lambda = 0.01$) | 0.948    | 7.127                 |
| $q$-DOT ($q = 0.75, \lambda = 0.10$) | 0.897    | 7.136                 |
| $q$-DOT ($q = 0.75, \lambda = 1.00$) | 0.647    | 7.245                 |
| Sinkhorn ($\lambda = 0.01$) | —        | —                     |
| Sinkhorn ($\lambda = 0.10$) | 0.000    | 7.164                 |
| Sinkhorn ($\lambda = 1.00$) | 0.000    | 7.788                 |

5.2. Runtime Comparison

We compared the runtimes of $q$-DOT and the Sinkhorn algorithm using the same dataset as in Section 5.1, but with different dataset sizes: we chose $n = m \in \{100, 300, 500, 1000\}$. The parameter choices were the same as in Section 5.1, except that the regularization parameter was fixed to $\lambda = 0.1$. The result is shown in Figure 4; the larger deformation parameter $q$ makes $q$-DOT converge faster when $n = m = 100$. When $n = m \geq 300$, the difference between $q = 0$, $q = 0.25$, and $q = 0.5$ was not as evident. This may be partly because we fixed the parameter choice $\kappa = 1 \times 10^{-6}$ for all the experiments, unlike the oracle parameter choice $\kappa = 2N\tau^4\lambda^{-1}$ (in Theorem 1) depending on $q$. Nonetheless, $q = 0.75$ is clearly superior to the smaller $q$. From these observations, the trade-off between the sparsity and computation speed resulting from the deformation parameter $q$ is theoretically established in Theorem 1 and it was empirically observed.

Figure 4. Cont.
5.3. Approximation of 1-Wasserstein Distance

Finally, we compared the approximation errors of the 1-Wasserstein distance $|\langle D, \hat{\Pi} \rangle - \langle D, \Pi^* \rangle|$ of q-DOT and the Sinkhorn algorithm with different $q$ and $\lambda$, where $\hat{\Pi}$ represents the computed transport matrix and $\Pi^* \in \arg \min_{\Pi \in L(\mu, \nu)} \langle D, \Pi \rangle$ represents the LP solution. We used the same dataset and stopping criterion $\varepsilon$ as described in Section 5.1. For the range of $q$, we used $q \in \{0.00, 0.25, 0.50, 0.75\}$. For the range of $\lambda$, we used $\lambda \in \{0.05, 0.1, 0.5\}$.

The result is shown in Figure 5. The difference was not significant when $q$ was small, such as $q \in \{0.00, 0.25\}$. Once $q$ became larger, such as $q \in \{0.50, 0.75\}$, the approximation error evidently worsened. The Sinkhorn algorithm always exhibited worse approximation errors than q-DOT with $q$ in the range used in this simulation regardless of $\lambda$. Formal guarantees for the 1-Wasserstein approximation error (such as Altschuler et al. [25] and Weed [40]) will be considered in future work.

![Figure 4. Runtime comparison of q-DOT and Sinkhorn algorithm ($q = 1$). The error bars indicate the standard errors of 20 trials.](image)

![Figure 5. Wasserstein approximation error of q-DOT and the Sinkhorn algorithm ($q = 1$). The line shades indicate the standard errors of 20 trials.](image)

**Author Contributions:** Conceptualization, H.B.; methodology, H.B.; validation, H.B. and S.S.; formal analysis, H.B. and S.S.; writing—original draft preparation, H.B.; writing—review and editing, H.B. and S.S.; funding acquisition, H.B. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the Hakubi Project, Kyoto University, and JST ERATO Grant JPMJER1903. The APC was covered by the Hakubi Project.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.
Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations
The following abbreviations are used in this manuscript:

| Abbreviation           | Description                                      |
|------------------------|--------------------------------------------------|
| BFGS                   | Broyden–Fletcher–Goldfarb–Shannon                |
| q-DOT                  | Deformed q-optimal transport                      |
| L-BFGS                 | Limited-memory BFGS                               |
| OT                     | Optimal transport                                 |

Appendix A. Derivation of Deformed $q$ Entropy

Given a functional relationship $\pi = \nabla^{\star} \Omega (\eta) = \exp_q (\eta / \lambda)$ in Equation (20), we derive the deformed $q$ entropy.

First, the derivative of the regularizer $\nabla \Omega$ is simply the inverse of the dual map $\nabla^{\star} \Omega$ by Danskin’s theorem [27]; hence, $\nabla \Omega (\pi) = \lambda \log_q (\pi)$. The (negative of) deformed $q$ entropy is recovered by integrating $\nabla \Omega$:

\[
\begin{align*}
\Omega (\pi) &= \lambda \int_0^p \log_q (p) dp = \lambda \int_0^p \frac{1 - q}{1 - q} dp = \frac{\lambda}{1 - q} \left( 2 - q - \pi \right) \\
&= \frac{\lambda}{2 - q} \left( \pi^{1 - q} - 1 - \pi \right) = \frac{\lambda}{2 - q} \left( \pi \log_q (\pi) - \pi \right). 
\end{align*}
\]

(A1)

Appendix B. Additional Lemmas

Note again that we let $M_1, M_2 > 0$ be the strong convexity and smoothness constants of $\tilde{F}$ over $\mathcal{Z}$, $N := \max \{ n, m \}$, and $z_\ast \in \arg\min_{z \in \mathcal{Z}} \tilde{F} (z)$.

Lemma A1. For all $k$,

\[ M_1 \| s^{(k)} \|^2 \leq z^{(k)} \top s^{(k)} \leq M_2 \| s^{(k)} \|^2. \]  

(A2)

In addition,

\[ \frac{\| \xi^{(k)} \|^2}{\xi^{(k)} \top s^{(k)}} \leq M_2. \]  

(A3)

Proof. Let $G^{(k)} := \int_0^1 \nabla^2 \tilde{F} (z^{(k)} + ts^{(k)}) dt$. Then, by the chain rule and the fundamental theorem of calculus,

\[ G^{(k)} s^{(k)} = \int_0^1 \frac{\partial \nabla \tilde{F} (z^{(k)} + ts^{(k)})}{\partial t} dt = \nabla \tilde{F} (z^{(k)} + s^{(k)}) - \nabla \tilde{F} (z^{(k)}) = g^{(k+1)} - g^{(k)} = z^{(k)}. \]  

(A4)

Because $\tilde{F}$ is $M_1$ strongly convex and $M_2$-smooth (over $\mathcal{Z}$), we have

\[ M_1 \| w \|^2 \leq w \top [\nabla^2 \tilde{F} (z)] w \leq M_2 \| w \|^2 \]  

(A5)

for all $z \in \mathcal{Z}$ and $w$. By choosing $z = z^{(k)} + ts^{(k)}$ and $w = s^{(k)}$, we have

\[ M_1 \| s^{(k)} \|^2 \leq \int_0^1 s^{(k)} \top [\nabla^2 \tilde{F} (z^{(k)} + ts^{(k)})] s^{(k)} dt = s^{(k)} \top G^{(k)} s^{(k)} = s^{(k)} \top z^{(k)} s^{(k)} \leq M_2 \| s^{(k)} \|^2. \]  

(A6)

Note that $z^{(k)} + ts^{(k)} \in \mathcal{Z}$ follows by the definition of $\mathcal{Z}$ in Equation (35). Thus, the first statement is proven.
The second statement is proven as follows:

\[
\frac{\|\xi^{(k)}\|_2^2}{\zeta^{(k)\top}s^{(k)}} = \frac{s^{(k)\top}G^{(k)}s^{(k)}}{s^{(k)\top}G^{(k)}s^{(k)}} = \frac{(s^{(k)\top}G^{(k)}s^{(k)})^2}{\|G^{(k)}s^{(k)}\|_2^2} = \int_0^1 \frac{(s^{(k)\top})(\nabla^2 \tilde{F}(z^{(k)}) + is^{(k)})(s^{(k)\top})}{\|s^{(k)\top}\|_2^2} \, dt \leq M_2,
\]

where \(s^{(k)\top} = G^{(k)1/2}s^{(k)}\). □

**Lemma A2.** For all \(k\),

\[
\frac{M_1}{2} \|z^{(k)} - z_*\|_2 \leq \|g^{(k)}\|_2.
\]

**Proof.** Because \(\tilde{F}\) is \(M_1\) strongly convex over \(Z\),

\[
\frac{M_1}{2} \|z^{(k)} - z_*\|_2 \leq \tilde{F}(z^{(k)}) - \tilde{F}(z_*) - \langle \nabla \tilde{F}(z^{(k)}), z^{(k)} - z_* \rangle 
\]

\[
\leq \|B^{(k)}\|_2 \|z^{(k)} - z_*\|_2,
\]

where it follows from the optimality of \(z_*\) and the Cauchy–Schwarz inequality. □

**Lemma A3.** The following equations hold:

\[
\det(B^{(k)}) \leq \left(\frac{c_2 K}{n + m}\right)^{n+m},
\]

\[
\prod_{k=0}^{K-1} \frac{\|B^{(k)}s^{(k)}\|_2^2}{\|s^{(k)\top}B^{(k)}s^{(k)}\|_2^2} \leq c_2,
\]

\[
\frac{\det(B^{(k)})}{\det(B^{(0)})} = \prod_{k=0}^{K-1} \frac{\zeta^{(k)\top}s^{(k)}}{s^{(k)\top}B^{(k)}s^{(k)}},
\]

where \(c_2 := \frac{n + m}{K} + M_2\) is defined in Lemma 2.

**Proof.** To prove Equation (A10), we use the linearity of the trace and \(\text{tr}(ba^\top) = a^\top b\) to evaluate \(\text{tr}(B^{(k+1)})\) as follows:

\[
\text{tr}(B^{(k+1)}) = \text{tr} \left( B^{(k)} - \frac{B^{(k)}s^{(k)}s^{(k)\top}B^{(k)}}{s^{(k)\top}B^{(k)}s^{(k)}} + \frac{\zeta^{(k)}\zeta^{(k)\top}}{\zeta^{(k)\top}s^{(k)}} \right)
\]

\[
= \text{tr}(B^{(k)}) - \text{tr} \left( \frac{B^{(k)}s^{(k)}s^{(k)\top}B^{(k)}}{s^{(k)\top}B^{(k)}s^{(k)}} \right) + \text{tr} \left( \frac{\zeta^{(k)}\zeta^{(k)\top}}{\zeta^{(k)\top}s^{(k)}} \right)
\]

\[
\leq \text{tr}(B^{(k)}) + \frac{\|\zeta^{(k)}\|_2^2}{\zeta^{(k)\top}s^{(k)}}
\]

\[
\leq \text{tr}(B^{(0)}) + \sum_{j=0}^{k} \frac{\|\zeta^{(j)}\|_2^2}{\zeta^{(j)\top}s^{(j)}}
\]

\[
\leq \text{tr}(B^{(0)}) + (k + 1)M_2,
\]

where Lemma A1 is used at the last inequality. Note that the trace is the sum of the eigenvalues, whereas the determinant is the product of the eigenvalues. Then, we can use the AM–GM inequality to translate the determinant into the trace as follows:
\[
\det(B^{(k+1)}) \leq \left( \frac{1}{n + m} \text{tr}(B^{(k+1)}) \right)^{n+m} \leq \left( \frac{\text{tr}(B^{(0)}) + M_2(k + 1)}{n + m} \right)^{n+m}.
\] (A14)

Hence, by substituting \( k = K - 1 \) and \( \text{tr}(B^{(0)}) = n + m \), Equation (A10) is proven.

To prove Equation (A11), we evaluate \( \text{tr}(B^{(k+1)}) \) in a way similar to that for Equation (A13). From Lemma A1,

\[
0 \leq \text{tr}(B^{(k+1)}) = \text{tr}(B^{(k)}) - \frac{\|B^{(k)}s^{(k)}\|_2^2}{s^{(k)} \cdot B^{(k)}s^{(k)}} + \frac{\|\zeta^{(k)}\|_2^2}{\zeta^{(k)} \cdot s^{(k)}}.
\]

By the AM–GM inequality,

\[
\prod_{j=0}^k \frac{\|B^{(j)}s^{(j)}\|_2^2}{s^{(j)} \cdot B^{(j)}s^{(j)}} \leq \left( \frac{1}{k+1} \sum_{j=0}^k \frac{\|B^{(j)}s^{(j)}\|_2^2}{s^{(j)} \cdot B^{(j)}s^{(j)}} \right)^{k+1}.
\] (A16)

Hence, by substituting \( k = K - 1 \) and \( \text{tr}(B^{(0)}) = n + m \), Equation (A11) is proven.

To prove Equation (A12), we use the matrix determinant lemma to expand \( \det(B^{(k+1)}) \) as follows:

\[
\det(B^{(k+1)}) = \det \left( B^{(k)} - \frac{B^{(k)}s^{(k)}(s^{(k)} \cdot B^{(k)})}{s^{(k)} \cdot B^{(k)}s^{(k)}} + \frac{\zeta^{(k)}(\zeta^{(k)} \cdot s^{(k)})}{\zeta^{(k)} \cdot s^{(k)}} \right)
\]

\[
= \left\{ 1 - \frac{1}{s^{(k)} \cdot B^{(k)}s^{(k)}} \cdot s^{(k)} \cdot B^{(k)} \left( B^{(k)} + \frac{\zeta^{(k)}(\zeta^{(k)} \cdot s^{(k)})}{\zeta^{(k)} \cdot s^{(k)}} \right)^{-1} B^{(k)}s^{(k)} \right\}
\]

\[
\cdot \det \left( B^{(k)} + \frac{\zeta^{(k)}(\zeta^{(k)} \cdot s^{(k)})}{\zeta^{(k)} \cdot s^{(k)}} \right).
\] (A17)

Further, by the Sherman–Morrison formula, we have

\[
\left( B^{(k)} + \frac{\zeta^{(k)}(\zeta^{(k)} \cdot s^{(k)})}{\zeta^{(k)} \cdot s^{(k)}} \right)^{-1} = B^{(k)} - \frac{B^{(k)-1}(\zeta^{(k)}(\zeta^{(k)} \cdot B^{(k)}s^{(k)}))^{-1}}{\zeta^{(k)} \cdot s^{(k)} + \zeta^{(k)}(\zeta^{(k)} \cdot B^{(k)}s^{(k)})^{-1} \zeta^{(k)}}.
\] (A18)

By plugging Equation (A18) into Equation (A17), we have

\[
\det(B^{(k+1)}) = \frac{(s^{(k)} \cdot B^{(k)}s^{(k)})^2}{(s^{(k)} \cdot B^{(k)}s^{(k)})^2 - (s^{(k)} \cdot B^{(k)}s^{(k)}) \zeta^{(k)}(\zeta^{(k)} \cdot B^{(k)-1} \zeta^{(k)})} \det \left( B^{(k)} + \frac{\zeta^{(k)}(\zeta^{(k)} \cdot s^{(k)})}{\zeta^{(k)} \cdot s^{(k)}} \right)
\]

\[
= \frac{(s^{(k)} \cdot B^{(k)}s^{(k)})^2}{(s^{(k)} \cdot B^{(k)}s^{(k)})^2 - (s^{(k)} \cdot B^{(k)}s^{(k)}) \zeta^{(k)}(\zeta^{(k)} \cdot B^{(k)-1} \zeta^{(k)})}
\]

\[
\cdot \left( 1 + \frac{\zeta^{(k)}(\zeta^{(k)} \cdot B^{(k)-1} \zeta^{(k)})}{\zeta^{(k)} \cdot s^{(k)}} \right) \det(B^{(k)})
\]

\[
= \det(B^{(k)}) \frac{\zeta^{(k)}(\zeta^{(k)} \cdot s^{(k)})}{s^{(k)} \cdot B^{(k)}s^{(k)}},
\]

where the matrix determinant lemma is invoked again at the second identity. Recursively applying Equation (A19) with \( \det(B^{(0)}) = 1 \), we obtain Equation (A12).
Lemma A4. For $k$, $\|s^{(k)}\|_2 \leq c_5 \|g^{(k)}\|_2 \cos \theta_k$, where $\theta_k$ is the angle between $s^{(k)}$ and $-g^{(k)}$, and $c_5 := \frac{2(1-\gamma')}{M_1}$ is defined in Lemma 2.

Proof. By the Armijo condition (30), we have

$$\tilde{F}(z^{(k+1)}) - \tilde{F}(z^{(k)}) \leq \gamma' \rho^{(k)} g^{(k)\top} d^{(k)} = \gamma' g^{(k)\top} s^{(k)}.$$  \hfill (A20)

Additionally, as $\tilde{F}$ is $M_1$-strongly convex over $Z$, it holds that $\tilde{F}(z^{(k+1)}) - \tilde{F}(z^{(k)}) \geq s^{(k)\top} g^{(k)} + \frac{1}{2} M_1 \|s^{(k)}\|^2_2$. Hence,

$$s^{(k)\top} g^{(k)} + \frac{1}{2} M_1 \|s^{(k)}\|^2_2 \leq \gamma' g^{(k)\top} s^{(k)}$$

$$\Rightarrow (1 - \gamma')(s^{(k)\top} g^{(k)}) \geq \frac{1}{2} M_1 \|s^{(k)}\|^2_2$$

$$\Rightarrow \|s^{(k)}\|_2 \leq \sqrt{\frac{2(1-\gamma')}{M_1} \|s^{(k)\top} g^{(k)}\|^2_2} = \cos \theta_k$$

which is the desired inequality. \hfill \qed

Lemma A5. For $k$, let $\theta_k$ be the angle between $s^{(k)}$ and $-g^{(k)}$. Then,

$$\frac{K-1}{k=0} \left(1 - \frac{\gamma' c_1 M_1 \cos^2 \theta_k}{2}\right) \leq \left(1 - \frac{\gamma' c_1 M_1}{2 c_4^2 c_5^2}\right)^{K/2},$$  \hfill (A22)

where $c_1$, $c_4$, and $c_5$ are defined in Lemma 2.

Proof. By multiplying each side of Equations A10–A12, we have

$$\prod_{k=0}^{K-1} \frac{\|B^{(k)}s^{(k)}\|^2_2}{s^{(k)\top} B^{(k)} s^{(k)}} \cdot \frac{\zeta^{(k)\top} s^{(k)}}{s^{(k)\top} B^{(k)} s^{(k)}} \leq c_3^K,$$  \hfill (A23)

where $c_3 := (\frac{K}{n+m})^{(n+m)/K} c_2^{-K}$ is defined in Lemma 2. By using $B^{(k)} s^{(k)} = -\rho^{(k)} g^{(k)}$ and $\zeta^{(k)\top} s^{(k)} \geq -(1-\gamma) g^{(k)\top} s^{(k)}$ (shown in Equation (A33)),

$$\prod_{k=0}^{K-1} \frac{\|B^{(k)}s^{(k)}\|^2_2}{s^{(k)\top} B^{(k)} s^{(k)}} \cdot \frac{\zeta^{(k)\top} s^{(k)}}{s^{(k)\top} B^{(k)} s^{(k)}} = \prod_{k=0}^{K-1} \frac{\|g^{(k)}\|^2_2 \zeta^{(k)\top} s^{(k)}}{(-s^{(k)\top} g^{(k)})^2}$$

$$\geq (1-\gamma)^K \prod_{k=0}^{K-1} \frac{\|g^{(k)}\|^2_2}{-s^{(k)\top} g^{(k)}}.$$  \hfill (A24)

Hence,

$$\prod_{k=0}^{K-1} \frac{\|g^{(k)}\|^2_2}{s^{(k)\top} g^{(k)}} \leq \left(\frac{c_3}{1-\gamma}\right)^K = c_4^K.$$  \hfill (A25)

By Lemma A4, we can confirm

$$\prod_{k=0}^{K-1} \cos^2 \theta_k \geq \prod_{k=0}^{K-1} \frac{\|g^{(k)}\|^2_2 \cos \theta_k}{s^{(k)\top} g^{(k)}} \geq \left(\frac{1}{c_4 c_5}\right)^K.$$  \hfill (A26)

Let $\bar{K}$ be the number of $k = 0, 1, \ldots, K-1$ such that $\cos \theta_k \leq \frac{1}{c_4 c_5}$, then

$$\left(\frac{1}{c_4 c_5}\right)^K \leq \prod_{k=0}^{K-1} \cos^2 \theta_k \leq \left(\frac{1}{c_4 c_5}\right)^{2\bar{K}}.$$  \hfill (A27)
implying that $\tilde{\kappa}$ is at most $\frac{\delta}{4}$ (note that $\frac{1}{c_4c_5} < 1$ from Equation (A26)). Therefore,

$$\prod_{k=0}^{K-1} \left( 1 - \frac{\gamma^* c_1 M_1 \cos^2 \theta_k}{2} \right) \leq \left( 1 - \frac{\gamma^* c_1 M_1}{2c_4^2 c_5^2} \right)^{K/2}.$$  \hfill (A28)

\[\Box\]

Appendix C. Deferred Proofs

Appendix C.1. Proof of Lemma 1

**Proof.** It is easy to confirm $M_1 = \kappa$ because $\tilde{\mathcal{F}}$ is the sum of $\mathcal{F}$ (convex) and $\frac{\kappa}{2} \|z\|^2_2$.

Because $\tilde{\mathcal{F}}$ is twice differentiable and $\mathcal{Z}$ is a closed convex set, we evaluate the smoothness parameter $M_2$ (over $Z$) by the eigenvalues of $\nabla^2 \tilde{\mathcal{F}}(z)$. We begin by evaluating the eigenvalues of $\nabla^2 \tilde{\mathcal{F}}(z)$, then evaluate the eigenvalues of $\nabla^2 \mathcal{F}(z)$ by $\nabla^2 \tilde{\mathcal{F}}(z) = \nabla^2 \mathcal{F}(z) + \kappa I$. Let $P(z) \in \mathbb{R}^{n \times n}$ be a matrix such that $P_{ij}(z) := \nabla \Omega^\tau(-D_{ij} - \alpha_i - \beta_j)$. Here, $P_{ij}(z)$ is the primal variable corresponding to the dual variables $(\alpha_i, \beta_j)$ (see Equation (15)). The gradient of $\mathcal{F}$ is

$$\nabla \mathcal{F}(z) = \begin{bmatrix} a_i - \sum_{j=1}^m \nabla \Omega^\tau(-D_{ij} - \alpha_i - \beta_j) \\ \vdots \\ b_j - \sum_{l=1}^n \nabla \Omega^\tau(-D_{lj} - \alpha_j - \beta_j) \end{bmatrix} = \begin{bmatrix} a_i - \sum_{j=1}^m a_j \\ \vdots \\ b_j - \sum_{j=1}^m b_j \end{bmatrix},$$  \hfill (A29)

and the Hessian of $\mathcal{F}$ is

$$\nabla^2 \mathcal{F}(z) = \frac{1}{\lambda} \begin{bmatrix} \text{diag} \left( \sum_j P_{ij}(z)^\tau \right) & P(z)^\tau \\ \left( P(z)^\tau \right)^\top & \text{diag} \left( \sum_j P_{ij}(z)^\tau \right) \end{bmatrix} \cdots : H,$$  \hfill (A30)

where $P(z)^\tau$ is the element-wise power of $P(z)$. Then, by invoking the Gershgorin circle theorem (Theorem 7.2.1 of [41]), the eigenvalues of $H$ can be upper bounded by the following value:

$$\max \left\{ \sum_{i=1}^m P_{ij}(z)^\tau \text{ center of } i\text{-th disc} + \left[ (P(z)^\tau)^\top \mathbf{1}_m \right]_{j=1}^n \right\} \text{ radius of } i\text{-th disc} \leq \max \left\{ 2 \sum_{j=1}^m P_{ij}(z)^\tau, 2 \sum_{j=1}^n P_{ij}(z)^\tau \right\} \leq 2N^4\kappa,$$  \hfill (A31)

where we use $0 \leq P_{ij}(z) \leq \kappa$ for all $i, j, \text{ and } z \in Z$ at the last inequality. Hence, $M_2 \leq \kappa + \frac{2N^4\kappa}{\lambda}$.

$M_2 \leq \kappa + \frac{2N^4\kappa}{\lambda}$ is confirmed by noting that $0 \leq P_{ij}(z) \leq \tau$ for all $i, j, \text{ and } z \in Z_\tau$ and that $Z_\tau$ is a closed convex set. \[\Box\]

Appendix C.2. Proof of Lemma 2

**Proof.** First, we evaluate the ratio between $\tilde{\mathcal{F}}(z^{(k+1)}) - \mathcal{F}_*$ and $\tilde{\mathcal{F}}(z^{(k)}) - \mathcal{F}_*$ for $k = 0, 1, 2, \ldots, K - 1$. Let $\theta_k$ be the angle between the vectors $s^{(k)}$ and $-g^{(k)}$. By the Armijo condition (Equation (30)), the difference $\tilde{\mathcal{F}}(z^{(k+1)}) - \tilde{\mathcal{F}}(z^{(k)})$ can be evaluated as follows:
\[ \tilde{F}(z^{(k+1)}) - \tilde{F}(z^{(k)}) \leq \gamma' \mu(z) g^{(k)} \top d^{(k)} \]
\[ = \gamma' g^{(k)} \top (z^{(k+1)} - z^{(k)}) \]
\[ = \gamma' g^{(k)} \top s^{(k)} \]
\[ = \gamma' (-\|s^{(k)}\|_2 \|g^{(k)}\|_2 \cos \theta_k). \] (A32)

In addition, by the curvature condition (Equation (31)),
\[ \zeta^{(k)} \top s^{(k)} = g^{(k+1) \top} s^{(k)} - g^{(k)} \top s^{(k)} \]
\[ = \rho^{(k)} g^{(k+1) \top} d^{(k)} \geq \rho^{(k)} \gamma g^{(k)} \top d^{(k)} \]
\[ \geq \gamma g^{(k)} \top s^{(k)} - g^{(k)} \top s^{(k)} \]
\[ = -(1 - \gamma) g^{(k)} \top s^{(k)}, \] (A33)

which implies \(\|s^{(k)}\|_2^2 \geq \frac{1}{M^2} \zeta^{(k)} \top s^{(k)} \geq -\frac{1-\gamma}{M^2} g^{(k)} \top s^{(k)} = \frac{1-\gamma}{M^2} \|s^{(k)}\|_2 \|g^{(k)}\|_2 \cos \theta_k\) together with Lemma A1. Hence, we have
\[ \|s^{(k)}\|_2 \geq c_1 \|g^{(k)}\|_2 \cos \theta_k, \] (A34)
where \(c_1 := \frac{1-\gamma}{M^2}\). Then,
\[ \tilde{F}(z^{(k+1)}) - \tilde{F}_* \leq (\tilde{F}(z^{(k)}) - \tilde{F}_*) + \gamma' (-\|s^{(k)}\|_2 \|g^{(k)}\|_2 \cos \theta_k) \]
\[ \leq (\tilde{F}(z^{(k)}) - \tilde{F}_*) - \gamma' c_1 \|g^{(k)}\|_2^2 \cos^2 \theta_k \]
\[ \leq (\tilde{F}(z^{(k)}) - \tilde{F}_*) - \gamma' c_1 (M/2) \|g^{(k)}\|_2 \|z^{(k)} - z_*\|_2 \cos^2 \theta_k \]
\[ \leq (\tilde{F}(z^{(k)}) - \tilde{F}_*) - \gamma' c_1 (M/2) \cos^2 \theta_k (\tilde{F}(z^{(k)}) - \tilde{F}_*) \]
\[ = (1 - \gamma' c_1 M \cos^2 \theta_k/2) (\tilde{F}(z^{(k)}) - \tilde{F}_*), \] (A35)

where Equation (A32) is used at the first inequality; Equation (A34) is used at the second inequality; Lemma A2 is used at the third inequality; a consequence of the convexity \(\tilde{F}(z^{(k)}) - \tilde{F}(z_*)) \leq (g^{(k)} \top z^{(k)} - z_*) \leq \|g^{(k)}\|_2 \|z^{(k)} - z_*\|_2\) is used at the fourth inequality.

Next, recursively invoking the inequality Equation (A35), we obtain
\[ \tilde{F}(z^{(K)}) - \tilde{F}_* \leq \left( \prod_{k=0}^{K-1} \left( 1 - \frac{\gamma' c_1 M \cos^2 \theta_k}{2} \right) \right) \left( \tilde{F}(z^{(0)}) - \tilde{F}_* \right) \]
\[ \leq \left( 1 - \frac{\gamma' c_1 M}{2c_3 c_5} \right)^{K/2} \left( \tilde{F}(z^{(0)}) - \tilde{F}_* \right), \] (A36)
which is the desired bound. The last inequality is due to Lemma 3. \(\square\)

Appendix C.3. Proof of Lemma 3

Proof. By substituting the definitions of the constants \(c_1, c_2, c_3, c_4,\) and \(c_5,\)
where, at the first inequality, we invoke 

\[
\frac{1 - \gamma^3}{M_2} \gamma' = \frac{1 - \gamma^3}{M_2} \gamma' = \frac{M_2^3}{4M_2^2(1 - \gamma')^2} \left\{ \left( \frac{n + m}{K} \right)^{2(n+m)} K^{-\frac{2(n+m+K)}{K}} \right\}^2
\]

At the second inequality, hence, the desired

\[
\left\{ \left( \frac{n + m}{K} \right) + M_2 \right\} \cdot K^{-\frac{2(n+m+K)}{K}} \geq \left( M_2 K^{-\frac{2(n+m+K)}{K}} \right)^{-\frac{2(n+m+K)}{K}} = M_2^{-2} K^{-\frac{2(n+m)}{K}} K^{-\frac{2(n+m)}{K}} \geq M_2^{2} K^{-\frac{2(n+m)}{K}}
\]

and we use \( K^{-\frac{2(n+m)}{K}} \geq e^{-\frac{2(n+m)}{K}} \) for all \( K \) at the second inequality. Hence, the desired inequality is proven. □

Appendix D. Additional Experiments

Appendix D.1. Comparison with Tsallis Entropy

In this study, we used the deformed \( q \) entropy instead of the Tsallis entropy [23] as the sparse regularization. Here, we briefly empirically analyze what happens if we use the Tsallis entropy instead. We compare the dual optimization objective in Definition 2 with the deformed \( q \) entropy and Tsallis entropy. We use the following convex regularizer formed by the Tsallis entropy:

\[
\Omega(\pi) = \lambda \sum_{i=1}^{n} \pi_i^q \log_q(\pi_i).
\]
Table A1. Comparison of the sparsity and absolute error on the synthetic dataset. Sparsity indicates the ratio of zero entries in each transport matrix. We counted the number of entries smaller than machine epsilon to measure the sparsity instead of imposing a small positive threshold for determining zero entries. Abs. error indicates the absolute error of the computed cost with respect to 1-Wasserstein distance. Tsallis-regularized OT with \( q = 0.00 \) does not work due to numerical instability.

| \( q \) | \( \lambda \) | Sparsity (\( q \)-DOT) | Abs. Error (\( q \)-DOT) | Sparsity (Tsallis) | Abs. Error (Tsallis) |
|---|---|---|---|---|---|
| 0.00, \( \lambda = 0.50 \) | 0.984 | 0.001 | — | — |
| 0.00, \( \lambda = 1.00 \) | 0.981 | 0.011 | — | — |
| 0.25, \( \lambda = 0.50 \) | 0.977 | 0.008 | 0.000 | 3.362 |
| 0.25, \( \lambda = 1.00 \) | 0.973 | 0.010 | 0.000 | 3.388 |
| 0.50, \( \lambda = 0.50 \) | 0.959 | 0.015 | 0.000 | 3.153 |
| 0.50, \( \lambda = 1.00 \) | 0.944 | 0.022 | 0.000 | 3.283 |
| 0.75, \( \lambda = 0.50 \) | 0.861 | 0.052 | 0.000 | 1.962 |
| 0.75, \( \lambda = 1.00 \) | 0.776 | 0.099 | 0.000 | 2.582 |

As can be seen from the results in Table A1, the Tsallis entropic regularizer neither induces sparsity nor achieves a better approximation of the 1-Wasserstein distance than the deformed \( q \) entropy. Note that the Tsallis entropy induces the dual map \( \nabla \Omega^*(\eta) = q^{1/(1-q)} / \exp_q(-\eta/\lambda) \) shown in Equation (25), which has dense support for \( q > 0 \) and becomes the source of dense transport matrices. This verifies that the design of the regularizer is important for regularized optimal transport.

Appendix D.2. Hyperparameter Sensitivity

In this section, we summarize more comprehensive experimental results of \( q \)-DOT and the Sinkhorn algorithm to show the performance dependence on hyperparameters \( q \) and \( \lambda \). Subsequently, we describe experiments to show the sparsity of transport matrices, absolute error of computed costs with respect to 1-Wasserstein distance, and runtime with differently-sized datasets.

The simulations in this section were executed on a 2.7 GHz Intel® Xeon® Gold 6258R processor (different from the processor that we used in Section 5). We used the following synthetic dataset: \( (x_i)_{i=1}^n \sim \mathcal{N}(\mathbf{1}_2, \mathbf{I}_2) \), \( (y_j)_{j=1}^m \sim \mathcal{N}(-\mathbf{1}_2, \mathbf{I}_2) \), with different \( \mathcal{N}(=n=m) \in \{100, 300, 500, 1000, 2000, 3000\} \). For \( q \)-DOT and Tsallis-regularized OT, different regularization parameters \( \lambda \in \{0.01, 0.1, 1\} \) were compared, and \( \epsilon = 1 \times 10^{-6} \) was used as the stopping criterion. We compared different deformation parameters \( q \in \{0, 0.25, 0.5, 0.75\} \). For the unregularized OT, we used the implementation of the Python optimal transport package [38]. For \( q \)-DOT, we used the L-BFGS-B method provided by the SciPy package [39]. To determine zero entries in the transport matrix, we regarded entries smaller than machine epsilon as zero.

The results are shown in Table A2. In these tables, the results with \( q = 1.00 \) correspond to the Sinkhorn algorithm. The results for \( (q, \lambda) = (1.00, 0.01) \) are missing because they did not work well due to numerical instability. In general, we observed similar behavior as we described in Section 5: sparsity intensified as \( q \) and \( \lambda \) decreased, thereby increasing runtime. As \( N \) increased, nonmonotonic trends in runtime were observed with respect to \( q \); for a fixed \( \lambda \), larger \( q \) accelerated the computation, while \( q = 0.25 \) seemed to be the slowest. This apparent discrepancy from Theorem 1 may be partly because Theorem 1 relies on an oracle parameter choice \( \kappa = 2N\tau^4\lambda^{-1} \) as we discussed in Section 5.2, which is hardly known in practice. Nevertheless, it is remarkable that even \( q = 0.75 \) gives very sparse solutions with a reasonable amount of runtime. Regarding the absolute error, smaller \( q \) tends to perform better with relatively small datasets, such as \( N \leq 1000 \), while \( q = 1.00 \) performs better for larger datasets, such as \( N = 2000 \) and 3000. As we mentioned in Section 5.3, theoretical analysis of the approximation error is still unclear, and will be left for future work.
Table A2. Hyperparameter sensitivity of $q$-DOT and Sinkhorn algorithm. In these tables, $q = 1.00$ corresponds to the Sinkhorn algorithm. $(q, \lambda) = (1.00, 0.01)$ did not work well because of numerical instability. The results shown in the tables are the means of 10 random trials. Bold typeface indicates the best result for each of sparsity, absolute error, and runtime.

| $(N = 100)$ | Sparsity | Abs. error | Runtime [ms] |
|-------------|----------|------------|--------------|
| $q = 0.00, \lambda = 0.01$ | 0.999 | $2.28 \times 10^{-2}$ | 4366.142 |
| $q = 0.00, \lambda = 0.10$ | 0.988 | $3.63 \times 10^{-3}$ | 1236.346 |
| $q = 0.00, \lambda = 1.00$ | 0.982 | $6.20 \times 10^{-3}$ | 842.253 |
| $q = 0.25, \lambda = 0.01$ | 0.989 | $8.18 \times 10^{-3}$ | 3182.535 |
| $q = 0.25, \lambda = 0.10$ | 0.986 | $5.54 \times 10^{-3}$ | 1131.784 |
| $q = 0.25, \lambda = 1.00$ | 0.973 | $1.16 \times 10^{-2}$ | 668.734 |
| $q = 0.50, \lambda = 0.01$ | 0.987 | $9.91 \times 10^{-3}$ | 2388.176 |
| $q = 0.50, \lambda = 0.10$ | 0.977 | $7.66 \times 10^{-3}$ | 1040.818 |
| $q = 0.50, \lambda = 1.00$ | 0.946 | $2.40 \times 10^{-2}$ | 339.978 |
| $q = 0.75, \lambda = 0.01$ | 0.979 | $1.16 \times 10^{-2}$ | 2396.353 |
| $q = 0.75, \lambda = 0.10$ | 0.950 | $1.31 \times 10^{-2}$ | 731.564 |
| $q = 0.75, \lambda = 1.00$ | 0.876 | $1.02 \times 10^{-1}$ | 200.654 |
| $q = 1.00, \lambda = 0.01$ | 0.987 | $9.91 \times 10^{-3}$ | 2388.176 |
| $q = 1.00, \lambda = 0.10$ | 0.977 | $7.66 \times 10^{-3}$ | 1040.818 |
| $q = 1.00, \lambda = 1.00$ | 0.946 | $2.40 \times 10^{-2}$ | 339.978 |

| $(N = 100)$ | Sparsity | Abs. error | Runtime [ms] |
|-------------|----------|------------|--------------|
| $q = 0.00, \lambda = 0.01$ | 0.999 | $2.48 \times 10^{-2}$ | 86046.395 |
| $q = 0.00, \lambda = 0.10$ | 0.997 | $3.91 \times 10^{-2}$ | 49523.995 |
| $q = 0.00, \lambda = 1.00$ | 0.996 | $4.10 \times 10^{-2}$ | 27357.659 |
| $q = 0.25, \lambda = 0.01$ | 0.998 | $2.36 \times 10^{-1}$ | 104364.641 |
| $q = 0.25, \lambda = 0.10$ | 0.996 | $5.12 \times 10^{-2}$ | 41810.473 |
| $q = 0.25, \lambda = 1.00$ | 0.994 | $4.22 \times 10^{-2}$ | 18415.400 |
| $q = 0.50, \lambda = 0.01$ | 0.996 | $4.52 \times 10^{-2}$ | 78618.996 |
| $q = 0.50, \lambda = 0.10$ | 0.994 | $4.50 \times 10^{-2}$ | 25512.371 |
| $q = 0.50, \lambda = 1.00$ | 0.984 | $4.92 \times 10^{-2}$ | 8266.048 |
| $q = 0.75, \lambda = 0.01$ | 0.994 | $4.55 \times 10^{-2}$ | 57839.639 |
| $q = 0.75, \lambda = 0.10$ | 0.997 | $5.07 \times 10^{-2}$ | 14257.452 |
| $q = 0.75, \lambda = 1.00$ | 0.890 | $1.00 \times 10^{-1}$ | 4362.478 |
| $q = 1.00, \lambda = 0.01$ | — | — | — |
| $q = 1.00, \lambda = 0.10$ | 0.990 | $7.92 \times 10^{-2}$ | 5731.333 |
| $q = 1.00, \lambda = 1.00$ | 0.982 | $8.53 \times 10^{-1}$ | 562.722 |

| $(N = 100)$ | Sparsity | Abs. error | Runtime [s] |
|-------------|----------|------------|--------------|
| $q = 0.00, \lambda = 0.01$ | 1.00 | $3.59 \times 10^{0}$ | 1386.554 |
| $q = 0.00, \lambda = 0.10$ | 0.999 | $2.25 \times 10^{-1}$ | 1243.867 |
| $q = 0.00, \lambda = 1.00$ | 0.999 | $1.85 \times 10^{-1}$ | 823.011 |
| $q = 0.25, \lambda = 0.01$ | 1.00 | $5.88 \times 10^{0}$ | 1555.064 |
| $q = 0.25, \lambda = 0.10$ | 0.999 | $1.86 \times 10^{-1}$ | 1201.656 |
| $q = 0.25, \lambda = 1.00$ | 0.998 | $1.86 \times 10^{-1}$ | 492.324 |
| $q = 0.50, \lambda = 0.01$ | 0.999 | $6.66 \times 10^{-1}$ | 1494.270 |
| $q = 0.50, \lambda = 0.10$ | 0.998 | $1.97 \times 10^{-1}$ | 589.379 |
| $q = 0.50, \lambda = 1.00$ | 0.994 | $1.85 \times 10^{-1}$ | 210.008 |
| $q = 0.75, \lambda = 0.01$ | 0.998 | $2.00 \times 10^{-1}$ | 1300.517 |
| $q = 0.75, \lambda = 0.10$ | 0.998 | $2.00 \times 10^{-1}$ | 321.221 |
| $q = 0.75, \lambda = 1.00$ | 0.937 | $2.08 \times 10^{-1}$ | 106.334 |
| $q = 1.00, \lambda = 0.01$ | — | — | — |
| $q = 1.00, \lambda = 0.10$ | 0.990 | $9.06 \times 10^{-2}$ | 147.372 |
| $q = 1.00, \lambda = 1.00$ | 0.990 | $8.62 \times 10^{-1}$ | 8.575 |
References

1. Villani, C. Optimal Transport: Old and New; Springer: Berlin/Heidelberg, Germany, 2009; Volume 338.

2. Shafieezadeh-Abadeh, S.; Mohajerin Esfahani, P.M.; Kuhn, D. Distributionally robust logistic regression. Adv. Neural Inf. Process. Syst. 2015, 28. https://dl.acm.org/doi/10.5555/2969239.2969415.

3. Courty, N.; Flamary, R.; Habrard, A.; Rakotomamonjy, A. Joint distribution optimal transportation for domain adaptation. Adv. Neural Inf. Process. Syst. 2017, 30. https://dl.acm.org/doi/10.5555/3294996.3295130.

4. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein generative adversarial networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; PMLR, pp. 214–223.

5. Kusner, M.; Sun, Y.; Kolkin, N.; Weinberger, K. From word embeddings to document distances. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 7–9 July 2015; PMLR, pp. 957–966.

6. Swanson, K.; Yu, L.; Lei, T. Rationalizing text matching: Learning sparse embeddings via optimal transport. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 5609–5626.

7. Otani, M.; Togashi, R.; Nakashima, Y.; Rahtu, E.; Heikkilä, J.; Satoh, S. Optimal correction cost for object detection evaluation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 21–24 June 2022; pp. 21107–21115.

8. Pele, O.; Werman, M. Fast and robust Earth Mover’s Distances. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; IEEE: New York, NY, USA, 2009; pp. 460–467.

9. Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. Adv. Neural Inf. Process. Syst. 2013, 26, 2292–2300.

10. Dessein, A.; Papadakis, N.; Rouas, J.L. Regularized optimal transport and the rot mover’s distance. J. Mach. Learn. Res. 2018, 19, 590–642.

11. Dvurechensky, P.; Gasnikov, A.; Kroshnin, A. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn’s algorithm. In Proceedings of the 36th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; PMLR, pp. 1367–1376.

12. Le, T.; Yamada, M.; Fukumizu, K.; Cuturi, M. Tree-sliced variants of Wasserstein distances. Adv. Neural Inf. Process. Syst. 2019, 32, 12304–12315.

13. Le, T.; Nguyen, T.; Phung, D.; Nguyen, V.A. Sobolev transport: A scalable metric for probability measures with graph metrics. In Proceedings of the 25th International Conference on Artificial Intelligence and Statistics, Online, 28–30 March 2022; PMLR, pp. 9844–9868.

14. Frogner, C.; Zhang, C.; Mobahi, H.; Araya, M.; Poggio, T.A. Learning with a Wasserstein loss. Adv. Neural Inf. Process. Syst. 2015, 28, 2053–2061.

15. Cuturi, M.; Teboul, O.; Vert, J.P. Differentiable ranking and sorting using optimal transport. Adv. Neural Inf. Process. Syst. 2019, 32, 6861–6871.

16. Blondel, M.; Martins, A.F.; Niculae, V. Learning with Fenchel-Young losses. J. Mach. Learn. Res. 2020, 21, 1–69.

17. Birkhoff, G. Tres observaciones sobre el algebra lineal. Univ. Nac. Tucum'an Rev. Ser. A 1946, 5, 147–154.

18. Brualdi, R.A. Combinatorial Matrix Classes; Cambridge University Press: Cambridge, UK, 2006; Volume 13.

19. Alvarez-Melis, D.; Jaakkola, T. Gromov–Wasserstein alignment of word embedding spaces. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 1881–1890.

20. Blondel, M.; Seguy, V.; Rolet, A. Smooth and sparse optimal transport. In Proceedings of the 21st International Conference on Artificial Intelligence and Statistics, Canary Islands, Spain, 9–11 April 2018; PMLR, pp. 880–889.

21. Liu, D.C.; Nocedal, J. On the limited memory BFGS method for large scale optimization. Math. Program. 1989, 45, 503–528. [CrossRef]

22. Amari, S.i.; Ohara, A. Geometry of q-exponential family of probability distributions. Entropy 2011, 13, 1170–1185. [CrossRef]

23. Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. J. Stat. Phys. 1988, 52, 479–487. [CrossRef]

24. Powell, M.J.D. Some global convergence properties of a variable metric algorithm for minimization without exact line searches. In Proceedings of the Nonlinear Programming, SIAM-AMS Proceedings, New York, NY, USA, 1 January 1976; Volume 9.

25. Altschuler, J.; Niles-Weed, J.; Rigollet, P. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. Adv. Neural Inf. Process. Syst. 2017, 30, 1961–1971.

26. Sinkhorn, R.; Knopp, P. Concerning nonnegative matrices and doubly stochastic matrices. Pac. J. Math. 1967, 21, 343–348. [CrossRef]

27. Danskin, J.M. The theory of max-min, with applications. SIAM J. Appl. Math. 1966, 14, 641–664. [CrossRef]

28. Bao, H.; Sugiyama, M. Fenchel-Young losses with skewed entropies for class-posterior probability estimation. In Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, San Diego, CA, USA, 13–15 April 2021; pp. 1648–1656.

29. Naudts, J. Deformed exponentials and logarithms in generalized thermostatistics. Phys. A Stat. Mech. Its Appl. 2002, 316, 323–334. [CrossRef]

30. Suyari, H. The unique non self-referential q-canonical distribution and the physical temperature derived from the maximum entropy principle in Tsallis statistics. Prog. Theor. Phys. Suppl. 2006, 162, 79–86. [CrossRef]

31. Ding, N.; Vishwanathan, S. t-Logistic regression. Adv. Neural Inf. Process. Syst. 2010, 23, 514–522.

32. Futami, F.; Sato, I.; Sugiyama, M. Expectation propagation for t-exponential family using q-algebra. Adv. Neural Inf. Process. Syst. 2017, 30. https://dl.acm.org/doi/10.5555/3294771.3294985.
33. Amid, E.; Warmuth, M.K.; Anil, R.; Koren, T. Robust bi-tempered logistic loss based on bregman divergences. *Adv. Neural Inf. Process. Syst.* 2019, 32, 15013–15022.

34. Martins, A.F.; Figueiredo, M.A.; Aguiar, P.M.; Smith, N.A.; Xing, E.P. Nonextensive entropic kernels. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008; pp. 640–647.

35. Muzellec, B.; Nock, R.; Patrini, G.; Nielsen, F. Tsallis regularized optimal transport and ecological inference. In Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.

36. Byrd, R.H.; Nocedal, J.; Yuan, Y.X. Global convergence of a class of quasi-Newton methods on convex problems. *SIAM J. Numer. Anal.* 1987, 24, 1171–1190. [CrossRef]

37. Schmitzer, B. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM J. Sci. Comput.* 2019, 41, A1443–A1481. [CrossRef]

38. 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.

39. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* 2020, 17, 261–272. doi: 10.1038/s41592-019-0686-2. [CrossRef] [PubMed]

40. Weed, J. An explicit analysis of the entropic penalty in linear programming. In Proceedings of the the 31st Conference on Learning Theory, Stockholm, Sweden, 5–9 July 2018; PMLR, pp. 1841–1855.

41. Golub, G.H.; van Loan, C.F. *Matrix Computations;* The Johns Hopkins University Press: Baltimore, MA, USA, 2013.