ERROR ESTIMATE FOR REGULARIZED OPTIMAL TRANSPORT PROBLEMS VIA BREGMAN DIVERGENCE

KEIICHI MORIKUNI, KOYA SAKAKIBARA, AND ASUKA TAKATSU

Abstract. Regularization by the Shannon entropy enables us to efficiently and approximately solve optimal transport problems on a finite set. This paper is concerned with regularized optimal transport problems via Bregman divergence. We introduce the required properties for Bregman divergences, provide a non-asymptotic error estimate for the regularized problem, and show that the error estimate becomes faster than exponentially.

1. Introduction

An *optimal transport theory* allows for measuring the difference between two probability measures. Innumerable applications of optimal transport theory include mathematics, physics, economics, statistics, computer science, and machine learning. This work focuses on the optimal transport theory on a finite set.

For $K \in \mathbb{N}$, define

$$\mathcal{P}_K := \left\{ z = (z_k) \in \mathbb{R}^K \;\middle|\; z_k \geq 0 \text{ for any } k,\ \sum_k z_k = 1 \right\}.$$

Here and hereafter, k runs over $1, 2, \ldots, K$. Fix $I, J \in \mathbb{N}$. Unless we indicate otherwise, i and j run over $1, 2, \ldots, I$ and $1, 2, \ldots, J$, respectively. For $x \in \mathcal{P}_I$ and $y \in \mathcal{P}_J$, define $x \otimes y \in \mathcal{P}_{I\times J}$ by

$$(x \otimes y)_{ij} := x_i y_j,$$

and set

$$\Pi(x, y) := \left\{ \Pi = (\pi_{ij}) \in \mathcal{P}_{I\times J} \;\middle|\; \sum_{l=1}^{J} \pi_{il} = x_i, \sum_{l=1}^{I} \pi_{lj} = y_j \text{ for any } i, j \right\},$$

where we identify $\mathcal{P}_{I\times J}$ with a subset of $\mathbb{R}^{I\times J}$. An element in $\Pi(x, y)$ is called a *transport plan* between x and y. Note that $\Pi(x, y)$ is a compact set, in particular, a convex polytope, and contains $x \otimes y$. Fix $C = (c^{ij}) \in \mathbb{R}^{I\times J}$ and define a map $\langle C, \cdot\rangle : \mathcal{P}_{I\times J} \to \mathbb{R}$ by

$$\langle C, \Pi\rangle := \sum_{i,j} c^{ij} \pi_{ij}.$$

Consider linear programs of the form

$$\inf_{\Pi\in\Pi(x,y)} \langle C, \Pi\rangle, \tag{1.1}$$

which is a so-called *optimal transport problem*. Since the function $\langle C, \cdot\rangle$ is linear, in particular continuous on a compact set $\Pi(x, y)$, the problem (1.1) always admits a minimizer, but a minimizer is not necessarily unique. A minimizer of the problem (1.1) is called an *optimal transport plan* between x and y.

In the context of the success of the regularized optimal transport problem by the Kullback–Leibler divergence, this paper considers a regularized optimal transport problem via Bregman divergence, which is a generalization of the Kullback–Leibler divergence through a strictly convex function.

Definition 1.1. Let U be a continuous, strictly convex function on $[0, 1]$ with $U \in C^1((0, 1])$. For $z, w \in \mathcal{P}_K$, the *Bregman divergence* associated with U of z with respect to w is given by

$$D_U(z, w) := \sum_k d_U(z_k, w_k),$$

where $d_U : [0, 1] \times (0, 1] \to \mathbb{R}$ is defined for $r \in [0, 1]$ and $r_0 \in (0, 1]$ by

$$d_U(r, r_0) := U(r) - U(r_0) - (r - r_0)U'(r_0)$$

and is naturally extended as a function on $[0,1]\times[0,1]$ valued in $[0,\infty]$ (see Lemma 2.1).

For example, the Bregman divergence associated with $U(r)=r\log r$ reduces to the Kullback–Leibler divergence.

Let us consider a regularized problem of the form

$$\inf_{\Pi\in\Pi(x,y)}\left(\langle C,\Pi\rangle+\varepsilon D_U(\Pi,x\otimes y)\right)\quad\text{for } \varepsilon>0. \tag{1.2}$$

By the continuity and strict convexity of U, $D_U(\cdot,x\otimes y)$ is continuous and strictly convex on a convex polytope $\Pi(x,y)$. Consequently, the problem (1.2) always admits a unique minimizer, denoted by $\Pi^U(C,x,y,\varepsilon)$. Then,

$$\lim_{\varepsilon\downarrow 0}\langle C,\Pi^U(C,x,y,\varepsilon)\rangle=\inf_{\Pi\in\Pi(x,y)}\langle C,\Pi\rangle \tag{1.3}$$

holds (see Subsection 2.4). To give a quantitative error estimate of (1.3), we require the following two assumptions. See Subsections 2.1 and 2.3 to verify that the assumptions are reasonable.

Assumption 1.2. $\Pi(x,y)\neq\operatorname{argmin}_{\Pi\in\Pi(x,y)}\langle C,\Pi\rangle$.

Assumption 1.3. Let $U\in C([0,1])\cap C^1((0,1])\cap C^2((0,1))$ satisfy $U''>0$ on $(0,1)$ and $\lim_{h\downarrow 0}U'(h)=-\infty$. In addition, $r\mapsto rU''(r)$ is non-decreasing in $(0,1)$.

We introduce notions to describe our quantitative error estimate of (1.3).

Definition 1.4. Let U be a continuous, strictly convex function on $[0,1]$ with $U\in C^1((0,1])$. Define $\mathfrak{D}_U(x,y)$ for $x\in\mathcal{P}_I$ and $y\in\mathcal{P}_J$ by

$$\mathfrak{D}_U(x,y):=\sup_{\Pi\in\Pi(x,y)}D_U(\Pi,x\otimes y).$$

Definition 1.5. The *suboptimality gap* of $x\in\mathcal{P}_I$ and $y\in\mathcal{P}_J$ with respect to $C\in\mathbb{R}^{I\times J}$ is defined by

$$\Delta_C(x,y):=\inf_{V'\in V(x,y)\setminus\operatorname{argmin}_{V\in V(x,y)}\langle C,V\rangle}\langle C,V'\rangle-\inf_{V\in V(x,y)}\langle C,V\rangle,$$

where $V(x,y)$ is the set of vertices of $\Pi(x,y)$ and set $\inf\emptyset:=\infty$.

In Subsection 2.1, we verify $\mathfrak{D}_U(x,y),\Delta_C(x,y)\in(0,\infty)$ under Assumption 1.2. We also confirm in Subsection 2.2 that Definition 1.6 below is well-defined.

Definition 1.6. Under Assumption 1.3, we denote by e_U the inverse function of $U':(0,1]\to U'((0,1])$. For $x\in\mathcal{P}_I$ and $y\in\mathcal{P}_J$, let $R_U(x,y)\in[1/2,1)$ satisfy

$$U'(R_U(x,y))-U'(1-R_U(x,y))=\mathfrak{D}_U(x,y),$$

which is uniquely determined. Define $\nu_U(x,y)\in\mathbb{R}$ by

$$\nu_U(x,y):=\sup_{r\in(0,R_U(x,y)]}\left(U'(1-r)+rU''(r)\right).$$

Our main result is as follows.

Theorem 1.7. *Under Assumptions 1.2 and 1.3, the interval*

$$\left(0,\frac{\Delta_C(x,y)R_U(x,y)}{\mathfrak{D}_U(x,y)}\right]\cap\left(0,\frac{\Delta_C(x,y)}{\mathfrak{D}_U(x,y)+\nu_U(x,y)-U'(1)}\right]$$

is well-defined and nonempty. In addition,

$$\langle C,\Pi^U(C,x,y,\varepsilon)\rangle-\inf_{\Pi\in\Pi(x,y)}\langle C,\Pi\rangle\le\Delta_C(x,y)\cdot e_U\left(-\frac{\Delta_C(x,y)}{\varepsilon}+\mathfrak{D}_U(x,y)+\nu_U(x,y)\right)$$

holds for ε in the above interval.

Let us review related results. Computing an exact solution of a large-scale optimal transport problem becomes problematic when, say, $N:=\max\{I,J\}>10^4$. The best-known practical complexity $\widetilde{O}(N^3)$ is attained by an interior point algorithm in [16, Section 5], where $\widetilde{O}$ omits polylogarithmic factors. Though Chen et al. [2, Informal Theorem I.3] improve this complexity to $(N^2)^{1+o(1)}$ and Jambulapati et al. [11, Theorem 2.4] provide an algorithm that finds an ϵ-approximation in $\widetilde{O}(N^2/\epsilon)$, their practical implementations have not been developed.

The tractability of the problem (1.1) is improved by introducing entropic regularization to its objective function, that is,

$$\inf_{\Pi\in\Pi(x,y)} \left(\langle C, \Pi\rangle - \varepsilon S(\Pi)\right),$$

where

$$S(z) := -\sum_k z_k \log z_k, \quad z \in \mathcal{P}_K$$

is the *Shannon entropy.* Here, we put $0\log 0 := 0$ due to the continuity

$$\lim_{r\downarrow 0} r\log r = 0.$$

Fang [7] introduces the Shannon entropy to regularize generic linear programs. By the continuity and the strict convexity of the Shannon entropy, the entropic regularized problem always has a unique minimizer for each value of the regularization parameter. Cominetti and SanMartín [3, Theorem 5.8] prove that the minimizer of the regularized problem converges exponentially to a certain minimizer of the given problem as the regularization parameter goes to zero. Weed [20, Theorem 5] provides a quantitative error estimate of the regularized problem, whose convergence rate is exponential. Note that the entropic regularization allows us to develop approximation algorithms for the problem (1.1). We refer to [17] and references therein. Different types of regularizers are introduced in recent studies. For example, Muzellec et al. [13] use the Tsallis entropy for ecological inference. Dessein et al. [5] and Daniels et al. [4] introduce the Bregman and f-divergences to regularize optimal transport problems, respectively. Apart from entropy and divergence, Klatt et al. [12] use convex functions of Legendre type for regularization.

Regularization by the Shannon entropy is equivalent to that by the Kullback–Libeler divergence. Here, the *Kullback–Leibler divergence* of $z \in \mathcal{P}_K$ with respect to $w \in \mathcal{P}_K$ is given by

$$D_{\mathrm{KL}}(z,w) := \sum_k^K z_k\left(\log z_k - \log w_k\right),$$

where we put $r\log 0 := \infty$ for $r > 0$. Note that the Kullback–Leibler divergence and its dual are the unique members that belong to both the Bregman and f-divergence classes (see [1] for instance). Let us define $U_o \in C([0,\infty)) \cap C^\infty((0,\infty))$ by

$$U_o(r) := \begin{cases} r\log r & \text{for } r \in (0,\infty), \\ 0 & \text{for } r = 0. \end{cases}$$

Then, $D_{U_o} = D_{\mathrm{KL}}$ holds on $\mathcal{P}_K \times \mathcal{P}_K$. There are other strictly convex functions U such that $D_U = D_{\mathrm{KL}}$ holds (see Subsection 4.1).

Our main result Theorem 1.7 with the case $U = U_o$ recovers Weed's work [20, Theorem 5]. Theorem 1.7 with the relation (2.1) guarantees that the regularized optimal value approaches the true optimal value faster than exponentially (see Subsection 2.3). Numerical experiments demonstrate that a Bregman divergence gives smaller errors than the Kullback–Leibler divergence.

This paper is organized as follows. In Section 2, we verify that Assumptions 1.2 and 1.3 are reasonable. Section 3 proves Theorem 1.7. In Section 4, we show that the normalization of U does not affect the error estimate in Theorem 1.7. We then consider the effect of scaling of data and the domain of U on the error estimate. Section 5 provides examples of U satisfying Assumption 1.3. In Section 6, we give numerical experiments and show, in particular, that faster convergence is achieved when regularizations other than the Kullback–Leibler divergence are considered. Finally, in Section 7, we summarize the contents of this paper and give directions for future research.

2. Preliminaries

In this section, we verify that Assumptions 1.2, 1.3 are reasonable and Definition 1.6 is well-defined. We also show that $\mathfrak{D}_U(x,y), \Delta_C(x,y) \in (0,\infty)$ under Assumption 1.2. Throughout, as in the introduction, we fix $I, J \in \mathbb{N}$ and take $C \in \mathbb{R}^{I\times J}$, $x \in \mathcal{P}_I$, and $y \in \mathcal{P}_J$. Let

$$\Omega := \mathbb{R}^{I\times J} \times \mathcal{P}_I \times \mathcal{P}_J \times (0,\infty)$$

and U denote a continuous, strictly convex function on $[0,1]$ with $U \in C^1((0,1])$, unless otherwise stated.

By the strict convexity of U on $[0,1]$,

$$d_U(r, r_0) := U(r) - U(r_0) - (r - r_0)U'(r_0) \geq 0$$

holds for $r \in [0,1]$ and $r_0 \in (0,1]$. In addition, for $r, r_0 \in (0,1]$, $d_U(r, r_0) = 0$ if and only if $r = r_0$. Recall the limiting behavior of U.

Lemma 2.1. *The limit*

$$U'(0) := \lim_{h \downarrow 0} U'(h)$$

exists in $[-\infty, \infty)$ *and* $\lim_{h\downarrow 0} hU'(h) = 0$ *holds.*

Proof. By the strict convexity of U on $[0,1]$, U' is strictly increasing on $(0,1]$ and $\lim_{h\downarrow 0} U'(h) \in [-\infty, \infty)$ holds. Thus, the first assertion follows.

If $U'(0) \in \mathbb{R}$, then $\lim_{h\downarrow 0} hU'(h) = 0$ holds. Assume $U'(0) = -\infty$. The Taylor expansion yields

$$U(r) - U(h) \geq (r - h)U'(h)$$

for all $r, h \in (0,1]$. By the continuity of U, taking the limit as $r \downarrow 0$ gives

$$U(0) - U(h) \geq -hU'(h)$$

for $h \in (0,1]$. If h is small enough, then $U'(h) < 0$ by the monotonicity of U' on $(0,1]$ together with $U'(0) = -\infty$. Thus, we conclude

$$0 = \lim_{h\downarrow 0}(U(h) - U(0)) \leq \liminf_{h\downarrow 0} hU'(h) \leq \limsup_{h\downarrow 0} hU'(h) \leq 0,$$

which leads to $\lim_{h\downarrow 0} hU'(h) = 0$. This completes the proof of the lemma. □

By Lemma 2.1, the limit

$$d_U(r, 0) := \lim_{r_0 \downarrow 0} d_U(r, r_0) \in [0, \infty]$$

exists. In the above relation and throughout, we adhere to the following natural convention:

$$u \pm (-\infty) = \mp\infty, \qquad \lambda \cdot (-\infty) = -\infty, \qquad -\infty \leq -\infty < u < \infty \leq \infty$$

and so on for $u \in \mathbb{R}$ and $\lambda > 0$. Thus, we can regard d_U (resp. D_U) as a function on $[0,1] \times [0,1]$ (resp. $\mathcal{P}_K \times \mathcal{P}_K$) valued in $[0, \infty]$. For $r \in [0,1]$, $d_U(r, 0) = 0$ if and only if $r = 0$. Moreover, $d_U(r, 0) = \infty$ for some $r \in (0,1]$ is equivalent to $U'(0) = -\infty$.

To consider the finiteness of $\mathfrak{D}_U(x, y)$, we define the *support* of $z \in \mathcal{P}_K$ by

$$\mathrm{spt}(z) := \{k \mid z_k > 0\}.$$

Lemma 2.2. *For* $\Pi \in \Pi(x, y)$, $\mathrm{spt}(\Pi) \subset \mathrm{spt}(x) \times \mathrm{spt}(y)$ *holds. Moreover,* $\mathrm{spt}(x) \times \mathrm{spt}(y) = \mathrm{spt}(x \otimes y)$ *follows.*

Proof. For $(i, j) \in \mathrm{spt}(\Pi)$, we have

$$x_i = \sum_{l=1}^{J} \pi_{il} \geq \pi_{ij} > 0, \qquad y_j = \sum_{l=1}^{I} \pi_{lj} \geq \pi_{ij} > 0,$$

which ensure that $i \in \mathrm{spt}(x)$ and $j \in \mathrm{spt}(y)$, that is, $(i, j) \in \mathrm{spt}(x) \times \mathrm{spt}(y)$. For (i, j), it turns out that

$$(i, j) \in \mathrm{spt}(x) \times \mathrm{spt}(y) \iff x_i > 0 \text{ and } y_j > 0 \iff x_i y_j > 0 \iff (i, j) \in \mathrm{spt}(x \otimes y).$$

This completes the proof of the lemma. □

By Lemma 2.2, we find that $D_U(\cdot, x \otimes y)$ is continuous on a compact set $\Pi(x, y)$ so that $\mathfrak{D}_U(x, y) < \infty$.

2.1. **On Assumption 1.2 and Definitions 1.4, 1.5.** There is nothing to prove on the optimal transport problem (1.1) in the case of $\Pi(x, y) = \mathrm{argmin}_{\Pi(x,y)}\langle C, \Pi\rangle$. Thus, we suppose Assumption 1.2, in which $\Pi(x, y)$ contains an element other than $x \otimes y$ and hence $\mathfrak{D}_U(x, y) > 0$ holds. Let $V(x, y)$ be the set of the vertices of $\Pi(x, y)$, that is, $V(x, y)$ is the set with the smallest cardinality among the sets whose convex hull coincides with $\Pi(x, y)$. Note that $\mathrm{argmin}_{V \in V(x,y)}\langle C, V\rangle = V(x, y)$ yields $\mathrm{argmin}_{\Pi \in \Pi(x,y)}\langle C, \Pi\rangle = \Pi(x, y)$. Thus, under Assumption 1.2, $V(x, y) \setminus \mathrm{argmin}_{V \in V(x,y)}\langle C, V\rangle$ is not empty and $\Delta_C(x, y) \in (0, \infty)$ holds.

2.2. **On Definition 1.6.** Let U satisfy Assumption 1.3. By the strict convexity of U on $[0,1]$ together with $U'(0)=-\infty$, the function U' on $(0,1]$ has the inverse function $e_U:(-\infty,U'(1)]\to(0,1]$. We observe from $U''>0$ on $(0,1)$ that the function $r\mapsto U'(r)-U'(1-r)$ is strictly increasing on $(0,1)$. This with the properties

$$U'\left(\frac{1}{2}\right)-U'\left(1-\frac{1}{2}\right)=0, \qquad \lim_{r\uparrow 1}(U'(r)-U'(1-r))=\infty$$

guarantees the unique existence of $R_U(x,y)$. Moreover, since $r\mapsto rU''(r)$ is non-decreasing in $(0,1)$, we find

$$\sup_{r\in(0,R_U(x,y)]}(U'(1-r)+rU''(r))\le U'(1)+R_U(x,y)U''(R_U(x,y))<\infty.$$

Thus, all the notion in Definition 1.6 is well-defined under Assumption 1.3.

2.3. **On Assumption 1.3.** Due to Aleksandrov's theorem (e.g., [6, Theorem 6.9]), U is twice differentiable almost everywhere on $[0,1]$. In the case of $U\in C^2((0,1))$, the strict convexity leads to $U''>0$ almost everywhere on $(0,1)$. Thus, the requirement $U\in C^2((0,1))$ together with $U''>0$ on $(0,1)$ is mild.

Let $U\in C([0,1])\cap C^1((0,1])\cap C^2((0,1))$ such that $U''>0$ on $(0,1)$. To apply some algorithms, such as gradient descent algorithms [17, Sections 4.4, 4.5, 9.3], we require that $\Pi^U(\omega)$ belongs to the interior of the convex polytope $\Pi(x,y)$ for any $\omega=(C,x,y,\varepsilon)\in\Omega$. It follows from [18, Lemma 3.7 and Remark 3.9] that $\Pi^U(\omega)$ belongs to the interior of $\Pi(x,y)$ for any $\omega=(C,x,y,\varepsilon)\in\Omega$ if and only if $U'(0)=-\infty$.

Let $U\in C^2((0,1))$ satisfy $U''>0$ on $(0,1)$. Define $q_U:(0,1)\to[-\infty,\infty]$ by

$$q_U(r):=rU''(r)\cdot\limsup_{h\downarrow 0}\frac{1}{h}\left(\frac{1}{U''(r+h)}-\frac{1}{U''(r)}\right), \qquad Q_U:=\sup_{r\in(0,1)}q_U(r).$$

If $Q_U<\infty$, then $U'(0)=-\infty$ yields $Q_U\ge 1$ by [9, Corollaries 2.6, 2.7]. Note that if $U\in C^3((0,1))$, then

$$q_U(r)=-\frac{rU'''(r)}{U''(r)} \quad \text{for } r\in(0,1).$$

In [9], the notion of q_U is introduced to determine the hierarchy of U in terms of concavity associated with U'. See also [15], where q_U is used to classify convex functions into displacement convex classes. For the definition of the displacement convex classes, see [19, Chapter 17]. It follows from [9, Theorem 2.4] that, for $W\in C^2((0,1))$ satisfying $W''>0$ on $(0,1)$, if $q_U<\infty, q_W>-\infty$ hold almost everywhere on $(0,1)$ and $q_U\le q_W$ holds on $(0,1)$, then there exist $\lambda>0$ and $\mu_1\in\mathbb{R}$ such that $U'\ge\lambda W'+\mu_1$ holds on $(0,1]$, consequently,

$$e_U(\tau)\le e_W(\lambda^{-1}(\tau-\mu_1)) \quad \text{on } \tau\in(U'(0),\lambda W'(1)+\mu_1].$$

Thus, under the assumption $U\in C([0,1])\cap C^1((0,1])\cap C^2((0,1))$ such that $U''>0$ on $(0,1)$ and $U'(0)=-\infty$, if $Q_U<\infty$, then the choice $Q_U=1$ is the best possible for the estimate in Theorem 1.7. Moreover, for $U\in C^2((0,1))$ satisfying $U''>0$ on $(0,1)$, $Q_U=1$ is equivalent to that $r\mapsto rU''(r)$ is non-decreasing on $(0,1)$ by [9, Corollary 2.6]. Thus, we confirm that Assumption 1.3 is reasonable.

Furthermore, if we choose $W=U_o$, then $q_W\equiv 1$ holds on $(0,1)$, and hence $Q_U=1$ implies the existence of $\lambda>0$ and $\mu_1\in\mathbb{R}$ such that

$$e_U(\tau)\le\exp\left(\lambda^{-1}(\tau-\mu_1)\right) \quad \text{for } \tau\in U'((0,1)). \tag{2.1}$$

Thus, the error estimate in Theorem 1.7 is faster than the exponential decay.

2.4. **Asymptotic behavior of the error.** Fix $\omega=(C,x,y,\varepsilon)\in\Omega$. Let $\Pi^*\in\Pi(x,y)$ be an optimal transport plan. We observe from the definition of $\Pi^U(\omega)$ that

$$\langle C,\Pi^U(\omega)\rangle+\varepsilon D_U(\Pi^U(\omega),x\otimes y)=\inf_{\Pi\in\Pi(x,y)}(\langle C,\Pi\rangle+\varepsilon D_U(\Pi,x\otimes y))\le\langle C,\Pi^*\rangle+\varepsilon D_U(\Pi^*,x\otimes y). \tag{2.2}$$

This with the nonnegativity of D_U yields

$$\langle C,\Pi^U(\omega)\rangle-\langle C,\Pi^*\rangle\le\varepsilon D_U(\Pi^*,x\otimes y),$$

proving (1.3). Moreover, the limit

$$\Pi^U(C,x,y,0):=\lim_{\varepsilon\downarrow 0}\Pi^U(C,x,y,\varepsilon)$$

exists and satisfies

$$\operatorname*{argmin}_{\Pi' \in \operatorname{argmin}_{\Pi \in \Pi(x,y)} \langle C, \Pi \rangle} D_U(\Pi', x \otimes y) = \{\Pi^U(C, x, y, 0)\}$$

(see [18, Theorem 3.11] for instance).

3. Proof of Theorem 1.7

Before proving Theorem 1.7, we consider the normalization of U since the correspondence $U \mapsto D_U$ is not injective.

Lemma 3.1. *For $\lambda > 0$ and $\mu_0, \mu_1 \in \mathbb{R}$, define $U_{\lambda,\mu_0,\mu_1} : [0,1] \to \mathbb{R}$ by*

$$U_{\lambda,\mu_0,\mu_1}(r) := \lambda U(r) + \mu_1 r + \mu_0.$$

Then, $d_{U_{\lambda,\mu_0,\mu_1}} = \lambda d_U$ holds on $[0,1] \times [0,1]$, consequently, $D_{U_{\lambda,\mu_0,\mu_1}} = \lambda D_U$ on $\mathcal{P}_K \times \mathcal{P}_K$. If U satisfies Assumption 1.3, then so does U_{λ,μ_0,μ_1} and $q_{U_{\lambda,\mu_0,\mu_1}} = q_U$ holds on $(0,1)$.

Since the proof is straightforward, we omit it. Let U satisfy Assumption 1.3. For $\mu_0, \mu_1 \in \mathbb{R}$ and $\omega = (C, x, y, \varepsilon) \in \Omega$, we have

$$\Pi^{U_{1,\mu_0,\mu_1}}(\omega) = \Pi^U(\omega), \qquad \mathfrak{D}_{U_{1,\mu_0,\mu_1}}(x,y) = \mathfrak{D}_U(x,y),$$
$$R_{U_{1,\mu_0,\mu_1}}(x,y) = R_U(x,y), \qquad \nu_{U_{1,\mu_0,\mu_1}}(x,y) = \nu_U(x,y) + \mu_1,$$

and $e_{U_{1,\mu_0,\mu_1}}(\tau) = e_U(\tau - \mu_1)$ for $\tau \in U'_{1,\mu_0,\mu_1}((0,1])$ together with $U'_{1,\mu_0,\mu_1}(1) = U'(1) + \mu_1$. Thus, in Theorem 1.7, we can normalize U to $U(0) = U(1) = 0$ as well as the case of $U = U_o$ without loss of generality. For the reason and the effect of choice of $\lambda > 0$, see Section 4.1.

Throughout the rest of this section, we suppose Assumptions 1.2 and 1.3 together with $U(0) = U(1) = 0$. We prepare two lemmas to prove Theorem 1.7. The following proof strategy is aligned with the argument of [20, Lemmas 6–8].

Lemma 3.2. *For $r, s, t \in [0,1]$ and $r_0 \in (0,1]$,*

$$U((1-t)r + ts) \geq (1-t)U(r) + tU(s) + rU(1-t) + sU(t),$$
$$d_U((1-t)r + ts, r_0) \geq (1-t)d_U(r, r_0) + t d_U(s, r_0) + rU(1-t) + sU(t).$$

Proof. For $r, s, t \in [0,1]$ and $r_0 \in (0,1]$, if the first inequality holds true, then it holds that

$$\begin{aligned} d_U((1-t)r + ts, r_0) &= U((1-t)r + ts) - U(r_0) - \{(1-t)r + ts - r_0\} U'(r_0) \\ &\geq (1-t)U(r) + tU(s) + rU(1-t) + sU(t) - U(r_0) - \{(1-t)r + ts - r_0\} U'(r_0) \\ &= (1-t)d_U(r, r_0) + t d_U(s, r_0) + rU(1-t) + sU(t), \end{aligned}$$

which is the second inequality.

To show the first inequality, set

$$G(r,s,t) := (1-t)U(r) + tU(s) + rU(1-t) + sU(t) - U((1-t)r + ts)$$

for $r, s, t \in [0,1]$. By the continuity of G, it is enough to show

$$\max_{r \in [0,1]} G(r,s,t) \leq 0 \quad \text{for } s, t \in (0,1). \tag{3.1}$$

Note that

$$\begin{aligned} \frac{\partial}{\partial r} G(r,s,t) &= U(1-t) + (1-t)\left(U'(r) - U'((1-t)r + ts)\right), \\ \frac{\partial^2}{\partial r^2} G(r,s,t) &= (1-t)\left[U''(r) - (1-t)U''((1-t)r + st))\right] \end{aligned}$$

for $r, s, t \in (0,1)$.

Let us now show

$$\max_{r \in [0,1]} G(r,s,t) = \max\{G(0,s,t), G(1,s,t)\} \quad \text{for } s, t \in (0,1). \tag{3.2}$$

Let $s,t \in (0,1)$. On one hand, for $r \in (0,1)$ with $r \le s$, since U' is strictly increasing on $(0,1)$, we have $U'(r) - U'((1-t)r + ts) \le 0$ and hence $\partial_r G(r,s,t) \le U(1-t)$. Note that $U(1-t) < 0$ follows from the strict convexity of U with the condition $U(0) = U(1) = 0$. This implies that $\partial_r G(r,s,t) < 0$ if $r \le s$ and

$$\max_{r\in[0,s]} G(r,s,t) = G(0,s,t), \quad \text{in particular,} \quad G(s,s,t) < G(0,s,t). \tag{3.3}$$

On the other hand, for $r \in (0,1)$ with $r > s$, we have $(1-t)r + ts < r$ and

$$(1-t)rU''((1-t)r+ts) < \ [(1-t)r+ts]U''((1-t)r+ts) \le rU''(r)$$

by $U'' > 0$ and the monotonicity of $r \mapsto rU''(r)$ on $(0,1)$. This yields $\partial_r^2 G(r,s,t) > 0$. If $\partial_r G(r_0,s,t) = 0$ holds for some $r_0 \in [s,1]$, then

$$\max_{r\in[s,1]} G(r,s,t) = \max\{G(s,s,t), G(1,s,t)\}. \tag{3.4}$$

In contrast, if $\partial_r G(r,s,t) < 0$ always holds, then

$$\max_{r\in[s,1]} G(r,s,t) = G(s,s,t). \tag{3.5}$$

Summarizing the above relations (3.3), (3.4), and (3.5), we have (3.2).

Since $U(0) = U(1) = 0$, a direct computation gives

$$\frac{\partial^2}{\partial s^2} G(0,s,t) = \frac{\partial^2}{\partial s^2}\left(tU(s) + sU(t) - U(ts)\right) = tU''(s) - t^2U''(ts) = \frac{t}{s}\left(sU''(s) - tsU''(ts)\right) \ge 0$$

for $s,t \in (0,1)$, where the inequality follows from the monotonicity of $r \mapsto rU''(r)$ on $(0,1)$. Thus, for $t \in (0,1)$, $G(0,\cdot,t)$ is convex on $[0,1]$ and

$$\max_{s\in[0,1]} G(0,s,t) = \max\{G(0,0,t), G(0,1,t)\} = 0. \tag{3.6}$$

Next, we find

$$\frac{\partial}{\partial s} G(1,s,t) = \frac{\partial}{\partial s}\left(tU(s) + U(1-t) + sU(t) - U(1-t+ts)\right) = tU'(s) + U(t) - tU'\left(1-t+ts\right) < U(t) < 0$$

for $s,t \in (0,1)$, where the first inequality follows from the monotonicity of U' on $(0,1)$. This leads to

$$\max_{s\in[0,1]} G(1,s,t) = G(1,0,t) = 0 \tag{3.7}$$

for $t \in (0,1)$. Thus, we deduce (3.1) from (3.2) together with (3.6) and (3.7). This proves the lemma. $\square$

Recall that for any $D > 0$, there exists $R \in (1/2,1)$ uniquely such that $U'(R) - U'(1-R) = D$ (see Subsection 2.2).

Lemma 3.3. *For $D > 0$ and $R \in (1/2,1)$ with $U'(R) - U'(1-R) = D$,*

$$r \mapsto Dr - U(r) - U(1-r)$$

is strictly increasing on $[0,R]$. Moreover, it holds that

$$-U(r) - U(1-r) \le -rU'(r) + r \sup_{\rho\in(0,R]} \left(U'(1-\rho) + \rho U''(\rho)\right)$$

for $r \in (0,R]$.

Proof. We calculate

$$\frac{\mathrm{d}^2}{\mathrm{d}r^2}(Dr - U(r) - U(1-r)) = -U''(r) - U''(1-r) < 0$$

for $r \in (0,1)$, consequently,

$$\frac{\mathrm{d}}{\mathrm{d}r}(Dr - U(r) - U(1-r)) > \frac{\mathrm{d}}{\mathrm{d}r}(Dr - U(r) - U(1-r))\bigg|_{r=R} = D - U'(R) + U'(1-R) = 0$$

for $r \in (0,R)$. This proves the first assertion.

We also find

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}r}\left(U(r)+U(1-r)-rU'(r)+r\sup_{\rho\in(0,R]}(U'(1-\rho)+\rho U''(\rho))\right)\\
&=-U'(1-r)-rU''(r)+\sup_{\rho\in(0,R]}(U'(1-\rho)+\rho U''(\rho))\\
&\geq 0
\end{aligned}$$

for $r\in(0,R]$. This together with Lemma 2.1 and the assumption $U(0)=U(1)=0$ yield

$$\begin{aligned}
&U(r)+U(1-r)-rU'(r)+r\sup_{\rho\in(0,R]}(U'(1-\rho)\\
&\geq \lim_{r\downarrow 0}\left(U(r)+U(1-r)-rU'(r)+r\sup_{\rho\in(0,R]}(U'(1-\rho)+\rho U''(\rho))\right)=0
\end{aligned}$$

for $r\in(0,R]$. This proves the second assertion of the lemma. $\square$

Proof of Theorem 1.7. Let $\omega=(C,x,y,\varepsilon)\in\Omega$. Then, $\Delta_C(x,y),\mathfrak{D}_U(x,y)\in(0,\infty)$ hold as mentioned in Subsection 2.1. We also have

$$U'(1)\leq U'(1)+\lim_{r\downarrow 0}rU''(r)\leq \nu_U(x,y)<\infty$$

and $\mathfrak{D}_U(x,y)+\nu_U(x,y)-U'(1)\in(0,\infty)$. Thus, the interval

$$\left(0,\frac{\Delta_C(x,y)R_U(x,y)}{\mathfrak{D}_U(x,y)}\right]\cap\left(0,\frac{\Delta_C(x,y)}{\mathfrak{D}_U(x,y)+\nu_U(x,y)-U'(1)}\right]$$

is well-defined and nonempty. Let us choose ε from the interval. Note that

$$\varepsilon\in\left(0,\frac{\Delta_C(x,y)}{\mathfrak{D}_U(x,y)+\nu_U(x,y)-U'(1)}\right]$$

is equivalent to

$$-\frac{\Delta_C(x,y)}{\varepsilon}+\mathfrak{D}_U(x,y)+\nu_U(x,y)\in U'((0,1]).$$

Recall that $V(x,y)$ is the vertex set of $\Pi(x,y)$. There exists a family $\{t_V\}_{V\in V(x,y)}\subset[0,1]$ uniquely such that

$$\sum_{V\in V(x,y)}t_V=1,\qquad \Pi^U(\omega)=\sum_{V\in V(x,y)}t_VV.$$

Set

$$V_0(x,y):=\operatorname*{argmin}_{V\in V(x,y)}\langle C,V\rangle,\quad t:=1-\sum_{V\in V_0(x,y)}t_V.$$

By Assumption 1.2, $V_0(x,y)\neq V(x,y)$ holds. Since $\Pi^U(\omega)$ belongs to the interior of the convex polytope $\Pi(x,y)$, we find that $t_V\in(0,1)$ for $V\in V(x,y)$, consequently, $t\in(0,1)$. We also set

$$\Pi^*:=\sum_{V\in V_0(x,y)}\frac{t_V}{1-t}V,\quad \Pi':=\sum_{V'\in V(x,y)\setminus V_0(x,y)}\frac{t_{V'}}{t}V'.$$

It turns out that $\Pi^*,\Pi'\in\Pi(x,y)$ and

$$\Pi^U(\omega)=(1-t)\Pi^*+t\Pi'.$$

We find that

$$\langle C,V\rangle=\inf_{\Pi\in\Pi(x,y)}\langle C,\Pi\rangle\quad\text{for } V\in V_0(x,y),\qquad\text{in particular}\quad \langle C,\Pi^*\rangle=\inf_{\Pi\in\Pi(x,y)}\langle C,\Pi\rangle.$$

Setting

$$r:=\frac{\langle C,\Pi^U(\omega)\rangle-\langle C,\Pi^*\rangle}{\Delta_C(x,y)},$$

we observe from (2.2) with the definition of $\mathfrak{D}_U(x,y)$ that

$$\langle C,\Pi^U(\omega)\rangle-\langle C,\Pi^*\rangle\leq\varepsilon\left(D_U(\Pi^*,x\otimes y)-D_U(\Pi^U(\omega),x\otimes y)\right)\leq\varepsilon\mathfrak{D}_U(x,y).\tag{3.8}$$

We also find

$$\langle C, \Pi^U(\omega)\rangle - \langle C, \Pi^*\rangle = t\langle C, \Pi' - \Pi^*\rangle \geq t\left(\inf_{V'\in V(x,y)\setminus V_0(x,y)}\langle C, V'\rangle - \inf_{V\in V_0(x,y)}\langle C, V\rangle\right) = t\Delta_C(x,y).$$

These with the condition $\varepsilon \in (0, \Delta_C(x,y)R_U(x,y)/\mathfrak{D}_U(x,y)]$ yield

$$t \leq r = \frac{\langle C, \Pi^U(\omega)\rangle - \langle C, \Pi^*\rangle}{\Delta_C(x,y)} \leq \frac{\varepsilon\mathfrak{D}_U(x,y)}{\Delta_C(x,y)} \leq R_U(x,y).$$

It follows from Lemma 2.2 with the second inequality in Lemma 3.2 that

$$\begin{aligned}
D_U(\Pi^U(\omega), x\otimes y) &= D_U((1-t)\Pi^* + t\Pi', x\otimes y)\\
&= \sum_{(i,j)\in \mathrm{spt}(x\otimes y)} d_U((1-t)\pi^*_{ij} + t\pi'_{ij}, x_i y_j)\\
&\geq \sum_{(i,j)\in \mathrm{spt}(x\otimes y)} \left\{(1-t)d_U(\pi^*_{ij}, x_i y_j) + t d_U(\pi'_{ij}, x_i y_j) + \pi^*_{ij}U(1-t) + \pi'_{ij}U(t)\right\}\\
&= (1-t)D_U(\Pi^*, x\otimes y) + tD_U(\Pi', x\otimes y) + U(1-t) + U(t).
\end{aligned}$$

This and Lemma 3.3 together with $t \leq r \leq R_U(x,y)$ yield

$$\begin{aligned}
D_U(\Pi^*, x\otimes y) - D_U(\Pi^U(\omega), x\otimes y) &\leq tD_U(\Pi^*, x\otimes y) - tD_U(\Pi', x\otimes y) - U(t) - U(1-t)\\
&\leq t\mathfrak{D}_U(x,y) - U(t) - U(1-t)\\
&\leq r\mathfrak{D}_U(x,y) - U(r) - U(1-r)\\
&\leq r\mathfrak{D}_U(x,y) - rU'(r) + r\nu_U(x,y)\\
&= r(\mathfrak{D}_U(x,y) - U'(r) + \nu_U(x,y)).
\end{aligned}$$

Combining this with (3.8), we find

$$\frac{\langle C, \Pi^U(\omega)\rangle - \langle C, \Pi^*\rangle}{\varepsilon} \leq D_U(\Pi^*, x\otimes y) - D_U(\Pi^U(\omega), x\otimes y) \leq r(\mathfrak{D}_U(x,y) - U'(r) + \nu_U(x,y)),$$

which leads to

$$U'(r) \leq -\frac{\langle C, \Pi^U(\omega)\rangle - \langle C, \Pi^*\rangle}{\varepsilon r} + \mathfrak{D}_U(x,y) + \nu_U(x,y) = -\frac{\Delta_C(x,y)}{\varepsilon} + \mathfrak{D}_U(x,y) + \nu_U(x,y).$$

It follows from the monotonicity of U' on $(0,1]$ that

$$\frac{\langle C, \Pi^U(\omega)\rangle - \langle C, \Pi^*\rangle}{\Delta_C(x,y)} = r = e_U(U'(r)) \leq e_U\left(-\frac{\Delta_C(x,y)}{\varepsilon} + \mathfrak{D}_U(x,y) + \nu_U(x,y)\right),$$

that is,

$$\langle C, \Pi^U(\omega)\rangle - \inf_{\Pi\in\Pi(x,y)}\langle C, \Pi\rangle \leq \Delta_C(x,y)\cdot e_U\left(-\frac{\Delta_C(x,y)}{\varepsilon} + \mathfrak{D}_U(x,y) + \nu_U(x,y)\right)$$

as desired. □

4. Normalization and scaling

In this section, we first show that the normalization of a strictly convex function does not affect the error estimate in Theorem 1.7. We then consider the effect of scaling of data and the domain of a strictly convex function on the error estimate.

4.1. **Normalization.** Let $U \in C([0,1]) \cap C^1((0,1]) \cap C^2((0,1))$ satisfy $U'' > 0$ on $(0,1)$. For $\mu_0, \mu_1 \in \mathbb{R}$ and $\lambda > 0$, define $U_{\lambda,\mu_0,\mu_1} \in C([0,1]) \cap C^1((0,1]) \cap C^2((0,1))$ by

$$U_{\lambda,\mu_0,\mu_1}(r) := \lambda U(r) + \mu_1 r + \mu_0.$$

By the *normalization* of U, we mean the choice of $\mu_0, \mu_1 \in \mathbb{R}$ and $\lambda > 0$ such that

$$U_{\lambda,\mu_0,\mu_1}(0) = u_0, \quad U_{\lambda,\mu_0,\mu_1}(1) = u_1, \quad U'_{\lambda,\mu_0,\mu_1}(1) = u'_1 \quad \text{for } u_0, u_1, u'_1 \in \mathbb{R} \text{ with } u_1 - u_0 < u'_1,$$

where the inequality on u_0, u_1, u_1' is required for U_{λ,μ_0,μ_1} to be strictly convex. Let $(C, x, y, \varepsilon) \in \Omega$. Then, we find $\Pi^{U_{\lambda,\mu_0,\mu_1}}(C, x, y, \varepsilon) = \Pi^U(C, x, y, \lambda\varepsilon)$ and that the interval

$$\left(0, \frac{\Delta_C(x,y) R_{U_{\lambda,\mu_0,\mu_1}}(x,y)}{\mathfrak{D}_{U_{\lambda,\mu_0,\mu_1}}(x,y)}\right] \cap \left(0, \frac{\Delta_C(x,y)}{\mathfrak{D}_{U_{\lambda,\mu_0,\mu_1}}(x,y) + \nu_{U_{\lambda,\mu_0,\mu_1}}(x,y) - U'_{\lambda,\mu_0,\mu_1}(1)}\right]$$

is well-defined (resp. contains ε) if and only if the interval

$$\left(0, \frac{\Delta_C(x,y) R_U(x,y)}{\mathfrak{D}_U(x,y)}\right] \cap \left(0, \frac{\Delta_C(x,y)}{\mathfrak{D}_U(x,y) + \nu_U(x,y) - U'(1)}\right]$$

is well-defined (resp. contains $\lambda\varepsilon$), in which the equality

$$e_{U_{\lambda,\mu_0,\mu_1}}\left(-\frac{\Delta_C(x,y)}{\varepsilon} + \mathfrak{D}_{U_{\lambda,\mu_0,\mu_1}}(x,y) + \nu_{U_{\lambda,\mu_0,\mu_1}}(x,y)\right) = e_U\left(-\frac{\Delta_C(x,y)}{\lambda\varepsilon} + \mathfrak{D}_U(x,y) + \nu_U(x,y)\right)$$

holds. Thus, in Theorem 1.7, we can normalize U as

$$U(0) = u_0, \quad U(1) = u_1, \quad U'(1) = u_1' \quad \text{for } u_0, u_1, u_1' \in \mathbb{R} \text{ with } u_1 - u_0 < u_1'$$

without loss of generality.

Let $W \in C([0,1]) \cap C^1((0,1]) \cap C^2((0,1))$ also satisfy $W'' > 0$ on $(0,1)$. Then, the following three conditions are equivalent to each other.

(C0) There exist $\mu_0, \mu_1 \in \mathbb{R}$ and $\lambda > 0$ such that $W = U_{\lambda,\mu_0,\mu_1}$ on $(0,1)$.
(C1) There exist $\mu_1 \in \mathbb{R}$ and $\lambda > 0$ such that $W' = \lambda U' + \mu_1$ on $(0,1)$.
(C2) There exists $\lambda > 0$ such that $W'' = \lambda U''$ on $(0,1)$.

Thus, under the normalization $U(0) = u_0$ and $U(1) = u_1$ for $u_0, u_1 \in \mathbb{R}$, we can use either U' or U'' instead of U itself. Note that each of (C0)–(C2) is equivalent to the following condition.

(D) There exists $\lambda > 0$ such that $D_W = \lambda D_U$ on $\mathcal{P}_K \times \mathcal{P}_K$ for $K \geq 3$.

The implication from (C0) to (D) is straightforward. Assume (D). For $K \geq 3$ and $r \in [0,1]$, define $z^r \in \mathcal{P}_K$ by

$$z_k^r := \begin{cases} r & \text{for } k = 1, \\ 1 - r & \text{for } k = 2, \\ 0 & \text{otherwise.} \end{cases}$$

We also define $z^* \in \mathcal{P}_K$ by $z_k^* = K^{-1}$ for all k. For $r \in (0,1)$, we calculate

$$W'(r) - W'(1-r) = \frac{\mathrm{d}}{\mathrm{d}r} D_W(z^r, z^*) = \lambda \frac{\mathrm{d}}{\mathrm{d}r} D_U(z^r, z^*) = \lambda\left(U'(r) - U'(1-r)\right).$$

Differentiating this with respect to r implies that

$$W''(r) - \lambda U''(r) = -\left(W''(1-r) - \lambda U''(1-r)\right), \quad \text{in particular} \quad W''(1/2) = \lambda U''(1/2). \tag{4.1}$$

For $r, s \in (0,1)$ with $r + s < 1$, we define $z^{r,s} \in \mathcal{P}_K$ by

$$z_k^{r,s} := \begin{cases} 1 - (r+s) & \text{for } k = 1, \\ r & \text{for } k = 2, \\ (K-2)^{-1} s & \text{otherwise.} \end{cases}$$

It turns out that

$$\begin{aligned}(r+s)W''(1-(r+s)) + rW''(r) &= \frac{\partial}{\partial r} D_W(z^1, z^{r,s}) \\ &= \lambda \frac{\partial}{\partial r} D_U(z^1, z^{r,s}) = \lambda\left[(r+s)U''(1-(r+s)) + rU''(r)\right],\end{aligned}$$

implying

$$W''(r) - \lambda U''(r) = -\frac{r+s}{r}\left(W''(1-(r+s)) - \lambda U''(1-(r+s))\right).$$

This with (4.1) provides

$$\frac{r+s}{r}\left(W''(1-(r+s)) - \lambda U''(1-(r+s))\right) = W''(1-r) - \lambda U''(1-r),$$

in particular, the choice of $r = 1/2$ leads to

$$W''\left(\frac{1}{2} - s\right) = \lambda U''\left(\frac{1}{2} - s\right) \quad \text{for } s \in \left(0, \frac{1}{2}\right).$$

This together with (4.1) gives $W'' = \lambda U''$ on $(0,1)$, which is nothing but (C2). Note that, under Assumption 1.2, we have $I, J \neq 1$ hence $IJ \geq 4$. Thus, the condition $K \geq 3$ in (D) is reasonable.

We also notice that (C2) leads to the following condition.

(C) $q_U = q_W$ on $(0,1)$.

By [9, Theorem 2.4], if q_U, q_W are finite almost everywhere on $(0,1)$, then (C) leads to (C2). Thus, all conditions (C0)–(C2), (D), and (C) are equivalent to each other.

To use U' instead of U, let us consider the following assumption.

Assumption 4.1. Let $L \in C((0,1]) \cap C^1((0,1))$ satisfy that $L' > 0$ on $(0,1)$, $\lim_{t\downarrow 0} L(t) = -\infty$, and $t \mapsto tL'(t)$ is non-decreasing on $(0,1)$.

Suppose Assumption 4.1. Then, there exists $t_0 \in (0,1]$ such that $L < 0$ on $(0,t_0]$. By the monotonicity of $t \mapsto tL'(t)$, we have

$$L(t_0) - L(t) = \int_t^{t_0} L'(s)\mathrm{d}s \leq t_0 L'(t_0) \int_t^{t_0} \frac{1}{s}\mathrm{d}s = t_0 L'(t_0)\left(\log t_0 - \log t\right)$$

for $t \in (0, t_0]$. Since L is monotone on $(0,t_0]$ and

$$\left(L(t_0) - t_0 L'(t_0)\log t_0\right)(t_0 - h) + t_0 L'(t_0) \int_h^{t_0} \log t \mathrm{d}t \leq \int_h^{t_0} L(t)\mathrm{d}t < 0$$

holds $h \in (0, t_0]$, the improper integral

$$U_L(r) := \int_0^r L(t)\mathrm{d}t$$

is well-defined for $r \in [0,1]$. It is easy to see that U_L satisfies Assumption 1.3. Note that

$$d_{U_L}(r, r_0) = \int_{r_0}^r (L(t) - L(r_0))\mathrm{d}t = \int_{r_0}^r \int_{r_0}^t L'(s)\mathrm{d}s\mathrm{d}t \quad \text{for } r \in [0,1], r_0 \in (0,1].$$

Conversely, if U satisfies Assumption 1.3, then $L = U'$ satisfies Assumption 4.1. Thus, we can use L satisfying Assumption 4.1 instead of U satisfying Assumption 1.3 for our regularization.

Remark 4.2. In Theorem 1.7, the range of the regularization parameter ε is given by the intersection of the two intervals. One interval

$$\left(0, \frac{\Delta_C(x,y)}{\mathfrak{D}_U(x,y) + \nu_U(x,y) - U'(1)}\right]$$

is needed to make

$$-\frac{\Delta_C(x,y)}{\varepsilon} + \mathfrak{D}_U(x,y) + \nu_U(x,y) \in U'((0,1])$$

as seen in the proof of Theorem 1.7. Hence, this interval is not needed if U is extended to a continuous, strictly convex function on $[0,\infty)$ and $U \in C^1((0,\infty))$ with $\lim_{r\uparrow\infty} U'(r) = \infty$. For example, under the normalization

$$U(0) = U(1) = 0, \qquad U'(1) = 1, \tag{4.2}$$

which is valid for the case $U = U_{\mathrm{o}}$, we can extend U by setting

$$U(r) := r\log r \quad \text{for } r > 1.$$

For L satisfying Assumption 4.1, if we set

$$\ell(t) := \left(L(1) - \int_0^1 L(s)\mathrm{d}s\right)^{-1}\left(L(t) - \int_0^1 L(s)\mathrm{d}s\right)$$

for $t \in [0,1]$, then ℓ satisfies Assumption 4.1 and U_ℓ satisfies the normalization (4.2).

4.2. **Scaling.** In the regularized problem (1.2), scaling can have two meanings: scaling of data and scaling of the domain of a strictly convex function. Let us show that they play an equivalent role.

In the optimal transport problem (1.1), although two given data x and y are normalized to be 1 with respect to the ℓ^1-norm, their ℓ^1-norms can be chosen arbitrarily if both are the same. For a subset $\mathcal{Z}$ of Euclidean space and $a > 0$, set

$$a\mathcal{Z} := \{az \mid z \in \mathcal{Z}\}.$$

For $x \in \mathcal{P}_I$ and $y \in \mathcal{P}_J$, we shall, by abuse of notation, define

$$\Pi(ax, ay) := \left\{ \widetilde{\Pi} = (\widetilde{\pi}_{ij}) \in a\mathcal{P}_{I\times J} \;\middle|\; \sum_{l=1}^{J} \widetilde{\pi}_{il} = ax_i \text{ and } \sum_{l=1}^{I} \widetilde{\pi}_{lj} = ay_j \text{ for any } i, j \right\}.$$

Then, we have $\Pi(ax, ay) = a\Pi(x, y)$ and hence $ax \otimes y \in \Pi(ax, ay)$. We denote by $V(ax, ay)$ the set of the vertices of $\Pi(ax, ay)$. Then, $V(ax, ay) = aV(x, y)$ follows. For $U \in C([0, 1]) \cap C^1((0, 1])$ being strictly convex on $[0, 1]$ and $a \in (0, 1]$, we can define $D_U : a\mathcal{P}_K \times a\mathcal{P}_K \to [0, \infty]$ by

$$D_U(az, aw) := \sum_k d_U(az_k, aw_k) \quad \text{for } z, w \in \mathcal{P}_K$$

and consider the regularized problem

$$\inf_{\widetilde{\Pi}\in\Pi(ax,ay)} \left(\langle C, \widetilde{\Pi}\rangle + \varepsilon D_U(\widetilde{\Pi}, ax \otimes y) \right) \quad \text{for } \omega = (C, x, y, \varepsilon) \in \Omega. \tag{4.3}$$

More generally, for $a, b > 0$ with $a \le b$ and $W \in C([0, b]) \cap C^1((0, b])$ being strictly convex on $[0, b]$, we define $d_W : [0, b] \times [0, b] \to [0, \infty]$ by

$$d_W(r, r_0) := W(r) - W(r_0) - (r - r_0)W'(r_0) \quad \text{for } r \in [0, b], r_0 \in (0, b]$$

and

$$d_W(r, 0) := \lim_{h\downarrow 0} d_W(r, h) \quad \text{for } r \in [0, b].$$

We also define $D_W : a\mathcal{P}_K \times a\mathcal{P}_K \to [0, \infty]$ by

$$D_W(az, aw) := \sum_k d_W(az_k, aw_k) \quad \text{for } z, w \in \mathcal{P}_K.$$

This enables us to consider the regularized problem on $a\mathcal{P}_I \times a\mathcal{P}_J$ as similar as (4.3) by using a strictly convex function $W \in C([0, b]) \cap C^1((0, b])$ with $a \le b$.

Next, let us scale the domain of $U \in C([0, 1]) \cap C^1((0, 1])$ being strictly convex on $[0, 1]$ by setting

$$U^b(r) := bU(b^{-1}r) : [0, b] \to \mathbb{R}$$

for $b > 0$. Following the notation in (4.4) below, U^b coincides with U_1^b. If U satisfies Assumption 1.3, then so does U^b if $b > 1$. We observe from

$$d_{U^b}(r, r_0) = b \cdot d_U(b^{-1}r, b^{-1}r_0) \quad \text{for } r, r_0 \in [0, b]$$

that

$$\begin{aligned}
\inf_{\Pi\in\Pi(x,y)} \left(\langle C, \Pi\rangle + \varepsilon D_{U^b}(\Pi, x \otimes y)\right) &= b \cdot \inf_{\Pi\in\Pi(x,y)} \left(\langle C, b^{-1}\Pi\rangle + \varepsilon D_U(b^{-1}\Pi, b^{-1}x \otimes y) \right), \\
\operatorname*{argmin}_{\Pi\in\Pi(x,y)} \left(\langle C, \Pi\rangle + \varepsilon D_{U^b}(\Pi, x \otimes y)\right) &= \operatorname*{argmin}_{\Pi\in\Pi(x,y)} \left(\langle C, b^{-1}\Pi\rangle + \varepsilon D_U(b^{-1}\Pi, b^{-1}x \otimes y) \right) \\
&= b \cdot \operatorname*{argmin}_{\widetilde{\Pi}\in\Pi(b^{-1}x,b^{-1}y)} \left(\langle C, \widetilde{\Pi}\rangle + \varepsilon D_U(\widetilde{\Pi}, b^{-1}x \otimes y) \right),
\end{aligned}$$

for $\omega = (C, x, y, \varepsilon) \in \Omega$. This means that the two scalings play an equivalent role.

More generally, for $b > 0$, let $W \in C([0, b]) \cap C^1((0, b]) \cap C^2((0, b))$ satisfy that $W'' > 0$ on $(0, b)$, $\lim_{h\downarrow 0} W'(h) = -\infty$, and $r \mapsto rW''(r)$ is non-decreasing on $(0, b)$. For $a > 0$, we define the function $W_b^a \in C([0, a]) \cap C^1((0, a]) \cap C^2((0, a))$ by

$$W_b^a(r) := ab^{-1}W(a^{-1}br). \tag{4.4}$$

Then, we see that ${W_b^a}'' > 0$ on $(0, a)$, $\lim_{h\downarrow 0} {W_b^a}'(h) = -\infty$ hold and $r \mapsto r{W_b^a}''(r)$ is non-decreasing on $(0, a)$. The normalization $W(0) = W(b) = 0$ with $W'(b) = 1$ is equivalent to the normalization $W_b^a(0) = W_b^a(a) = 0$

with ${W_b^a}'(a) = 1$. In particular, $U := W_b^1$ satisfies Assumption 1.3. Furthermore, for $\omega = (C, x, y, \varepsilon) \in \Omega$, we have

$$\operatorname*{argmin}_{\widetilde{\Pi} \in \Pi(ax, ay)} \left(\langle C, \widetilde{\Pi} \rangle + \varepsilon D_{W_b^a}(\widetilde{\Pi}, ax \otimes y) \right) = a \cdot \operatorname*{argmin}_{\Pi \in \Pi(x,y)} \left(\langle C, \Pi \rangle + \varepsilon D_U(\Pi, x \otimes y) \right).$$

This means that the left-hand side is a singleton. We denote by $\Pi^{W_b^a}(C, ax, ay, \varepsilon)$ the unique element. Then, $\Pi^{W_b^a}(C, ax, ay, \varepsilon) = a\Pi^U(\omega)$ holds. Let us define all notions provided to state Theorem 1.7 as follows:

$$\begin{aligned}
\mathfrak{D}_{W_b^a}(ax, ay) &:= \sup_{\widetilde{\Pi} \in \Pi(ax, ay)} D_{W_b^a}(\widetilde{\Pi}, ax \otimes y), \\
\Delta_C(ax, ay) &:= \inf_{\widetilde{V}' \in V(ax,ay) \setminus \operatorname{argmin}_{\widetilde{V} \in V(ax,ay)} \langle C, \widetilde{V}' \rangle} \langle C, \widetilde{V}' \rangle - \inf_{\widetilde{V} \in V(ax,ay)} \langle C, \widetilde{V} \rangle, \\
R_{W_b^a}(ax, ay) &\in [a/2, a] \quad \text{such that} \quad {W_b^a}'(R_{W_b^a}(ax, ay)) - {W_b^a}'(a - R_{W_b^a}(ax, ay)) = a^{-1} \mathfrak{D}_{W_b^a}(ax, ay), \\
\nu_{W_b^a}(ax, ay) &:= \sup_{r \in (0, R_{W_b^a}(ax, ay)]} \left({W_b^a}'(a - r) + r {W_b^a}''(r) \right).
\end{aligned} \tag{4.5}$$

We also denote by $e_{W_b^a} : {W_b^a}'((0, a]) \to (0, a]$ the inverse function of ${W_b^a}' : (0, a] \to {W_b^a}'((0, a])$. It follows from

$$W_b^a(ar) = aU(r), \quad {W_b^a}'(ar) = U'(r), \quad {W_b^a}''(ar) = a^{-1}U''(r), \quad d_{W_b^a}(ar, ar_0) = a d_U(r, r_0)$$

for $r, r_0 \in [0, 1]$ that

$$\mathfrak{D}_{W_b^a}(ax, ay) = a\mathfrak{D}_U(x, y), \quad R_{W_b^a}(ax, ay) = aR_U(x, y), \quad \nu_{W_b^a}(ax, ay) = \nu_U(x, y),$$

and $e_{W_b^a} = ae_U$ on ${W_b^a}'((0, a]) = U'((0, 1])$. We also find $\Delta_C(ax, ay) = a\Delta_C(x, y)$. Thus, under Assumption 1.2, it holds that

$$\begin{aligned}
&\left(0, \frac{\Delta_C(ax, ay) R_{W_b^a}(ax, ay)}{a \mathfrak{D}_{W_b^a}(ax, ay)}\right] \cap \left(0, \frac{a^{-1}\Delta_C(ax, ay)}{a^{-1}\mathfrak{D}_{W_b^a}(ax, ay) + \nu_{W_b^a}(ax, ay) - {W_b^a}'(a)}\right] \\
&= \left(0, \frac{\Delta_C(x, y) R_U(x, y)}{\mathfrak{D}_U(x, y)}\right] \cap \left(0, \frac{\Delta_C(x, y)}{\mathfrak{D}_U(x, y) + \nu_U(x, y) - U'(1)}\right]
\end{aligned}$$

and

$$\begin{aligned}
&\langle C, \Pi^{W_b^a}(C, ax, ay, \varepsilon) \rangle - \inf_{\widetilde{\Pi} \in \Pi(ax, ay)} \langle C, \widetilde{\Pi} \rangle \\
&= \langle C, a\Pi^U(\omega) \rangle - \inf_{\Pi \in \Pi(x,y)} \langle C, a\Pi \rangle \\
&= a \left(\langle C, \Pi^U(\omega) \rangle - \inf_{\Pi \in \Pi(x,y)} \langle C, \Pi \rangle \right) \\
&\le a \left(\Delta_C(x, y) e_U \left(-\frac{\Delta_C(x, y)}{\varepsilon} + \mathfrak{D}_U(x, y) + \nu_U(x, y) \right) \right) \\
&= \Delta_C(x, y) e_{W_b^a} \left(-\frac{\Delta_C(x, y)}{\varepsilon} + \mathfrak{D}_U(x, y) + \nu_U(x, y) \right) \\
&= \frac{\Delta_C(ax, ay)}{a} e_{W_b^a} \left(-\frac{\Delta_C(ax, ay)}{a\varepsilon} + \frac{1}{a} \mathfrak{D}_{W_b^a}(ax, ay) + \nu_{W_b^a}(ax, ay) \right)
\end{aligned}$$

for ε in the interval above. This estimate can be derived directly in a similar way to the proof of Theorem 1.7.

Thus, in the problem (1.2), if we simultaneously scale the data and the domain of a strictly convex function, we obtain essentially the same error estimate in Theorem 1.7, where the domain of a strictly convex function has no effect on the estimate.

4.3. **Invariance of the Kullback–Leibler divergence under scaling of data.** In contrast to the previous subsection, let us now consider the case that the scaling of data does not match the domain of a strictly convex function. For this sake, let $W \in C([0, b]) \cap C^1((0, b]) \cap C^2((0, b))$ satisfy that $W'' > 0$ on $(0, b)$, $\lim_{h \downarrow 0} W'(h) = -\infty$, and $r \mapsto rW''(r)$ is non-decreasing on $(0, b)$ as in the previous subsection and $\widetilde{a}, a > 0$ with $\widetilde{a} < a$.

First, we find that

$$\begin{aligned}\operatorname*{argmin}_{\widetilde{\Pi}\in\Pi(\widetilde{a}x,\widetilde{a}y)}\left(\langle C,\widetilde{\Pi}\rangle+\varepsilon D_{W_b^a}(\widetilde{\Pi},\widetilde{a}x\otimes y)\right)&=\operatorname*{argmin}_{\widetilde{\Pi}\in\Pi(\widetilde{a}x,\widetilde{a}y)}\left(\langle C,a^{-1}\widetilde{\Pi}\rangle+\varepsilon D_{W_b^1}(a^{-1}\widetilde{\Pi},a^{-1}\widetilde{a}x\otimes y)\right)\\&=\operatorname*{argmin}_{\Pi\in\Pi(x,y)}\left(\langle C,\Pi\rangle+\varepsilon D_{W_b^{a/\widetilde{a}}}(\Pi,x\otimes y)\right)\end{aligned}$$

for $\omega=(C,x,y,\varepsilon)\in\Omega$. Thus, scaling the data is equivalent to scaling the domain of a strictly convex function. Then, we can use U satisfying Assumption 1.3 instead of W.

The following proposition suggests that the regularization effect by a Bregman divergence does not vary under scaling of data unless the Bregman divergence is the Kullback–Leibler divergence.

Proposition 4.3. *Let U satisfy Assumption 1.3. Let $\widetilde{a},a>0$ with $\widetilde{a}<a\leq 1$. If there exists $\kappa>0$ such that*

$$D_U(\widetilde{a}z,\widetilde{a}w)=\kappa D_U(az,aw)\quad\text{for } z,w\in\mathcal{P}_K\quad\text{for } K\geq 3. \tag{4.6}$$

Then, there exist $\mu_0,\mu_1\in\mathbb{R}$ and $\lambda>0$ such that $U=(U_o)_{\lambda,\mu_0,\mu_1}$ on $(0,a]$.

Proof. By a similar argument in the implication (D) to (C2), it follows from (4.6) that

$$\widetilde{a}^2U''(\widetilde{a}r)=\kappa a^2U''(ar)\quad\text{for } r\in(0,1).$$

Let $\theta:=\widetilde{a}a^{-1}<1$. Then, the above relation is equivalent to

$$\theta rU''(\theta r)=\kappa\theta^{-1}rU''(r)\quad\text{for } r\in(0,a).$$

The monotonicity of $r\mapsto rU''(r)$ on $(0,a)$ yields $\kappa\theta^{-1}\leq 1$. For $N\in\mathbb{N}$, it turns out that

$$\begin{aligned}U'(a\theta)-U'(a\theta^{N+1})&=\int_{a\theta^{N+1}}^{a\theta}U''(r)\mathrm{d}r=\sum_{n=1}^{N}\int_{a\theta^{n+1}}^{a\theta^n}U''(r)\mathrm{d}r=\sum_{n=1}^{N}\int_{a\theta^{n+1}}^{a\theta^n}rU''(r)\cdot\frac{1}{r}\mathrm{d}r\\&\leq\sum_{n=1}^{N}a\theta^nU''(a\theta^n)\cdot\frac{1}{a\theta^{n+1}}\int_{a\theta^{n+1}}^{a\theta^n}1\mathrm{d}r=\sum_{n=1}^{N}\left(\kappa\theta^{-1}\right)^{n-1}a\theta U''(a\theta)\left(\theta^{-1}-1\right)\\&=\widetilde{a}U''(\widetilde{a})\left(\theta^{-1}-1\right)\sum_{n=1}^{N}\left(\kappa\theta^{-1}\right)^{n-1}.\end{aligned}$$

If $\kappa\theta^{-1}<1$, then

$$\lim_{N\to\infty}(U'(a\theta)-U'(a\theta^{N+1}))\leq\widetilde{a}U''(\widetilde{a})\left(\theta^{-1}-1\right)\cdot\lim_{N\to\infty}\sum_{n=1}^{N}\left(\kappa\theta^{-1}\right)^{n-1}=\widetilde{a}U''(\widetilde{a})\left(\theta^{-1}-1\right)\cdot\frac{1}{1-\kappa\theta^{-1}}<\infty,$$

which contradicts the condition $\lim_{h\downarrow 0}U'(h)=-\infty$. Hence, $\kappa\theta^{-1}=1$ and

$$rU''(r)=\theta rU''(\theta r)\leq rU''(r)\quad\text{for } r\in(0,a),$$

that is, $r\mapsto rU''(r)$ is constant on $(0,a)$. This is equivalent to that there exist $\mu_0,\mu_1\in\mathbb{R}$ and $\lambda>0$ such that $U(r)=\lambda r\log r+\mu_1r+\mu_0$ on $(0,a]$. This completes the proof of the proposition. □

Let U satisfy Assumption 1.3 and $a\in(0,1)$. For the regularized problem of the form

$$\inf_{\widetilde{\Pi}\in\Pi(\widetilde{a}x,\widetilde{a}y)}\left(\langle C,\widetilde{\Pi}\rangle+\varepsilon D_U(\widetilde{\Pi},ax\otimes y)\right)\quad\text{for } \omega=(C,x,y,\varepsilon)\in\Omega,$$

the choice of a is important to give a similar estimate as in Theorem 1.7, since the quantities such as (4.5) may be involved in the estimate. Indeed, if we define

$$\mathfrak{D}_U(ax,ay):=\sup_{\widetilde{\Pi}\in\Pi(ax,ay)}D_U(\widetilde{\Pi},ax\otimes y)$$

then, for $\widetilde{a} > 0$, it turns out that

$$\begin{aligned} d_U(ar, ar_0) = a d_{U^{a^{-1}}}(r, r_0) = a \int_{r_0}^{r} \int_{r_0}^{t} \frac{\mathrm{d}^2}{\mathrm{d}s^2} U^{a^{-1}}(s) \mathrm{d}s\mathrm{d}t = a \int_{r_0}^{r} \int_{r_0}^{t} aU''(as)\mathrm{d}s\mathrm{d}t \\ \geq a \int_{r_0}^{r} \int_{r_0}^{t} \widetilde{a}U''(\widetilde{a}s)\mathrm{d}s\mathrm{d}t = a\widetilde{a}^{-1} d_U(\widetilde{a}r, \widetilde{a}r_0), \end{aligned}$$

consequently,

$$\widetilde{a}^{-1}\mathfrak{D}_U(\widetilde{a}x, \widetilde{a}y) \leq a^{-1}\mathfrak{D}_U(ax, ay)$$

holds with equality if and only if $U = (U_o)_{\lambda,\mu_0,\mu_1}$ holds for some $\mu_0, \mu_1 \in \mathbb{R}$ and $\lambda > 0$.

5. Examples and Comparison

We give examples of U satisfying Assumption 1.3.

5.1. **Model case.**

Recall that $U_o \in C([0,\infty)) \cap C^\infty((0,\infty))$ is defined as

$$U_o(r) := \begin{cases} r\log r & \text{for } r > 0, \\ 0 & \text{for } r = 0. \end{cases}$$

Obviously, U_o satisfies Assumption 1.3 and the normalization (4.2). For $r, r_0 > 0$, we see that

$$d_{U_o}(r, r_0) = U_o(r) - U_o(r_0) - (r - r_0)U_o'(r_0) = r(\log r - \log r_0) - (r - r_0),$$

which yields $D_{U_o}(z, w) = D_{\mathrm{KL}}(z, w)$ for $z, w \in \mathcal{P}_K$.

Let us see that Theorem 1.7 for the case $U = U_o$ coincides with the error estimate given in [20, Theorem 5] interpreted as

$$\langle C, \Pi^{U_o}(\omega)\rangle - \inf_{\Pi\in\Pi(x,y)} \langle C, \Pi\rangle \leq \Delta_C(x, y) \exp\left(-\frac{\Delta_C(x, y)}{\varepsilon} + \mathfrak{D}_{U_o}(x, y) + 1\right) \tag{5.1}$$

for $\varepsilon \in (0, \Delta_C(x, y)/(1 + \mathfrak{D}_{U_o}(x, y))]$, where $\omega = (C, x, y, \varepsilon) \in \Omega$. In our setting, the ℓ^1-radius of $\Pi(x, y)$ defined in [20, Definition 2] is calculated as

$$\max_{\Pi\in\Pi(x,y)} \sum_{i,j} \pi_{ij} = 1.$$

The entropic radius defined in [20, Definition 3] is calculated as

$$\begin{aligned} \sup_{\Pi,\Pi'\in\Pi(x,y)} (S(\Pi) - S(\Pi')) &= \sup_{\Pi,\Pi'\in\Pi(x,y)} (-D_{U_o}(\Pi, x\otimes y) + D_{U_o}(\Pi', x\otimes y)) \\ &= \sup_{\Pi\in\Pi(x,y)} D_{U_o}(\Pi, x\otimes y) - \inf_{\Pi\in\Pi(x,y)} D_{U_o}(\Pi, x\otimes y) \\ &= \mathfrak{D}_{U_o}(x, y), \end{aligned}$$

thanks to the relation $D_{U_o}(\Pi, x\otimes y) = -S(\Pi) + S(x) + S(y)$ for $\Pi \in \Pi(x, y)$. A direct calculation provides

$$e_{U_o}(\tau) = \exp(\tau - 1) : (-\infty, 1] \to (0, 1], \qquad R_{U_o}(x, y) = \frac{e^{\mathfrak{D}_{U_o}(x,y)}}{1 + e^{\mathfrak{D}_{U_o}(x,y)}}, \qquad \nu_{U_o}(x, y) = 2, \qquad U_o'(1) = 1.$$

Thus, our estimate in Theorem 1.7 coincides with (5.1), where the range of the regularization parameter in Theorem 1.7 is given by

$$\left(0, \frac{\Delta_C(x, y)R_{U_o}(x, y)}{\mathfrak{D}_{U_o}(x, y)}\right] \cap \left(0, \frac{\Delta_C(x, y)}{\mathfrak{D}_{U_o}(x, y) + \nu_{U_o}(x, y) - U_o'(1)}\right] = \left(0, \frac{\Delta_C(x, y)}{1 + \mathfrak{D}_{U_o}(x, y)}\right],$$

which coincides with that in [20, Theorem 5].

5.2. **q-logarithmic function.** Let us consider an applicable example other then the model case. From the equivalent between (C0)–(C2), any of U, U', and U'' is on the table for consideration. If we regard $1/U''$ as a deformation function, U' is called a deformed logarithmic function and $-U$ corresponds to the density function of an entropy (see [14, Chapters 10, 11] for details). One typical example of deformed logarithmic functions is the q-logarithmic function. For $q \in \mathbb{R}$, define the *q-logarithmic function* $\ln_q\colon (0,\infty) \to \mathbb{R}$ by

$$\ln_q(t) := \int_1^t s^{-q}\,\mathrm{d}s = \begin{cases} \dfrac{t^{1-q}-1}{1-q} & \text{if } q \neq 1, \\ \log t & \text{if } q = 1. \end{cases}$$

The entropy associated to $\ln_q$ is called the Tsallis entropy (see [14, Chapter 8] for instance). We see that $\lim_{h\downarrow 0} \ln_q(h) = -\infty$ is equivalent to $q \geq 1$. Moreover, the function $t \mapsto t \ln_q'(t) = t^{1-q}$ is non-decreasing on $(0,1)$ if and only if $q \leq 1$ holds. Thus, $\ln_q$ satisfies Assumption 4.1 if and only if $q = 1$ holds, where $U_{\ln_1} = (U_o)_{1,0,-1}$. This means that, if we regard a power function as a deformation function, that is, $1/U''$, only the power function of exponent 1 is applicable to Theorem 1.7. To obtain an example that satisfies Theorem 1.7, we need modify the power function of exponent 1 other than power functions of general exponent.

5.3. **Upper incomplete gamma function.** For $\alpha \in \mathbb{R}$, define $L_\alpha : (0,1) \to \mathbb{R}$ by

$$L_\alpha(t) := -\ln_{1-\alpha}(-\log t) = \begin{cases} -\dfrac{(-\log t)^\alpha - 1}{\alpha} & \text{if } \alpha \neq 0, \\ -\log(-\log t) & \text{if } \alpha = 0. \end{cases}$$

It turns out that

$$\frac{1}{L_\alpha'(t)} = t \cdot (-\log t)^{1-\alpha} > 0 \quad \text{for } t \in (0,1),$$

which can be regarded as a refinement of the power function of exponent 1 since the logarithmic function is referred to as the power function of exponent 0. We see that $\lim_{t\uparrow 1} L_\alpha(t)$ is finite if and only if $\alpha > 0$ holds, and $\lim_{h\downarrow 0} L_\alpha(h) = -\infty$ if and only if $\alpha \geq 0$ holds. Moreover, the function $t \mapsto tL_\alpha'(t) = (-\log t)^{\alpha-1}$ is non-decreasing on $(0,1)$ if and only if $\alpha \leq 1$ holds. Thus, L_α satisfies Assumption 4.1 if and only if $\alpha \in (0,1]$ holds, where $U_{L_1} = U_o$.

In what follows, $\alpha \in (0,1]$ is assumed. We set

$$L_\alpha(1) := \lim_{t\uparrow 1} L_\alpha(t) = \frac{1}{\alpha}.$$

It follows from the change of variables $-\log t = \tau$ that

$$\int_0^1 L_\alpha(t)\mathrm{d}t = -\frac{1}{\alpha}\Gamma(\alpha+1) + \frac{1}{\alpha},$$

where $\Gamma(\cdot)$ is the gamma function. As mentioned in Remark 4.2, if we set

$$\ell_\alpha(t) := -\frac{(-\log t)^\alpha}{\Gamma(\alpha+1)} + 1,$$

for $t \in (0,1]$, then ℓ_α satisfies Assumption 4.1 and

$$U_\alpha(r) := \int_0^r \ell_\alpha(t)\mathrm{d}t = -\frac{1}{\Gamma(\alpha+1)}\Gamma(\alpha+1, -\log r) + r \tag{5.2}$$

satisfies the normalization (4.2), where

$$\Gamma(p,\tau) = \int_\tau^\infty t^{p-1}\exp(-t)\mathrm{d}t$$

is the upper incomplete gamma function for $p > 0$ and $\tau \geq 0$. Note that $\Gamma(p,0) = \Gamma(p)$. Since the inverse function $e_{U_\alpha} : U_\alpha'((0,1]) \to (0,1]$ of $U_\alpha : (0,1] \to U_\alpha'((0,1]) = (-\infty, 1]$ is given by

$$e_{U_\alpha}(\tau) = \exp\left(-\left[-\Gamma(\alpha+1)(\tau-1)\right]^{\frac{1}{\alpha}}\right),$$

the error estimate in Theorem 1.7 for $U = U_\alpha$ is the exponential decay in the case of $\alpha = 1$, which is the same as (5.1), and is more tight if $\alpha \in (0,1)$ as we observed in Subsection 2.3.

The function L_α is introduced to analyze the preservation of concavity by the Dirichlet heat flow in a convex domain on Euclidean space (see [8, 10]).

5.4. **Complementary error function.** Let us give an example of a function L satisfying Assumption 4.1 except for the continuity at $t = 1$. In this case,

$$W(r) := \int_0^r L(t)\mathrm{d}t$$

satisfies Assumption 1.3 except for the continuity and differentiability at $r = 1$ but

$$W_1^a(r) = aW(a^{-1}r) = a\int_0^{a^{-1}r} L(t)\mathrm{d}t = \int_0^r L(a^{-1}t)\mathrm{d}t$$

satisfies Assumption 1.3 if $a > 1$. Then, it might be worth to consider the effect of $a > 1$ on the regularized solution of the form

$$\operatorname*{argmin}_{\Pi\in\Pi(x,y)} \left(\langle C, \Pi\rangle + \varepsilon D_{W_1^a}(\Pi, x\otimes y)\right),$$

as we mentioned in Subsection 4.3.

Let us give an example of such L. Define $H : (0, 1) \to \mathbb{R}$ by the inverse function of

$$\tau \mapsto \frac{1}{\sqrt{\pi}}\int_{-\tau/2}^{\infty} e^{-\sigma^2} d\sigma = \frac{1}{2}\operatorname{erfc}\left(-\frac{\tau}{2}\right),$$

where $\operatorname{erfc}(\tau) = 1 - \operatorname{erf}(\tau)$ is the complementary error function with the error function

$$\operatorname{erf}(\tau) = \frac{2}{\sqrt{\pi}}\int_0^{\tau} e^{-\sigma^2} d\sigma \quad \text{for } \tau \in \mathbb{R},$$

and we used the properties

$$\lim_{\tau\downarrow-\infty} \frac{1}{2}\operatorname{erfc}\left(-\frac{\tau}{2}\right) = 0, \qquad \lim_{\tau\uparrow\infty} \frac{1}{2}\operatorname{erfc}\left(-\frac{\tau}{2}\right) = 1.$$

It is easily seen that $H \in C^\infty((0, 1))$ and $\lim_{t\downarrow0} H(t) = -\infty$. Since we have

$$\frac{1}{2}\operatorname{erfc}\left(-\frac{H(t)}{2}\right) = t \quad \text{for } t \in (0, 1), \qquad \frac{\mathrm{d}}{\mathrm{d}\tau}\left(\frac{1}{2}\operatorname{erfc}\left(-\frac{\tau}{2}\right)\right) = \frac{1}{\sqrt{4\pi}}e^{-\frac{\tau^2}{4}} \quad \text{for } \tau \in \mathbb{R},$$

we find that

$$\begin{aligned}
1 &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}\operatorname{erfc}\left(-\frac{H(t)}{2}\right)\right) = \frac{1}{\sqrt{4\pi}}e^{-\frac{H(t)^2}{4}}H'(t),\\
0 &= \frac{\mathrm{d}^2}{\mathrm{d}t^2}\left(\frac{1}{2}\operatorname{erfc}\left(-\frac{H(t)}{2}\right)\right) = \frac{1}{\sqrt{4\pi}}e^{-\frac{H(t)^2}{4}}\left(-\frac{H(t)}{2}H'(t)^2 + H''(t)\right),
\end{aligned}$$

consequently,

$$\frac{\mathrm{d}}{\mathrm{d}t}(tH'(t)) = H'(t) + tH''(t) = H'(t)\left(1 + \frac{tH(t)}{2}H'(t)\right),$$

for $t \in (0, 1)$. It was proved in [9, Section 4.3] that

$$\inf_{t\in(0,1)} \frac{tH(t)}{2}H'(t) = \lim_{t\downarrow0} \frac{tH(t)}{2}H'(t) = -1$$

in terms of the inverse function of H, and hence $t \mapsto tH'(t)$ is non-decreasing on $(0, 1)$. Thus, H satisfies Assumption 4.1 except for the continuity at $t = 1$.

The function H is also introduced to analyze the preservation of concavity by the Dirichlet heat flow in a convex domain on Euclidean space (see [10]). Although we do not detail here the definition of "F-concavity is preserved by the Dirichlet heat flow in a convex domain on Euclidean space", for $F \in C^2((0, 1))$, if F-concavity is preserved by the Dirichlet heat flow in a convex domain on Euclidean space, then F satisfies Assumption 4.1 except for the continuity at $t = 1$ by [10, Theorems 1.5 and 1.6] and [9, Theorem 2.4 and Subsection 4.3].

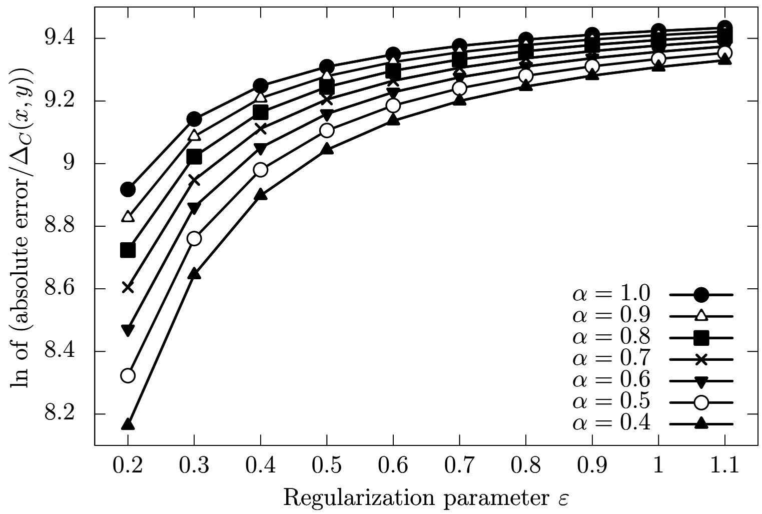

FIGURE 1. Absolute error/$\Delta_C(x, y)$ vs. regularization parameter ε.

6. Numerical experiments

Let $\omega = (C, x, y, \varepsilon) \in \Omega$. Numerical experiments demonstrate that the error $\langle C, \Pi^U(\omega)\rangle - \inf_{\Pi\in\Pi(x,y)}\langle C, \Pi\rangle$ follows the estimate in Theorem 1.7. The function U tested is associated with the function U_α for $\alpha \in (0, 1]$, which is defined in (5.2). The test problem used has the values of the entries of a matrix $C \in \mathbb{R}^{5\times5}$ and test vectors $x, y \in \mathcal{P}_5$ generated in MATLAB.

The regularized problem (1.2) is solved by using the gradient descent method presented in [18, Corollary 4.3]. The method terminates when the Frobenius norm of the gradient becomes 10^{-8}. All computations are performed on a computer with an Intel Core i7-8565U 1.80 GHz central processing unit (CPU), 16 GB of random-access memory (RAM), and the Microsoft Windows 11 Pro 64 bit Version 22H2 Operating System. All programs for implementing the method were coded and run in MATLAB R2020b for double precision floating-point arithmetic with unit roundoff $2^{-53} \simeq 1.1 \cdot 10^{-16}$.

Figure 1 shows the natural logarithm of the ratio of the absolute error $\langle C, \Pi^{U_\alpha}(\omega)\rangle - \inf_{\Pi\in\Pi(x,y)}\langle C, \Pi\rangle$ and $\Delta_C(x, y)$ versus the value of the regularization parameter ε. Here, $\Delta_C(x, y) \simeq 4.6 \cdot 10^{-6}$. We observe that as ε decreases for each value of α, the error decreases. As the value of α decreases for each value of ε, the error decreases. As the value of ε decreases, the methods tend to take more iterations. This is because the regularized problem approaches the given problem as ε approaches zero.

7. Concluding remarks

In this paper, we considered regularization of optimal transport problems via Bregman divergence. We proved that the optimal value of the regularized problem converges to that of the given problem. More precisely, our error estimate becomes faster than exponentially. Numerical experiments showed that regularization by a Bregman divergence outperforms that by the Kullback–Leibler divergence.

There are several future directions subsequent to this study. The time complexity of our regularized problem is left open. It would also be interesting to extend the setting of this paper from a finite set to Euclidean space.

Acknowledgements. KM was supported in part by JSPS KAKENHI Grant Numbers JP20K14356, JP21H03451. KS was supported in part by JSPS KAKENHI Grant Numbers JP22K03425, JP22K18677, JP23H00086. AT was supported in part by JSPS KAKENHI Grant Numbers JP19K03494, JP19H01786. The authors are sincerely grateful to Maria Matveev and Shin-ichi Ohta for helpful discussion.

References

[1] S.-I. Amari, *α-divergence is unique, belonging to both f-divergence and bregman divergence classes*, IEEE Trans. Inf. Theory **55** (2009), no. 11, 4925–4931.

[2] L. Chen, R. Kyng, Y. P. Liu, R. Peng, M. P. Gutenberg, and S. Sachdeva, *Maximum flow and minimum-cost flow in almost-linear time*, 2022 IEEE 63rd Annual Symposium on Foundations of Computer Science, 2022, pp. 612–623.

[3] R. Cominetti and J. S. Martín, *Asymptotic analysis of the exponential penalty trajectory in linear programming*, Math. Program. **67** (1994), 169–187.

[4] M. Daniels, T. Maunu, and P. Hand, *Score-based generative neural networks for large-scale optimal transport*, Proceedings of the Advances in Neural Information Processing Systems, 2021, pp. 12955–12965.

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

[6] L. C. Evans and R. F. Gariepy, *Measure theory and fine properties of functions*, Revised, Textbooks in Mathematics, CRC Press, Boca Raton, FL, 2015.

[7] S. C. Fang, *An unconstrained convex programming view of linear programming*, Z. Oper. Res. **36** (1992), no. 2, 149–161.

[8] K. Ishige, P. Salani, and A. Takatsu, *To logconcavity and beyond*, Commun. Contemp. Math. **22** (2020), no. 2, 1950009, 17.

[9] K. Ishige, P. Salani, and A. Takatsu, *Hierarchy of deformations in concavity*, Info. Geo. (2022).

[10] K. Ishige, P. Salani, and A. Takatsu, *Characterization of F-concavity preserved by the Dirichlet heat flow*, arXiv:2207.13449 (2022).

[11] A. Jambulapati, A. Sidford, and K. Tian, *A direct $\tilde{O}(1/\epsilon)$ iteration parallel algorithm for optimal transport*, Proceedings of the Advances in Neural Information Processing Systems, 2019.

[12] M. Klatt, C. Tameling, and A. Munk, *Empirical regularized optimal transport: Statistical theory and applications*, SIAM J. Math. Data Sci. **2** (2020), no. 2, 419–443.

[13] B. Muzellec, R. Nock, G. Patrini, and F. Nielsen, *Tsallis regularized optimal transport and ecological inference*, Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017, pp. 2387–2393.

[14] J. Naudts, *Generalised Thermostatistics*, Springer-Verlag, Berlin, Germany, 2011.

[15] S.-I. Ohta and A. Takatsu, *Displacement convexity of generalized relative entropies. II*, Comm. Anal. Geom. **21** (2013), no. 4, 687–785.

[16] O. Pele and M. Werman, *Fast and robust earth mover's distances*, 2009 IEEE 12th International Conference on Computer Vision, 2009, pp. 460–467.

[17] G. Peyré and M. Cuturi, *Computational Optimal Transport: With Applications to Data Science*, Vol. 11, Now Publishers, Delft, Netherlands, 2019.

[18] A. Takatsu, *Relaxation of optimal transport problem via strictly convex functions*, arXiv:2102.07336 (2021).

[19] C. Villani, *Optimal Transport: Old and New*, Springer-Verlag, Berlin, Germany, 2009.

[20] J. Weed, *An explicit analysis of the entropic penalty in linear programming*, Proceedings of the 31st Conference on Learning Theory, 2018, pp. 1841–1855.

(K. Morikuni) Institute of Systems and Information Engineering, University of Tsukuba, 1-1-1 Tennodai, Tsukuba-shi, Ibaraki 305-8573, Japan
Email address: morikuni@cs.tsukuba.ac.jp

(K. Sakakibara) Faculty of Mathematics and Physics, Institute of Science and Engineering, Kanazawa University, Kakuma-machi, Kanazawa-shi, Ishikawa 920-1192, Japan; RIKEN iTHEMS, 2-1 Hirosawa, Wako-shi, Saitama 351-0198, Japan
Email address: ksakaki@se.kanazawa-u.ac.jp

(A. Takatsu) Department of Mathematical Sciences, Tokyo Metropolitan University, 1-1 Minami-osawa, Hachioji-shi, Tokyo 192-0397, Japan; RIKEN Center for Advanced Intelligence Project (AIP), Nihonbashi 1-chome Mitsui Building, 15th floor, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan.
Email address: asuka@tmu.ac.jp