

# STOCHASTIC OPTIMAL TRANSPORT IN BANACH SPACES FOR REGULARIZED ESTIMATION OF MULTIVARIATE QUANTILES

BERNARD BERCU, JÉRÉMIE BIGOT AND GAUTHIER THURIN

*Université de Bordeaux  
Institut de Mathématiques de Bordeaux et CNRS (UMR 5251)*

**ABSTRACT.** We introduce a new stochastic algorithm for solving entropic optimal transport (EOT) between two absolutely continuous probability measures  $\mu$  and  $\nu$ . Our work is motivated by the specific setting of Monge-Kantorovich quantiles where the source measure  $\mu$  is either the uniform distribution on the unit hypercube or the spherical uniform distribution. Using the knowledge of the source measure, we propose to parametrize a Kantorovich dual potential by its Fourier coefficients. In this way, each iteration of our stochastic algorithm reduces to two Fourier transforms that enables us to make use of the Fast Fourier Transform (FFT) in order to implement a fast numerical method to solve EOT. We study the almost sure convergence of our stochastic algorithm that takes its values in an infinite-dimensional Banach space. Then, using numerical experiments, we illustrate the performances of our approach on the computation of regularized Monge-Kantorovich quantiles. In particular, we investigate the potential benefits of entropic regularization for the smooth estimation of multivariate quantiles using data sampled from the target measure  $\nu$ .

**Keywords :** Entropic Optimal Transport; Monge-Kantorovich quantiles; Multivariate quantiles; Stochastic optimization in a Banach space; Multiple Fourier Series.

**MSCcodes :** 62H12, 62G20, 62L20

## 1. INTRODUCTION

Consider a probability distribution  $\nu$  supported on a subset  $\mathcal{Y} \subset \mathbb{R}^d$ . In the scalar case  $d = 1$ , the quantile function of  $\nu$  is nothing else than the generalized inverse  $F_\nu^{-1}$  of the cumulative distribution function  $F_\nu$  of  $\nu$ . However, in the multi-dimensional case  $d \geq 2$ , there does not exist a standard notion of multivariate quantiles as there is no canonical ordering in  $\mathbb{R}^d$ . Therefore, various notions of quantiles in dimension  $d \geq 2$  have been proposed in the statistical literature, some of them being inspired by the notion of data depth introduced in [41] and other based on geometric principles [11]. We refer the reader to Section 1.2 in [24] for a recent survey of the many existing concepts of multivariate quantiles.

The aim of this paper is to investigate the notion of Monge-Kantorovich (MK) quantiles using the theory of quadratic optimal transport (OT) that has been introduced in [12]. The basic concepts of MK quantiles can be summarized as follows. For  $\mathcal{P}_d$  the set of Lebesgue-absolutely continuous probability measures on  $\mathbb{R}^d$ , one first considers a reference distribution  $\mu \in \mathcal{P}_d$ , supported on a convex and compact set  $\mathcal{X} \subset \mathbb{R}^d$ . As discussed in [12], this reference measure  $\mu$  is typically either the uniform distribution on  $\mathcal{X} = [0, 1]^d$  or

---

The authors gratefully acknowledge financial support from the Agence Nationale de la Recherche (MaS-DOL grant ANR-19-CE23-0017). Jérémie Bigot is a member of Institut Universitaire de France (IUF), and this work has also been carried out with financial support from the IUF.

the *spherical uniform*<sup>1</sup> distribution on the unit ball. Then, the MK quantile function of a square integrable probability measure  $\nu$ , with respect to  $\mu$ , is defined as the optimal transport map  $Q : \mathcal{X} \rightarrow \mathcal{Y}$  between  $\mu$  and  $\nu$ . More precisely, let  $X$  be a random vector with distribution  $\mu$ . Then,  $Q$  is the optimal mapping satisfying

$$(1.1) \quad Q = \operatorname{argmin}_{T : T\#\mu = \nu} \mathbb{E} \left( \frac{1}{2} \|X - T(X)\|^2 \right),$$

the notation  $T\#\mu = \nu$  meaning that  $T : \mathcal{X} \rightarrow \mathcal{Y}$  is a push-forward map from  $\mu$  to  $\nu$ , and  $\|\cdot\|$  standing for the usual Euclidean norm in  $\mathbb{R}^d$ . There, one can rely on the well-known Kantorovich duality (see e.g. [39, 42]) of optimal transport to characterize  $Q$ . Since  $\mu$  is absolutely continuous, it is well-known [7, 17] that  $Q$  can be rewritten as

$$(1.2) \quad Q(x) = x - \nabla u_0(x) = \nabla \psi_0(x) \quad \text{with} \quad \psi_0(x) = x - u_0(x),$$

for  $\mu$ -almost every  $x \in \mathcal{X}$ . In the above equation,  $u_0$  denotes the unique solution, up to a scalar translation, of the Kantorovich dual formulation of OT

$$(1.3) \quad u_0 \in \arg \max_{u \in L^1(\mu)} \int_{\mathcal{X}} u(x) d\mu(x) + \int_{\mathcal{Y}} u^c(y) d\nu(y),$$

where  $u^c : \mathcal{Y} \rightarrow \mathbb{R}$  is the  $c$ -conjugate of a function  $u \in L^1(\mu)$  in the sense that

$$u^c(y) = \inf_{x \in \mathcal{X}} \{c(x, y) - u(x)\} \quad \text{with} \quad c(x, y) = \frac{1}{2} \|x - y\|^2.$$

Based on a sample  $(Y_1, \dots, Y_n)$  from  $\nu$ , it is natural to estimate  $Q$  by the plug-in estimator

$$(1.4) \quad \widehat{Q}_n = \operatorname{argmin}_{T : T\#\mu = \widehat{\nu}_n} \mathbb{E} \left( \frac{1}{2} \|X - T(X)\|^2 \right) \quad \text{where} \quad \widehat{\nu}_n = \frac{1}{n} \sum_{j=1}^n \delta_{Y_j}.$$

Alternatively, one has from (1.2) and (1.3) that for all  $x \in \mathcal{X}$ ,  $\widehat{Q}_n(x) = x - \nabla \widehat{u}_n(x)$  where

$$(1.5) \quad \widehat{u}_n \in \arg \max_{u \in L^1(\mu)} \int_{\mathcal{X}} u(x) d\mu(x) + \int_{\mathcal{Y}} u^c(y) d\widehat{\nu}_n(y).$$

Finding a numerical solution to the problem (1.5) involves the use of optimization techniques in the Banach space  $L^1(\mu)$  which is a delicate issue that is tackled in the present paper. More precisely, we propose a new stochastic algorithm in order to estimate the dual potential  $u_0$  using computational optimal transport [35] based on entropic regularization [18], which also yields a new regularized estimator of the MK quantile function  $Q$ .

In the last years, the benefit of this regularization has been to allow the use of OT based methods in statistics and machine learning. In this paper, we also advocate the use of entropic OT (EOT) to obtain an estimator of the dual potential  $u_0$  that is smoother than  $\widehat{u}_n$ , leading to an estimator of the MK quantile function  $Q$  that is also smoother than  $\widehat{Q}_n$ . More precisely, we recall that the dual formulation of EOT as formulated in [21] is

$$(1.6) \quad \max_{u \in L^1(\mu)} \int_{\mathcal{X}} u(x) d\mu(x) + \int_{\mathcal{Y}} u^{c,\varepsilon}(y) d\nu(y) - \varepsilon$$

---

<sup>1</sup>*Spherical uniform* refers to the distribution  $\mu_S$  of a random vector  $X = R\Phi$  where  $R$  and  $\Phi$  are independent and drawn uniformly from  $[0,1]$  and the unit hypersphere  $\mathbb{S}^{d-1} = \{\varphi \in \mathbb{R}^d : \|\varphi\| = 1\}$ , respectively.

where  $\varepsilon \geq 0$  stands for a regularization parameter and  $u^{c,\varepsilon}$  is the smooth conjugate of  $u \in L^1(\mu)$  defined, for  $\varepsilon > 0$ , by

$$(1.7) \quad u^{c,\varepsilon}(y) = -\varepsilon \log \left( \int_{\mathcal{X}} \exp \left( \frac{u(x) - c(x,y)}{\varepsilon} \right) d\mu(x) \right)$$

and  $u^{c,0}(y) = u^c(y)$ . In what follows, a function that can be expressed as a smooth conjugate will be called a *regularized  $c$ -transform*. The quadratic cost function  $c$  belongs to  $L^1(\mu \otimes \nu)$  as soon as  $\nu$  has a finite second moment. Thus, it is known that, up to an additive constant, the solution of (1.6) is unique for any  $\varepsilon > 0$ , see e.g. the discussion in [4, Section 2]. The estimation of such a solution is the target of this work. To this end, we mainly focus on the setting where  $\mu$  is the uniform distribution on  $\mathcal{X} = [0, 1]^d$ , and we parametrize a dual function  $u \in L^1(\mu)$  by its decomposition in the standard Fourier basis  $\phi_\lambda(x) = e^{2\pi i \langle \lambda, x \rangle}$ , for  $\lambda \in \mathbb{Z}^d$ , that is

$$(1.8) \quad u(x) = \sum_{\lambda \in \Lambda} \theta_\lambda \phi_\lambda(x).$$

Based on a sample  $(Y_1, \dots, Y_n)$  from  $\nu$ , we estimate the Fourier coefficients  $\theta = (\theta_\lambda)_{\lambda \in \Lambda}$  via a stochastic algorithm  $\hat{\theta}_n = (\hat{\theta}_{n,\lambda})_{\lambda \in \Lambda}$ , which allows us to propose a natural plug-in estimator

$$\hat{u}_\varepsilon^n(x) = \sum_{\lambda \in \Lambda} \hat{\theta}_{n,\lambda} \phi_\lambda(x).$$

An estimator of  $Q$  is then induced from the entropic analog of (1.2) using a regularized  $c$ -transform of  $\hat{u}_\varepsilon^n$ , and the notion of barycentric projection (see e.g. [36, Section 3]). From a computational point of view, our stochastic algorithm, described in Section 2, mainly involves the use of two Fast Fourier Transforms (FFT) and the choice of a regular grid of  $p$  points in  $\mathcal{X}$  to estimate a set of  $p$  Fourier coefficients. The computational cost of our recursive procedure at each iteration is thus of order  $\mathcal{O}(p \log(p))$ . Therefore, its numerical cost, at each iteration, is *independent* of the sample size  $n$  for which multivariate quantiles need to be computed.

## 1.1. Relation to previous works.

1.1.1. *Comparison to other algorithms for solving OT.* The estimation of  $Q$  using the plug-in estimator  $\hat{Q}_n$  based on the empirical measure  $\hat{\nu}_n$  can begin with various computational strategies to solve OT between  $\mu$  and  $\hat{\nu}_n$ . One can replace  $\mu$  by a discrete measure  $\hat{\mu}_n$  on a regular grid and then solve a *discrete* OT problem between  $\hat{\mu}_n$  and  $\hat{\nu}_n$  as in [12, 24]. However, the computational cost of such a discrete OT problem is potentially very high because it scales cubically in the number of observations [35]. It is also proposed in [23] to compute the semi-dual problem (1.5) using the Newton-type algorithms proposed in [28].

In the present paper, we suggest a new strategy which relies on a parametrization of the dual function  $u$  by its Fourier coefficients, which allows us to better make use of the knowledge of the reference distribution  $\mu$ . Beyond the context of multivariate quantiles, the estimation of OT maps is an active area of research. Dual potentials were parameterized by wavelets expansions in [25], and another popular approach is based on neural networks, see e.g. [9, 29, 30]. Other recent contributions on the estimation of OT maps also include [19, 31, 34]. In [36], the entropic map has been studied as a natural alternative with respect to entropic regularization, and we follow this line of work in the quantiles' context.

Stochastic algorithms for solving the *semi-discrete* OT problem (1.5) between an absolutely continuous measure  $\mu$  and the empirical measure  $\hat{\nu}_n$  have already been proposed

in [4, 5, 21]. Dual functions  $v \in L^1(\widehat{\nu}_n)$  can be identified to their values  $v(Y_i)$  for  $1 \leq i \leq n$ , which yields, for  $\varepsilon \geq 0$ , the following stochastic optimization problem

$$(1.9) \quad \min_{v \in \mathbb{R}^n} \int_{\mathcal{X}} \varepsilon \log \left( \frac{1}{n} \sum_{j=1}^n \exp \left( \frac{v_j - c(x, Y_j)}{\varepsilon} \right) \right) d\mu(x) - \frac{1}{n} \sum_{j=1}^n v_j + \varepsilon.$$

However, these approaches are based on a sample  $(X_1, \dots, X_m)$  from  $\mu$ , to solve the OT problem between *the absolutely continuous measure  $\mu$  and the discrete measure  $\widehat{\nu}_n$*  when  $m \rightarrow +\infty$  and  $n$  is held fixed. The originality of our approach is to make use of a sample  $(Y_1, \dots, Y_n)$  from  $\nu$  to solve the regularized OT problem between *two absolutely continuous measures  $\mu$  and  $\nu$*  when  $n \rightarrow +\infty$ . In this continuous setting, [21] also proposed a RKHS parametrization of a pair of dual potentials, which is much different from the Fourier decomposition of a dual potential in the semi-dual formulation as proposed in this paper.

**1.1.2. Comparison with existing works for MK quantiles estimation.** The convergence properties of the empirical transport map (1.4) to estimate the un-regularized MK quantile map (1.1) have been studied in [12, 23]. Nevertheless, these estimators take their values in the sample  $(Y_1, \dots, Y_n)$ , and regularizing is required to interpolate between these observations. This was done in [3, 24] based on optimal couplings  $(X_n, Y_n)$ , inherited from discrete OT. The use of Moreau envelopes in [24] preserves the cyclical monotonicity as well as the couplings  $(X_n, Y_n)$ . These are ideal theoretical properties, but a supplementary gradient descent is required when computing a single  $Q(x)$  for  $x \in \mathcal{X}$ . This is alleviated in [3] with an approximation of  $Q$  rather than an interpolation. More precisely, given the unregularized solution  $v$  of the problem (1.9) for  $\varepsilon = 0$ , the authors approximate its  $c$ -transform  $v^c$  by a LogSumExp. This yields a smooth estimator that is cyclically monotone, but based on an un-regularized dual potential  $v$ . In comparison, the use of EOT in our procedure represents a step towards more regularization, with a cyclically monotone estimator related to recent advances in computational OT. Note that EOT was also recently used in the MK quantiles' framework in [10, 32].

**1.2. Organization of the paper.** Our paper is organized as follows. Section 2 details the formulation of our algorithm in the space of Fourier coefficients. The main results about the convergence of our stochastic algorithm are given in Section 3. In Section 4, we state various keystone properties of the objective functions involved in the stochastic formulation of EOT in the space of Fourier coefficients. Then, in Section 5, we illustrate the performances of our new algorithm on simulated data. In particular, the methodology to obtain a map from the *spherical uniform* distribution instead of the uniform distribution on the unit hypercube is explained. In these numerical experiments, by letting  $\varepsilon$  varying, we also study the effect of the entropic regularization on the estimation of the MK quantile function  $Q$ . A conclusion and a discussion on some perspectives are given in Section 6. All the proofs are postponed to a technical Appendix. Finally, additional proofs on the differentiability of the objective functions are given in supplement materials.

For the sake of reproducible research, the Python codes for the experiments carried out in this paper are available at [https://github.com/gauthierthurin/SGD\\_Space\\_Fourier\\_coeffs](https://github.com/gauthierthurin/SGD_Space_Fourier_coeffs).

## 2. A NEW STOCHASTIC ALGORITHM IN THE SPACE OF FOURIER COEFFICIENTS

**2.1. Our approach.** From now on and throughout the paper,  $\mu$  is assumed to be the uniform distribution on  $\mathcal{X} = [0, 1]^d$ , except in some of the numerical experiments carried out in Section 5 where a change of variable enables to consider the spherical uniform

distribution for which  $\mathcal{X} = \mathbb{B}^d$ . Then, we consider the normalization condition for the dual potentials

$$(2.1) \quad \int_{\mathcal{X}} u(x) d\mu(x) = 0.$$

Taking the support of  $\mu$  to be equal to  $[0, 1]^d$  is motivated by the choice to parametrize a dual function  $u \in L^1(\mu)$ , satisfying the identifiability condition (2.1), by its decomposition in the standard Fourier basis  $\phi_\lambda(x) = e^{2\pi i \langle \lambda, x \rangle}$ , for  $\lambda \in \mathbb{Z}^d$ , that is

$$u(x) = \sum_{\lambda \in \Lambda} \theta_\lambda \phi_\lambda(x),$$

where  $\Lambda = \mathbb{Z}^d \setminus \{0\}$  and  $\theta = (\theta_\lambda)_{\lambda \in \Lambda}$  are the Fourier coefficients of  $u$ ,

$$\theta_\lambda = \int_{\mathcal{X}} \overline{\phi_\lambda(x)} u(x) d\mu(x).$$

We refer to [40] for an introduction to multiple Fourier series on the *flat torus*  $\mathbb{T}^d = \mathbb{R}^d / \mathbb{Z}^d$ . Hereafter,  $\mathbb{T}^d$  stands for the set of equivalence classes  $[x] = \{x + k ; k \in \mathbb{Z}^d\}$  for all  $x \in [0, 1]^d$ . With a slight abuse of notation, we identify  $\mathbb{T}^d$  to its fundamental domain  $[0, 1]^d$ , so that integration on  $\mathbb{T}^d$  is Lebesgue-integration on  $[0, 1]^d$ , see [40] or [15, 31] in the OT literature. Then, for a given regularization parameter  $\varepsilon > 0$ , we rewrite the dual problem (1.6) with this parametrization, to consider, for  $\ell_1(\Lambda)$  defined hereafter, the following *stochastic convex minimisation* problem

$$(2.2) \quad \theta^\varepsilon = \underset{\theta \in \ell_1(\Lambda)}{\operatorname{argmin}} H_\varepsilon(\theta) \quad \text{with} \quad H_\varepsilon(\theta) = \mathbb{E} [h_\varepsilon(\theta, Y)]$$

where  $Y$  is a random vector with distribution  $\nu$  and

$$h_\varepsilon(\theta, y) = \varepsilon \log \left( \int_{\mathcal{X}} \exp \left( \frac{\sum_{\lambda \in \Lambda} \theta_\lambda \phi_\lambda(x) - c(x, y)}{\varepsilon} \right) d\mu(x) \right) + \varepsilon.$$

There, we refer to [26, Chapter 8] for a basic course on Fréchet differentiability and Taylor formulas for functions between Banach spaces. In Section 4, it is shown that, for every  $y \in \mathcal{Y}$ , the function  $\theta \mapsto h_\varepsilon(\theta, y)$  is Fréchet differentiable only if  $\theta$  belongs to the convex set

$$\ell_1(\Lambda) = \left\{ \theta = (\theta_\lambda)_{\lambda \in \Lambda} \in \mathbb{C}^\Lambda : \theta_{-\lambda} = \overline{\theta_\lambda} \text{ and } \|\theta\|_{\ell_1} = \sum_{\lambda \in \Lambda} |\theta_\lambda| < +\infty \right\}.$$

Moreover, its differential  $D_\theta h_\varepsilon(\theta, y)$  is identified as an element of the dual Banach space

$$\ell_\infty(\Lambda) = \left\{ v = (v_\lambda)_{\lambda \in \Lambda} \in \mathbb{C}^\Lambda : v_{-\lambda} = \overline{v_\lambda} \text{ and } \|v\|_{\ell_\infty} = \sup_{\lambda \in \Lambda} |v_\lambda| < +\infty \right\}.$$

The components of the first order Fréchet derivative  $D_\theta h_\varepsilon(\theta, y)$  are the partial derivatives

$$(2.3) \quad \frac{\partial h_\varepsilon(\theta, y)}{\partial \theta_\lambda} = \int_{\mathcal{X}} \overline{\phi_\lambda(x)} F_{\theta, y}(x) d\mu(x)$$

that are the Fourier coefficients of the function

$$(2.4) \quad F_{\theta, y}(x) = \frac{\exp \left( \frac{\sum_{\lambda \in \Lambda} \theta_\lambda \phi_\lambda(x) - c(x, y)}{\varepsilon} \right)}{\int_{\mathcal{X}} \exp \left( \frac{\sum_{\lambda \in \Lambda} \theta_\lambda \phi_\lambda(x) - c(x, y)}{\varepsilon} \right) d\mu(x)}.$$

One can observe that  $F_{\theta,y}$  is a probability density function, which is a key property that we shall repeatedly use. In this paper, we shall analyze (2.2) as a stochastic convex minimisation problem over the Banach space  $(\ell_1(\Lambda), \|\cdot\|_{\ell_1})$ , that corresponds to the formulation of a regularized dual problem of OT in the space of Fourier coefficients.

Imposing that the Fourier coefficients  $\theta = (\theta_\lambda)_{\lambda \in \Lambda}$  form an absolutely convergent series implicitly requires that the optimal dual potential minimizing (1.6) satisfy periodic conditions at the boundary of  $[0, 1]^d$ . For readability of the paper, a detailed discussion on sufficient conditions for the un-regularized optimal dual potential  $u_0$  to be periodic is postponed to D in the supplementary material.

Let  $(Y_n)$  be a sequence of independent random vectors sharing the same distribution  $\nu$ . In the spirit of [38], we propose to estimate the solution of (2.2) by considering the stochastic algorithm in the Banach space  $(\ell_1(\Lambda), \|\cdot\|_{\ell_1})$  defined, for all  $n \geq 0$ , by

$$(2.5) \quad \hat{\theta}_{n+1} = \hat{\theta}_n - \gamma_n W D_{\theta} h_{\varepsilon}(\hat{\theta}_n, Y_{n+1})$$

where  $\gamma_n = \gamma n^{-c}$  with  $\gamma > 0$  and  $1/2 < c \leq 1$ , which clearly implies the standard conditions

$$(2.6) \quad \sum_{n=0}^{\infty} \gamma_n = +\infty \quad \text{and} \quad \sum_{n=0}^{\infty} \gamma_n^2 < +\infty.$$

Moreover,  $W$  is the following linear operator

$$\begin{cases} W : (\ell_{\infty}(\Lambda), \|\cdot\|_{\ell_{\infty}}) & \rightarrow (\ell_1(\Lambda), \|\cdot\|_{\ell_1}) \\ v = (v_\lambda)_{\lambda \in \Lambda} & \mapsto w \odot v = (w_\lambda v_\lambda)_{\lambda \in \Lambda} \end{cases}$$

where  $w = (w_\lambda)_{\lambda \in \Lambda}$  is a *deterministic sequence of positive weights* satisfying the normalizing condition

$$(2.7) \quad \|w\|_{\ell_1} = \sum_{\lambda \in \Lambda} w_\lambda < +\infty.$$

A main difficulty arising here is that the space  $\ell_1(\Lambda)$  of parameters differs from its dual space  $\ell_{\infty}(\Lambda)$  to which the Fréchet derivative  $D_{\theta} h_{\varepsilon}(\theta, y)$  belongs. This is a classical issue when considering convex optimization in Banach spaces, see e.g. [8], and this is the reason why we introduce the linear operator  $W$  in (2.5) that maps  $\ell_{\infty}(\Lambda)$  to  $\ell_1(\Lambda)$ . The use of the linear operator  $W$  also induces two weighted norms on the space

$$\ell_2(\Lambda) = \left\{ \theta = (\theta_\lambda)_{\lambda \in \Lambda} \in \mathbb{C}^\Lambda : \theta_{-\lambda} = \overline{\theta_\lambda} \text{ and } \|\theta\|_{\ell_2}^2 = \sum_{\lambda \in \Lambda} |\theta_\lambda|^2 < +\infty \right\}.$$

One can observe that we clearly have  $\ell_1(\Lambda) \subset \ell_2(\Lambda)$ .

**Definition 2.1.** *For every  $\theta \in \ell^2(\Lambda)$  and for a sequence  $w = (w_\lambda)_{\lambda \in \Lambda}$  of positive weights satisfying (2.7), we define the two weighted norms*

$$(2.8) \quad \|\theta\|_W^2 = \sum_{\lambda \in \Lambda} w_\lambda |\theta_\lambda|^2 \quad \text{and} \quad \|\theta\|_{W^{-1}}^2 = \sum_{\lambda \in \Lambda} w_\lambda^{-1} |\theta_\lambda|^2.$$

The aim of this paper is to establish consistency results for the stochastic algorithm given by (2.5). Hereafter, a *regularized estimator* of the optimal potential defined, for  $x \in \mathcal{X}$ , by

$$(2.9) \quad u_{\varepsilon}(x) = \sum_{\lambda \in \Lambda} \theta_{\lambda}^{\varepsilon} \phi_{\lambda}(x)$$

is naturally given by

$$(2.10) \quad \hat{u}_\varepsilon^n(x) = \sum_{\lambda \in \Lambda} \hat{\theta}_{n,\lambda} \phi_\lambda(x).$$

In practice, our numerical procedure starts by considering a discretization of the dual potential  $u$  over a regular grid  $\mathcal{X}_p = \{x_1, \dots, x_p\}$  of points in  $\mathcal{X}$ . This allows us to compute the corresponding set of Fourier coefficients at frequencies  $\Lambda_p$  of size  $p$  by the Fast Fourier Transform (FFT). Then, the sequence  $(\hat{\theta}_{n,\lambda})_{\lambda \in \Lambda_p}$  satisfying (2.5) is easily implemented using, at each iteration, the FFT and its inverse, see Algorithm 1 below. Hence, the computational cost, at each iteration, of our algorithm is of order  $\mathcal{O}(p \log(p))$ , while the cost of the celebrated Sinkhorn algorithm [18] is  $\mathcal{O}(pn)$ , using a discrete source measure supported on  $\mathcal{X}_p$ , and the one of the stochastic algorithms proposed in [4, 5, 21] is  $\mathcal{O}(n)$  at each iteration.

In our approach, the computational cost depends on the size  $p$  of the grid on  $\mathcal{X}_p$  that is fixed by the user. This size  $p$  does not require to be particularly large, as showed by numerical experiments. However, we stress that this appealing computational cost of  $\mathcal{O}(p \log(p))$  comes with a drawback regarding the dimension. Indeed, the number  $p$  of points in a uniform grid on  $[0, 1]^d$  grows exponentially with  $d$ . Thus, a standard implementation of the FFT on a uniform grid becomes difficult for medium dimensions such as  $d = 10$ . Extending our work to the high-dimensional setting would require the study of more sophisticated FFTs as proposed in [37], but this issue is beyond the scope of this paper.

---

**Algorithm 1** Stochastic algorithm (2.5)

---

```

Initialize  $N \in \mathbb{N}$ ,  $\mathcal{X}_p = \{x_1, \dots, x_p\}$ ,  $u \in \mathbb{R}^p$  and  $W \in \mathbb{R}^{p \times p}$ 
 $\theta \leftarrow \text{FFT}(u)$ 
while  $n \leq N$  do
   $y \leftarrow Y_n$ 
   $u \leftarrow \text{IFFT}(\theta)$ 
  for  $i \in \{1, \dots, p\}$  do
     $F[i] \leftarrow \exp\left((u[i] - c(x_i, y))/\varepsilon\right)$ 
  end for
   $F \leftarrow F / \text{mean}(F)$  ▷ estimate of (2.4)
   $\text{grad} \leftarrow \text{FFT}(F)$ 
   $\theta \leftarrow \theta - \gamma_n W \cdot \text{grad}$ 
end while

```

---

**2.2. The barycentric projection.** Inspired by (1.2), we could propose to estimate the MK quantile function via the *regularized estimator*  $\hat{Q}_\varepsilon^n(x) = x - \nabla \hat{u}_\varepsilon^n(x)$ . However,  $\hat{u}_\varepsilon^n$  is not necessarily a concave function, and thus  $\hat{Q}_\varepsilon^n$  does not correspond to the gradient of a convex function, that is the desired multivariate monotonicity for a quantile function, as argued in [24]. To the contrary, the *entropic map* studied in [36] is the gradient of a convex function as shown in [13][Lemma 1]. Since the entropic map can be estimated from any solution of the EOT problem (1.6), we propose in this paper the following estimator



derived from (2.10),

$$(2.11) \quad \hat{Q}_\varepsilon^n(x) = \sum_{j=1}^n \hat{F}_j(x) Y_j \quad \text{where} \quad \hat{F}_j(x) = \frac{\exp\left(\frac{(\hat{u}_\varepsilon^n)^{c,\varepsilon}(Y_j) - c(x, Y_j)}{\varepsilon}\right)}{\sum_{\ell=1}^n \exp\left(\frac{(\hat{u}_\varepsilon^n)^{c,\varepsilon}(Y_\ell) - c(x, Y_\ell)}{\varepsilon}\right)},$$

that is obtained by computing the smooth conjugate  $(\hat{u}_\varepsilon^n)^{c,\varepsilon} \in \mathbb{R}^n$  of  $\hat{u}_\varepsilon^n$ . Note that if one denotes by  $((\hat{u}_\varepsilon^n)^{c,\varepsilon})^{c,\varepsilon}(x)$  the smooth conjugate of  $(\hat{u}_\varepsilon^n)^{c,\varepsilon}$  at  $x$ , then our estimator can also be expressed as

$$\hat{Q}_\varepsilon^n(x) = x - \nabla((\hat{u}_\varepsilon^n)^{c,\varepsilon})^{c,\varepsilon}(x).$$

Recall that an alternative algorithm to solve the semi-discrete EOT problem is to consider the formulation (1.9) as studied in [4, 5, 21]. Based on independent samples  $X_1, \dots, X_m$  from  $\mu$ , these works approach the unique solution  $\tilde{v}_n \in \mathbb{R}^n$  of the problem (1.9) when  $m \rightarrow +\infty$  and  $n$  is held fixed. Then, one can estimate the entropic map using, for all  $x \in \mathcal{X}$ ,

$$(2.12) \quad \tilde{Q}_\varepsilon^n(x) = \sum_{j=1}^n \tilde{F}_j(x) Y_j \quad \text{where} \quad \tilde{F}_j(x) = \frac{\exp\left(\frac{\tilde{v}_{n,j} - c(x, Y_j)}{\varepsilon}\right)}{\sum_{\ell=1}^n \exp\left(\frac{\tilde{v}_{n,\ell} - c(x, Y_\ell)}{\varepsilon}\right)}.$$

The numerical performances of  $\hat{Q}_\varepsilon^n(x)$  are compared to those of  $\tilde{Q}_\varepsilon^n$  in Section 5.

### 3. MAIN RESULTS

In order to state our main results, it is necessary to introduce two suitable assumptions related to the optimal sequence of Fourier coefficients  $\theta^\varepsilon = (\theta_\lambda^\varepsilon)_{\lambda \in \Lambda}$  and the second order Fréchet derivative of the function  $H_\varepsilon$  given by (2.2).

**Assumption 3.1.** *The sequence of Fourier coefficients  $(\theta_\lambda^\varepsilon)_{\lambda \in \Lambda}$  satisfies  $\|\theta^\varepsilon\|_{W^{-1}} < +\infty$ .*

**Assumption 3.2.** *For any regularization parameter  $\varepsilon > 0$ , there exists a positive constant  $c_\varepsilon$  such that the second order Fréchet derivative of the function  $H_\varepsilon$  evaluated at the optimal value  $\theta^\varepsilon$  satisfies, for any  $\tau \in \ell_1(\Lambda)$ ,*

$$(3.1) \quad D^2 H_\varepsilon(\theta^\varepsilon)[\tau, \tau] \geq c_\varepsilon \|\tau\|_{\ell^2}^2.$$

Our main theoretical result is devoted to the almost sure convergence of the random sequence  $(\hat{\theta}_n)_n$  defined by (2.5).

**Theorem 3.1.** *Suppose that the initial value  $\hat{\theta}_0$  is any random element in  $\ell^2(\Lambda)$  such that  $\|\hat{\theta}_0\|_{W^{-1}} < +\infty$ . Then, under Assumptions 3.1 and 3.2, the sequence  $(\hat{\theta}_n)$  converges almost surely in  $\ell_2$  towards the solution  $\theta^\varepsilon$  of the stochastic convex minimisation problem (2.2), i.e.*

$$(3.2) \quad \lim_{n \rightarrow \infty} \|\hat{\theta}_n - \theta^\varepsilon\|_{\ell_2} = 0 \quad a.s.$$

Equivalently, we also have that

$$(3.3) \quad \lim_{n \rightarrow \infty} \int_{\mathcal{X}} |\hat{u}_\varepsilon^n(x) - u_\varepsilon(x)|^2 d\mu(x) = 0 \quad a.s.$$

Assumption 3.1 can be made more explicit by the choice of a specific sequence of weights  $w = (w_\lambda)_{\lambda \in \Lambda}$  and by imposing regularity assumptions on the function  $u_\varepsilon \in L^1(\mu)$  given



by (2.9). For example, one may assume in dimension  $d = 2$  that  $u_\varepsilon$  is differentiable (with periodic conditions on the boundary on  $\mathcal{X}$ ) and that its gradient is square integrable,

$$\int_{\mathcal{X}} \|\nabla u(x)\|^2 d\mu(x) < +\infty.$$

Then, under such assumptions, one may use the fact that  $\nabla u(x) = \sum_{\lambda \in \Lambda} 2\pi i \lambda \theta_\lambda^\varepsilon \phi_\lambda(x)$  and Parseval's identity, [40][Theorem 1.7], to obtain that

$$\sum_{\lambda \in \Lambda} \|\lambda\|^2 |\theta_\lambda^\varepsilon|^2 < +\infty.$$

Consequently, for the specific choice  $w_\lambda = \|\lambda\|^{-2}$ , we find that Assumption 3.1 holds properly. In higher dimension  $d$ , it is necessary to make additional assumptions on the differentiability of  $u_\varepsilon$ . Note that we shall also prove in Lemma A.2 that for any  $\theta, \tau \in \bar{\ell}_1(\Lambda)$ ,

$$D^2 H_\varepsilon(\theta^\varepsilon)[\tau, \tau] \geq \frac{1}{\varepsilon} \left( 2 - \int_{\mathcal{Y}} \int_{\mathcal{X}} F_{\theta^\varepsilon, y}^2(x) d\mu(x) d\nu(y) \right) \|\tau\|_{\ell_2}^2.$$

Therefore a sufficient condition for Assumption 3.2 to hold is to assume that

$$\int_{\mathcal{Y}} \int_{\mathcal{X}} F_{\theta^\varepsilon, y}^2(x) d\mu(x) d\nu(y) < 2 \quad \text{with} \quad c_\varepsilon = \frac{1}{\varepsilon} \left( 2 - \int_{\mathcal{Y}} \int_{\mathcal{X}} F_{\theta^\varepsilon, y}^2(x) d\mu(x) d\nu(y) \right).$$

#### 4. PROPERTIES OF THE OBJECTIVE FUNCTION $H_\varepsilon$

The purpose of this section is to discuss various keystone properties of the functions  $h_\varepsilon$  and  $H_\varepsilon$  that are needed to establish our main result on the convergence of our stochastic algorithm  $\hat{\theta}_n$ .

Throughout this section, it is assumed that  $\varepsilon > 0$ . Moreover, all the results stated below are valid for any cost function  $c$  that is lower semi-continuous and that belongs to  $L^1(\mu \otimes \nu)$  so that regularized OT is well defined. Consequently, the restriction to the quadratic cost is no longer needed in this section.

Let us first discuss the first and second order Fréchet differentiability of the functions  $H_\varepsilon$  and  $h_\varepsilon$  that are functions from the Banach space  $(\bar{\ell}_1(\Lambda), \|\cdot\|_{\ell_1})$  to  $\mathbb{R}$ . The following proposition gives the expression of the first order Fréchet derivative, that we shall sometimes refer to as the gradient, of  $h_\varepsilon$  and  $H_\varepsilon$ , as well as upper bounds on their operator norm.

**Proposition 4.1.** *For any  $y \in \mathcal{Y}$ , the first order Fréchet derivative of the function  $h_\varepsilon(\cdot, y)$  at  $\theta \in \bar{\ell}_1(\Lambda)$  is the linear operator  $D_\theta h_\varepsilon(\theta, y) : \bar{\ell}_1(\Lambda) \rightarrow \mathbb{R}$  defined for any  $\tau \in \bar{\ell}_1(\Lambda)$  as*

$$(4.1) \quad D_\theta h_\varepsilon(\theta, y)[\tau] = \sum_{\lambda \in \Lambda} \frac{\overline{\partial h_\varepsilon(\theta, y)}}{\partial \theta_\lambda} \tau_\lambda$$

where

$$(4.2) \quad \frac{\partial h_\varepsilon(\theta, y)}{\partial \theta_\lambda} = \int_{\mathcal{X}} \overline{\phi_\lambda(x)} F_{\theta, y}(x) d\mu(x).$$

Moreover, the linear operator  $D_\theta h_\varepsilon(\theta, y)$  can be identified as an element of  $\bar{\ell}_\infty(\Lambda)$  and its operator norm satisfies, for any  $\theta \in \bar{\ell}_1(\Lambda)$  and  $y \in \mathcal{Y}$ ,

$$(4.3) \quad \|D_\theta h_\varepsilon(\theta, y)\|_{op} = \sup_{\|\tau\|_{\ell_1} \leq 1} |D_\theta h_\varepsilon(\theta, y)[\tau]| \leq \sup_{\lambda \in \Lambda} \left| \frac{\partial h_\varepsilon(\theta, y)}{\partial \theta_\lambda} \right| \leq 1.$$

The first order Fréchet derivative of the function  $H_\varepsilon$  at  $\theta \in \bar{\ell}_1(\Lambda)$  is the linear operator  $DH_\varepsilon(\theta) : \bar{\ell}_1(\Lambda) \rightarrow \mathbb{R}$  defined for any  $\tau \in \bar{\ell}_1(\Lambda)$  as

$$(4.4) \quad DH_\varepsilon(\theta)[\tau] = \sum_{\lambda \in \Lambda} \frac{\partial H_\varepsilon(\theta)}{\partial \theta_\lambda} \tau_\lambda$$

where

$$\frac{\partial H_\varepsilon(\theta)}{\partial \theta_\lambda} = \int_{\mathcal{Y}} \frac{\partial h_\varepsilon(\theta, y)}{\partial \theta_\lambda} d\nu(y).$$

Moreover, the operator norm of the linear operator  $DH_\varepsilon(\theta)$  satisfies, for any  $\theta \in \bar{\ell}_1(\Lambda)$ ,

$$(4.5) \quad \|DH_\varepsilon(\theta)\|_{op} = \sup_{\|\tau\|_{\ell_1} \leq 1} |DH_\varepsilon(\theta)[\tau]| \leq \sup_{\lambda \in \Lambda} \left| \frac{\partial H_\varepsilon(\theta)}{\partial \theta_\lambda} \right| \leq 1.$$

The proposition below gives the expression of the second order Fréchet derivative, that we shall sometimes refer to as the Hessian, of  $h_\varepsilon$  and  $H_\varepsilon$  and upper bounds on their operator norm.

**Proposition 4.2.** *For any  $y \in \mathcal{Y}$ , the second order Fréchet derivative of the function  $h_\varepsilon(\cdot, y)$  at  $\theta \in \bar{\ell}_1(\Lambda)$  is the following symmetric bilinear mapping from  $\bar{\ell}_1(\Lambda) \times \bar{\ell}_1(\Lambda)$  to  $\mathbb{R}$*

$$(4.6) \quad \begin{aligned} D^2_{\theta} h_\varepsilon(\theta, y)[\tau, \tau'] &= \frac{1}{\varepsilon} \sum_{\lambda' \in \Lambda} \sum_{\lambda \in \Lambda} \tau'_{\lambda'} \overline{\tau_\lambda} \int_{\mathcal{X}} \phi_{\lambda'}(x) \overline{\phi_\lambda(x)} F_{\theta, y}(x) d\mu(x) \\ &\quad - \frac{1}{\varepsilon} \left( \sum_{\lambda \in \Lambda} \tau'_\lambda \int_{\mathcal{X}} \phi_\lambda(x) F_{\theta, y}(x) d\mu(x) \right) \overline{\left( \sum_{\lambda \in \Lambda} \tau_\lambda \int_{\mathcal{X}} \phi_\lambda(x) F_{\theta, y}(x) d\mu(x) \right)}. \end{aligned}$$

and its operator norm satisfies, for any  $\theta \in \bar{\ell}_1(\Lambda)$  and  $y \in \mathcal{Y}$ ,

$$(4.7) \quad \|D^2_{\theta} h_\varepsilon(\theta, y)\|_{op} = \sup_{\|\tau\|_{\ell_1} \leq 1, \|\tau'\|_{\ell_1} \leq 1} |D^2_{\theta} h_\varepsilon(\theta, y)[\tau, \tau']| \leq \frac{1}{\varepsilon}.$$

Moreover, the second order Fréchet derivative of  $H_\varepsilon : \bar{\ell}_1(\Lambda) \rightarrow \mathbb{R}$  is the symmetric bilinear mapping from  $\bar{\ell}_1(\Lambda) \times \bar{\ell}_1(\Lambda)$  to  $\mathbb{R}$  defined by

$$(4.8) \quad \begin{aligned} D^2 H_\varepsilon(\theta)[\tau, \tau'] &= \frac{1}{\varepsilon} \sum_{\lambda' \in \Lambda} \sum_{\lambda \in \Lambda} \tau'_{\lambda'} \overline{\tau_\lambda} \int_{\mathcal{Y}} \int_{\mathcal{X}} \phi_{\lambda'}(x) \overline{\phi_\lambda(x)} F_{\theta, y}(x) d\mu(x) d\nu(y) \\ &\quad - \frac{1}{\varepsilon} \int_{\mathcal{Y}} \left( \sum_{\lambda \in \Lambda} \tau'_\lambda \int_{\mathcal{X}} \phi_\lambda(x) F_{\theta, y}(x) d\mu(x) \right) \overline{\left( \sum_{\lambda \in \Lambda} \tau_\lambda \int_{\mathcal{X}} \phi_\lambda(x) F_{\theta, y}(x) d\mu(x) \right)} d\nu(y), \end{aligned}$$

and its operator norm satisfies, for any  $\theta \in \bar{\ell}_1(\Lambda)$ ,

$$(4.9) \quad \|D^2 H_\varepsilon(\theta)\|_{op} = \sup_{\|\tau\|_{\ell_1} \leq 1, \|\tau'\|_{\ell_1} \leq 1} |D^2 H_\varepsilon(\theta)[\tau, \tau']| \leq \frac{1}{\varepsilon}.$$

We now provide useful results on the regularity of  $H_\varepsilon$ .

**Proposition 4.3.** *For any  $y \in \mathcal{Y}$ , the functions  $h_\varepsilon(\cdot, y)$  and  $H_\varepsilon$  are strictly convex on  $\bar{\ell}_1(\Lambda)$ .*

As already noticed in previous works [4, 21] dealing with related objective functions, the function  $H_\varepsilon$  is not strongly convex. Nevertheless, one can obtain a local strong convexity property of the function  $H_\varepsilon$  in the neighborhood of its minimizer  $\theta^\varepsilon$ . This result is a consequence of the notion of generalized self-concordance introduced in [2], which has been shown to hold for regularized semi-discrete OT in [4], and which we extend to the setting of the functional  $H_\varepsilon$  on the Banach space  $\bar{\ell}_1(\Lambda)$ .

**Proposition 4.4.** *For all  $\theta \in \bar{\ell}_1(\Lambda)$ , we have*

$$(4.10) \quad H_\varepsilon(\theta) - H_\varepsilon(\theta^\varepsilon) \leq \frac{1}{\varepsilon} \|\theta - \theta^\varepsilon\|_{\ell_1}^2.$$

Moreover, for any  $\theta \in \bar{\ell}_1(\Lambda)$ , the following local strong convexity property holds

$$(4.11) \quad DH_\varepsilon(\theta)[\theta - \theta^\varepsilon] \geq g\left(\frac{2}{\varepsilon} \|\theta - \theta^\varepsilon\|_{\ell_1}\right) D^2 H_\varepsilon(\theta^\varepsilon)[\theta - \theta^\varepsilon, \theta - \theta^\varepsilon],$$

where, for all  $x > 0$ ,

$$(4.12) \quad g(x) = \frac{1 - \exp(-x)}{x}.$$

## 5. NUMERICAL EXPERIMENTS

**5.1. Influence of the dimension  $d$ .** We first investigate the convergence of our numerical scheme for the estimation of the entropic map using various values of the dimension  $d$  to analyse its impact of the computational performances of our approach.

To do so, our estimator  $\hat{Q}_\varepsilon^n(x)$  in (2.11) is compared to  $\tilde{Q}_\varepsilon^n(x)$  in (2.12) where the dual potential  $\tilde{v}_n \in \mathbb{R}^n$  needed to compute  $\tilde{Q}_\varepsilon^n(x)$  is obtained with either the Sinkhorn algorithm [18] or a stochastic algorithm as proposed in [4, 21]. Starting from the uniform distribution on  $[0, 1]^d$ , we consider the map  $Q : x \mapsto L^T Lx + b$  where  $L$  is a lower triangular matrix and  $b \in \mathbb{R}^d$ , both filled with ones. Trivially,  $Q$  is the gradient of a convex function, so that it is the MK quantile function of  $\nu = Q_\# \mu$ . Thus, by Monte-Carlo sampling, we are able to approximate the mean squared error of any estimator  $\hat{Q}$  defined as

$$(5.1) \quad \text{MSE}(\hat{Q}) = \mathbb{E} \left[ \|\hat{Q}(X) - Q(X)\|^2 \right].$$

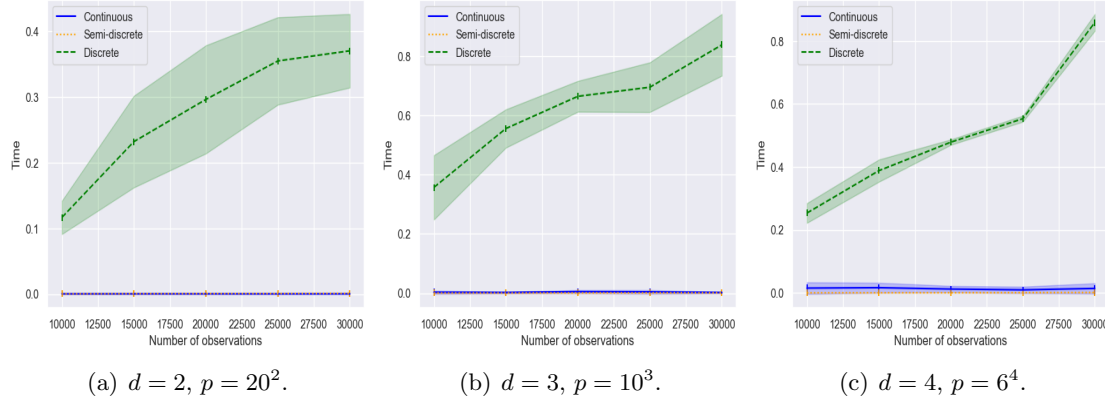
The three ways of estimating  $Q$  are based on iterative schemes that we let running until convergence of the MSE below the value  $10^{-2}$  for  $d = 2, 3, 4$ , and by taking  $\varepsilon = 0.005$ .

Fig. 1 illustrates the time before convergence, in seconds, as a function of  $n$  (the number of observations). In what follows, the continuous, semi-discrete, and discrete approaches refer to Algorithm 1 with  $W$  the identity matrix, the stochastic algorithm from [4, 21], and the Sinkhorn algorithm [18] respectively. For  $\mathcal{X}_p$  given in Algorithm 1, the uniform distribution on  $\mathcal{X}_p$  is taken as a discrete reference measure for the Sinkhorn algorithm to ensure a fair comparison with our algorithm. The MSE is estimated through  $m = 500$  other random samples from  $\mu$ . The size  $p$  of the grid  $\mathcal{X}_p$  is maintained comparable in every considered dimensions. Results are averaged over 10 experiments for several samples  $(Y_1, \dots, Y_n)$ , and standard deviation is indicated around each MSE curve. Overall, these numerical experiments reveal a potentially faster convergence for approaches based on stochastic algorithms when the number of observations grows. Moreover, our continuous approach slightly outperforms the semi-discrete one in term of computational performances.

**5.2. Numerical experiments in dimension  $d = 1$ .** The univariate setting allows us an explicit knowledge of the ground truth  $Q$ . There, we study our algorithm with either the standard quadratic cost in  $\mathbb{R}^d$  given by  $c(x, y) = \frac{1}{2} \|x - y\|^2$  or the quadratic cost on the flat torus  $\mathbb{T}^d = \mathbb{R}^d / \mathbb{Z}^d$  that is

$$(5.2) \quad c(x, y) = \frac{1}{2} d_{\mathbb{T}^d}(x, y), \quad \text{with} \quad d_{\mathbb{T}^d}(x, y) = \min_{\lambda \in \mathbb{Z}^d} \|x - y + \lambda\|.$$

The choice of the quadratic cost on the torus is motivated by the discussion in the supplementary material Appendix D on sufficient conditions related to the summability of the Fourier coefficients of an optimal dual potential.



**Figure 1.** Overall time, in seconds, until convergence of the MSE below  $10^{-2}$  for different solvers for EOT.

For the learning rate  $\gamma_n = \gamma n^{-c}$ , we took  $\gamma = \varepsilon$  and  $c = 3/4$ . The sequence of weights  $w = (w_\lambda)_{\lambda \in \Lambda}$  is chosen as  $w_\lambda = |\lambda|^{-2}$  for  $\lambda \in \mathbb{Z} \setminus \{0\}$ . Taking a larger exposant than 2 results in smoother estimators of the optimal dual potential  $u_\varepsilon$ . For various values of  $\varepsilon \in [0.005, 0.5]$ , we consider a beta( $a, b$ ) distribution  $\nu$  on  $\mathcal{Y} = [0, 1]$  with parameters  $a = 5$  and  $b = 5$ . The optimal dual potential  $u_0$  and quantile function  $Q_0$  are straightforward to compute when  $d = 1$  for the standard quadratic cost. For a sample of size  $n = 10^5$ ,  $\hat{u}_\varepsilon^n$  and  $\hat{Q}_\varepsilon^n$  are displayed in Fig. 2 using either the standard quadratic cost or the quadratic cost of the torus. One can observe that the choice of the cost yields a different regularization effect. Choosing  $\varepsilon = 0.005$  yields values of  $\hat{u}_\varepsilon^n$  and  $\hat{Q}_\varepsilon^n$  that are very close to  $u_0$  and  $Q_0$  respectively.

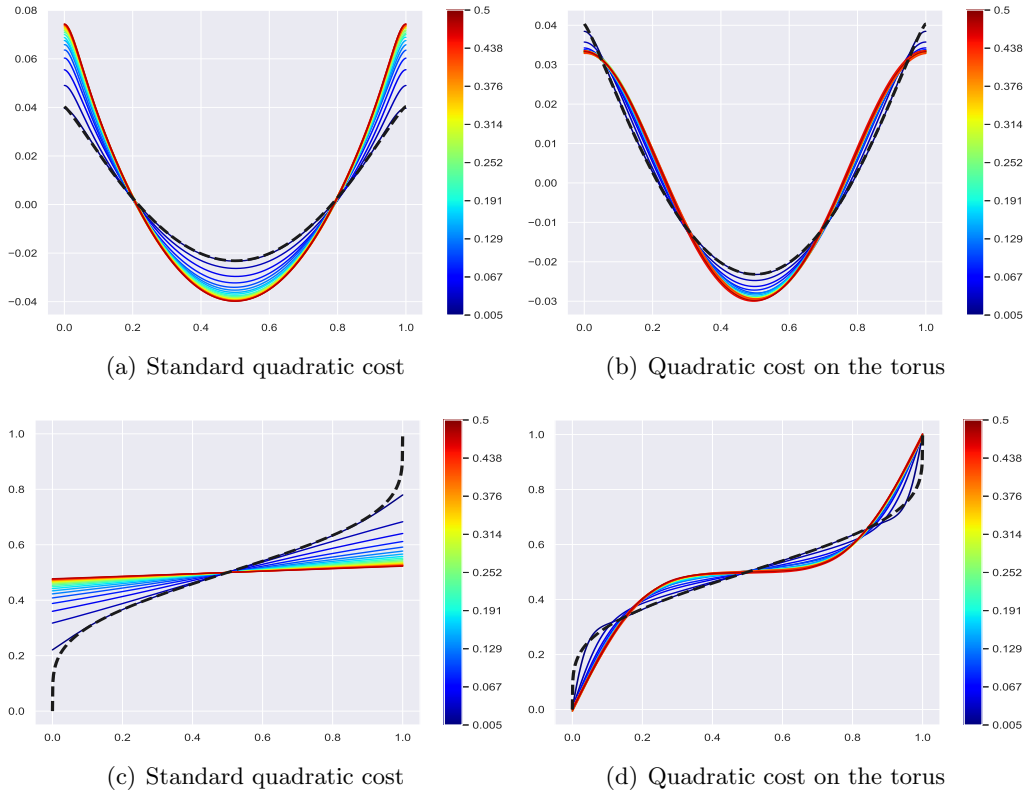
From now on, let us consider a sample  $(Y_1^*, \dots, Y_J^*)$  of small size  $J = 100$  of the same beta( $a, b$ ) distribution. We illustrate the potential benefits of using regularized OT to obtain a smoother estimator than the usual empirical quantile function  $\hat{Q}_0^J$  defined as the generalized inverse of the empirical cumulative distribution function

$$\hat{F}_0^J(x) = \frac{1}{J} \sum_{j=1}^J \mathbb{1}_{\{Y_j^* \leq x\}}.$$

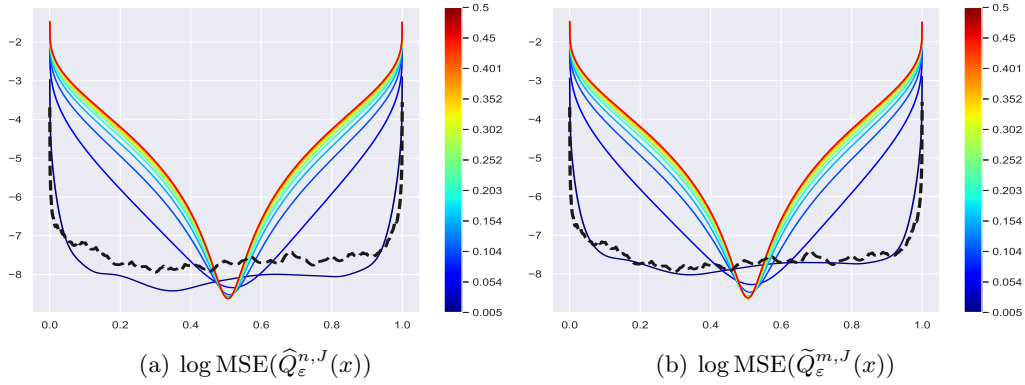
To this end, for various values of  $\varepsilon \in [0.005, 0.5]$ , we compute the two estimators  $\hat{Q}_\varepsilon^{n,J}$  from (2.11) and  $\tilde{Q}_\varepsilon^{m,J}$  from (2.12) with sequences of  $n = m = 10^5$  random variables sampled from the discrete measure  $\hat{\nu}_J^*$  or the uniform measure on  $[0, 1]$  respectively. In Fig. 3, we display in logarithmic scale the point-wise mean-squared errors

$$\text{MSE}(\hat{Q}_\varepsilon^{n,J}(x)) = \mathbb{E} \left[ |\hat{Q}_\varepsilon^{n,J}(x) - Q_0(x)|^2 \right] \quad \text{and} \quad \text{MSE}(\tilde{Q}_\varepsilon^{m,J}(x)) = \mathbb{E} \left[ |\tilde{Q}_\varepsilon^{m,J}(x) - Q_0(x)|^2 \right],$$

where the above expectations are approximated using Monte-Carlo experiments from 100 repetitions of the above described procedure. The MSE of these regularized estimators is then compared to the MSE of the usual empirical quantile function  $\hat{Q}_0^J$  defined accordingly. For all values of  $\varepsilon$ , it can be seen, from Fig. 3, that regularization always improves the estimation of  $Q_0(x)$  by  $\hat{Q}_0^J(x)$  around the median location  $x = 0.5$ . For the smallest values of  $\varepsilon$ , regularization also improves the estimation of  $Q_0(x)$  for  $x \in [0.1, 0.9]$ , and the best results are obtained with the stochastic algorithm based on the FFT.



**Figure 2.** Estimators  $\hat{u}_\varepsilon^n$  and  $\hat{Q}_\varepsilon^n$  on the first and second lines respectively. The black and dashed curves are either the un-regularized optimal dual potential  $u_0$  or the un-regularized quantile function  $Q_0$  of the beta distribution for the standard quadratic cost.



**Figure 3.** Point-wise error of the regularized estimators  $\hat{Q}_\varepsilon^{n,J}$  and  $\tilde{Q}_\varepsilon^{m,J}$  for various values of  $\varepsilon \in [0.005, 0.5]$ . The black and dashed curve is the point-wise error of the un-regularized empirical quantile function  $\hat{Q}_0^J$ .

**5.3. Numerical experiments in dimension  $d = 2$ .** As argued in [24], taking as reference the spherical uniform distribution  $\mu_S$  on the unit ball  $\mathbb{B}^d$  induces different properties for MK quantiles. Thanks to a change in polar coordinates, one can parametrize on  $\mathbb{B}^d$  instead of  $[0, 1]^d$ . By definition, a random vector  $X$  with spherical uniform distribution is given by  $X = R\Phi$  where  $R$  and  $\Phi$  are independent and drawn uniformly from  $[0, 1]$  and the unit hypersphere  $\mathbb{S}^{d-1}$ , respectively. In dimension  $d = 2$ ,  $X$  writes in polar coordinates as

$$X = \begin{pmatrix} R \cos(2\pi\Psi) \\ R \sin(2\pi\Psi) \end{pmatrix} \in \mathbb{B}^2,$$

where  $(R, \Psi)$  is uniform on  $[0, 1]^2$ . Then, for a function  $u \in L^1(\mathbb{B}^d, \mu_S)$ , its parametrization in polar coordinates is given, for all  $(r, \psi) \in [0, 1] \times [0, 1]$ , by

$$\bar{u}(r, \psi) = u \begin{pmatrix} r \cos(2\pi\psi) \\ r \sin(2\pi\psi) \end{pmatrix}.$$

Hence, by definition of  $\mu_S$ , the function  $\bar{u}$  is an element of  $L^1([0, 1]^2, \mu)$  where  $\mu$  is the uniform distribution on  $\mathcal{X} = [0, 1]^2$ . Consequently, thanks to this re-parametrization, we propose to solve in the Fourier domain, for  $\Lambda = \mathbb{Z}^2 \setminus \{0\}$ , the following regularized OT problem

$$(5.3) \quad \theta^\varepsilon = \underset{\theta \in \ell_1(\Lambda)}{\operatorname{argmin}} \bar{H}_\varepsilon(\theta) \quad \text{with} \quad \bar{H}_\varepsilon(\theta) = \mathbb{E} [\bar{h}_\varepsilon(\theta, Y)],$$

where  $Y = (Y_1, Y_2) \in \mathbb{R}^2$  is a random vector with distribution  $\nu$ , and  $\bar{h}_\varepsilon$  is given by

$$\bar{h}_\varepsilon(\theta, y) = \varepsilon \log \left( \int_{\mathcal{X}} \exp \left( \frac{\sum_{\lambda \in \Lambda} \theta_\lambda \phi_\lambda(r, \psi) - c_y(r, \psi)}{\varepsilon} \right) d\mu(r, \psi) \right) + \varepsilon,$$

with  $\phi_\lambda(r, \psi) = e^{2\pi i(\lambda_1 r + \lambda_2 \psi)}$  for  $\lambda = (\lambda_1, \lambda_2) \in \mathbb{Z}^2$ , and  $c_y$  refers to the quadratic cost,

$$c_y(r, \psi) = \frac{1}{2} ((r \cos(2\pi\psi) - y_1)^2 + (r \sin(2\pi\psi) - y_2)^2).$$

In order to solve (5.3), we adapt the stochastic algorithm (2.5) which yields, after  $n$  iterations, the sequence  $\bar{\theta}_n$  and the estimator, in polar coordinates,

$$(5.4) \quad \bar{u}_\varepsilon^n(r, \psi) = \sum_{\lambda \in \Lambda} \bar{\theta}_{n,\lambda} \phi_\lambda(r, \psi).$$

In practice, we discretize  $[0, 1]^2$  by choosing equi-spaced radius points  $0 \leq r_1 < \dots < r_{p_1} \leq 1$  and angles  $0 \leq \psi_1 < \dots < \psi_{p_2} < 1$  which results in taking a grid of  $p = p_1 p_2$  points

$$\mathcal{X}_p = \{(r_{\ell_1}, \psi_{\ell_2})_{(\ell_1, \ell_2) \in \{1, p_1\} \times \{1, p_2\}}\} \subset [0, 1]^2.$$

Finally, the stochastic algorithm (2.5) is implemented on this polar grid using the weight sequence  $w_\lambda = 1$  for all  $\lambda \in \Lambda_p$  that is with  $\alpha = 0$ . Of course, Assumption 3.1 is always verified if  $(w_\lambda) \equiv (1, 1, \dots, 1, 0, 0, \dots)$ . This is motivated by the fact that choosing  $w_\lambda = \|\lambda\|^{-\alpha}$  with  $\alpha \geq 1$  would impose periodic constraints on the dual potentials  $\bar{u}(r, \psi)$  along the radius coordinate. However, as shown by the following numerical experiments, an optimal dual potential typically does not satisfy the polar periodic conditions  $\bar{u}(0, \psi) = \bar{u}(1, \psi)$  for all  $\psi \in [0, 1]$ . The counterpart of  $\hat{Q}_\varepsilon^n$  in (2.11) directly follows from (5.4), that is

$$(5.5) \quad \bar{Q}_\varepsilon^n(x) = \sum_{j=1}^n \bar{F}_j(x) Y_j \quad \text{where} \quad \bar{F}_j(x) = \frac{\exp \left( \frac{(\bar{u}_\varepsilon^n)^{c, \varepsilon}(Y_j) - c(x, Y_j)}{\varepsilon} \right)}{\sum_{\ell=1}^n \exp \left( \frac{(\bar{u}_\varepsilon^n)^{c, \varepsilon}(Y_\ell) - c(x, Y_\ell)}{\varepsilon} \right)},$$

where the integral in the computation  $(\bar{u}_\varepsilon^n)^{c,\varepsilon}(\cdot)$  is approximated with the polar grid  $\mathcal{X}_p$ . In what follows, we report numerical experiments for the banana-shaped distribution  $\nu$  considered in [12]. It corresponds to sampling  $Y$  as the random vector

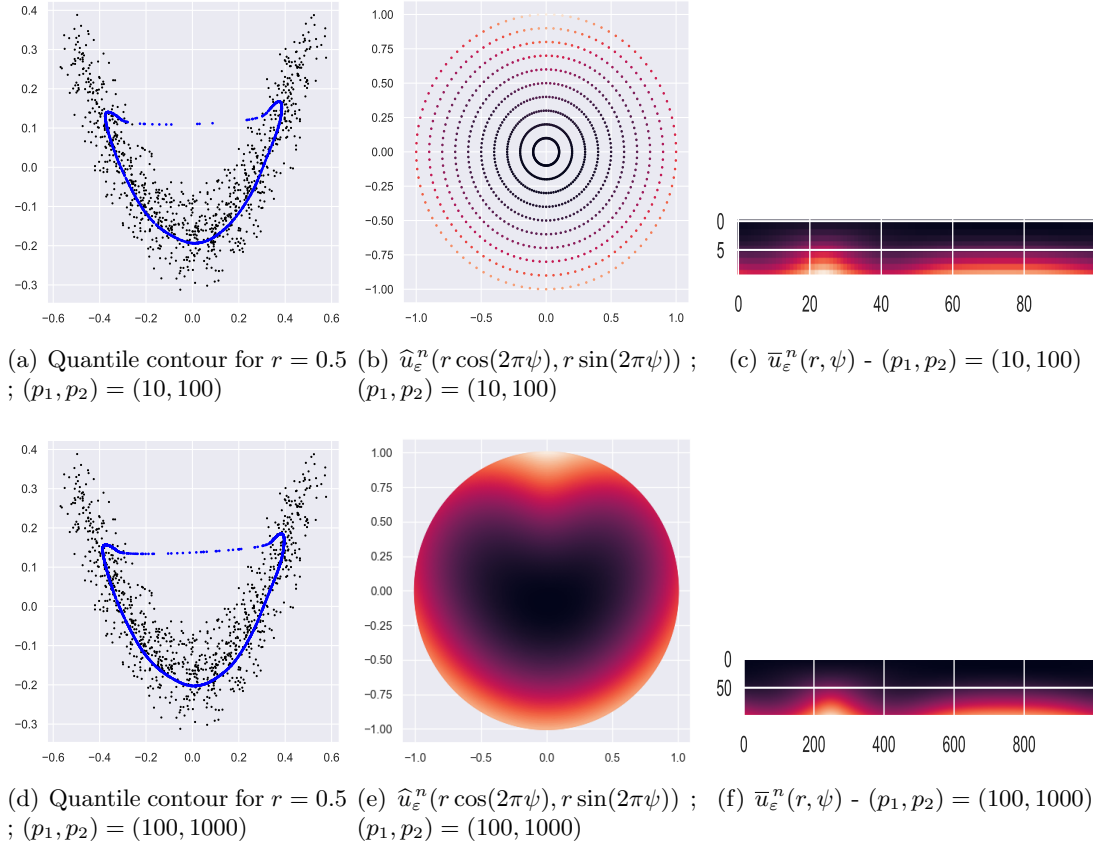
$$Y = \begin{pmatrix} U + R \cos(2\pi\Phi) \\ U^2 + R \sin(2\pi\Phi) \end{pmatrix},$$

where  $U$  is uniform on  $[-1, 1]$ ,  $\Phi$  is uniform on  $[0, 1]$ ,  $R = 0.2Z(1 - (1 - |U|)/2)$  with  $Z$  uniform on  $[0, 1]$ , and  $U, \Phi$  and  $Z$  independent. In these simulations, the random variable  $Y$  is also centered and scaled so that it takes its values within the subset  $[-0.6, 0.6] \times [-0.4, 0.5] \subset [0, 1]^2$ .

We first consider a sample  $Y_1^*, \dots, Y_J^*$  of size  $J = 10^3$  that is held fixed and displayed in Fig. 4. Then, we draw  $n = 10^5$  random variables  $Y_1, \dots, Y_n$  from the associated discrete distribution  $\hat{\nu}_J^*$ , and we run the stochastic algorithm (2.5) with different sizes  $(p_1, p_2) = (10, 100)$  and  $(p_1, p_2) = (100, 1000)$  for the discretization  $\mathcal{X}_p$ . Note that the cost of each iteration of the stochastic algorithm is of order  $\mathcal{O}(p \log(p))$  for  $p = p_1 p_2$ . Therefore, the choice of discretization of the polar coordinates greatly influences the computational cost of the algorithm. In Fig. 4, we display the resulting regularized dual potentials  $\hat{u}_\varepsilon^n$  in cartesian coordinates for  $\varepsilon = 0.005$ . We also draw the resulting MK contour quantiles of level  $r = 0.5$  for each choice of discretization. It can be seen that the resulting MK contour quantiles are very similar with a much lowest computational cost for the discretization of size  $(p_1, p_2) = (10, 100)$ .

Fig. 5 contains a comparison of the convergence between our FFT-based scheme (2.11) and (2.12), based on the stochastic gradient descent from [4], that we refer to as the regularized SGD. The reference distribution is taken to be the spherical uniform. Also, we compare these regularized approaches (using  $\varepsilon = 0.005$ ) with classical un-regularized ones. To this end, we implement a subgradient descent for un-regularized OT, namely the same Robbins-Monro scheme as (1.9) with  $\varepsilon = 0$ , that is a semi-discrete scheme advocated in [12][Section 4]. Finally, we use the OT network simplex solver from the Python library [20] to compute the solution of un-regularized OT between two empirical discrete distributions with supports  $\mathcal{X}_p = \{x_1, \dots, x_p\}$  and  $(Y_1^*, \dots, Y_J^*)$ . We first consider a sample  $Y_1^*, \dots, Y_J^*$  of size  $J = 10^4$  that is held fixed. For our FFT approach, we let  $p_1 = 20$ ,  $p_2 = 500$ , so that  $p = p_1 p_2 = 10^4$ . For the three iterative schemes, the number of iterations varies between  $10^4$ ,  $10^5$  and  $10^6$ . This corresponds, for our FFT approach, to a stochastic algorithm with 1, 10 and 100 epochs, whereas the other approaches sample from the reference distribution  $\mu_S$ . The first line of Fig. 5 contains the corresponding quantile contours of order  $r = 0.5$  for each of these methods, for several number of iterations. The colored dots are obtained by transporting points of radius  $r = 0.5$ , while the lines between them are visual artefacts. Unlike regularized estimators, the quantile function estimated from an un-regularized semi-discrete scheme is restricted to take its values in the set of observations  $(Y_1^*, \dots, Y_J^*)$ . On another hand, with the simplex solver, the obtained empirical quantile map is not a function, rather a collection of points. The use of stochastic algorithms is more targeted to this task. Still, it is represented here as a benchmark, indicating where the quantile contours shall be. Furthermore, the second line of Fig. 5 deals with convergence depending on the number of iterations. As customary, we consider a recursive estimation of the values of our objectives, respectively  $H_\varepsilon$  for (2.2),  $\tilde{H}_\varepsilon$  for (1.9) and  $\tilde{H}_0$  for (1.9) with  $\varepsilon = 0$ . These objectives are recursively estimated along the iterations by gradual averaging in order to account for convergence, as proposed in [4]. For  $J = p = 10^4$ , the computational cost at each iteration of the two regularized procedures is

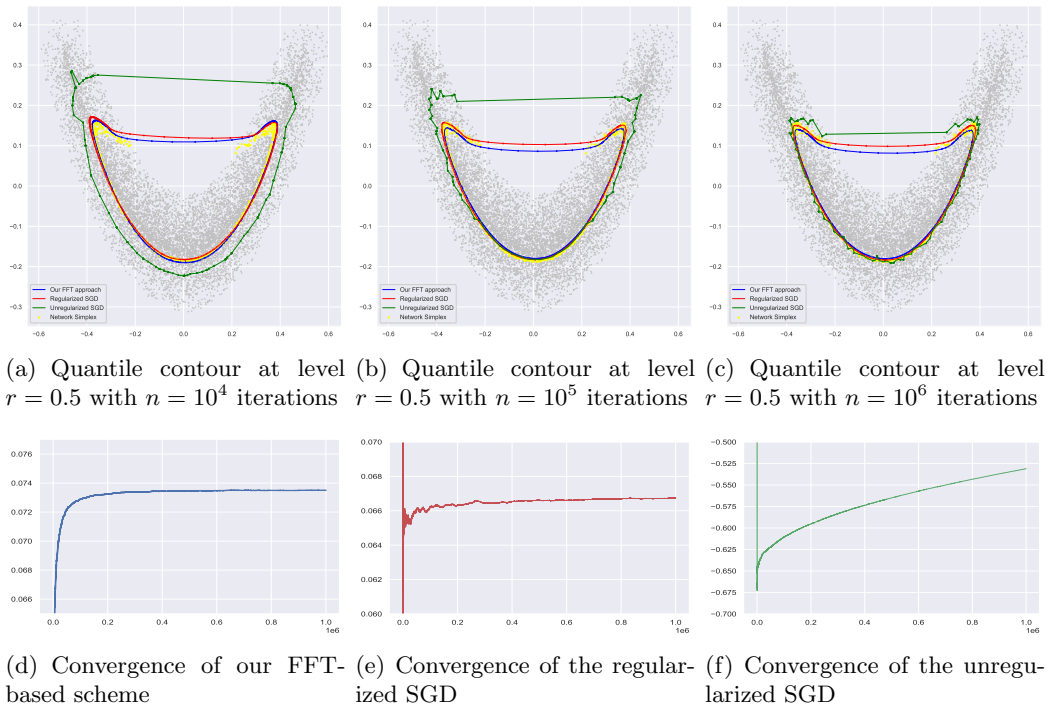




**Figure 4.** The blue curves are regularized MK quantile contours at level  $r = 0.5$  for  $\varepsilon = 0.005$  from the discrete measure  $\hat{\nu}_J$  (displayed with black points) using two different discretizations  $(p_1, p_2) = (10, 100)$  (first row) and  $(p_1, p_2) = (100, 1000)$  (second row). The second (resp. third) columns represent the values of the regularized dual potentials in cartesian (resp. polar) coordinates for each choice of discretization.

of the same order. It can be seen that the un-regularized SGD has not converged with  $10^6$  iterations, whereas the regularized approaches (2.11) and (2.12) have similar convergence behavior. Together with the first line of Fig. 5, these results illustrate that entropically regularized methods converge faster towards a more suitable solution.

We finally propose a last numerical experiment to highlight the behavior of EOT when varying the regularization parameter  $\varepsilon$ . We chose to draw  $n = 10^7$  random variables  $Y_1, \dots, Y_n$  from the banana-shaped distribution, and we ran the stochastic algorithm (2.5) for the discretization  $(p_1, p_2) = (10, 1000)$ . Doing so, the obtained sample is very close to the true density, and the various resulting contours only depend on  $\varepsilon$ . In Fig. 6, we display the resulting regularized MK quantile contours of levels  $r \in \{0.2, 0.3, \dots, 1\}$  for different values of  $\varepsilon \in [0.002, 0.5]$ . This visualization warns on the choice of the regularization parameter that must be chosen small enough, as usual with EOT. Note that, for  $n = 10^7$  observations, we have not been able to implement the Sinkhorn algorithm. Moreover,



**Figure 5.** Comparison between regularized (with  $\varepsilon = 0.005$ ) and unregularized approaches.

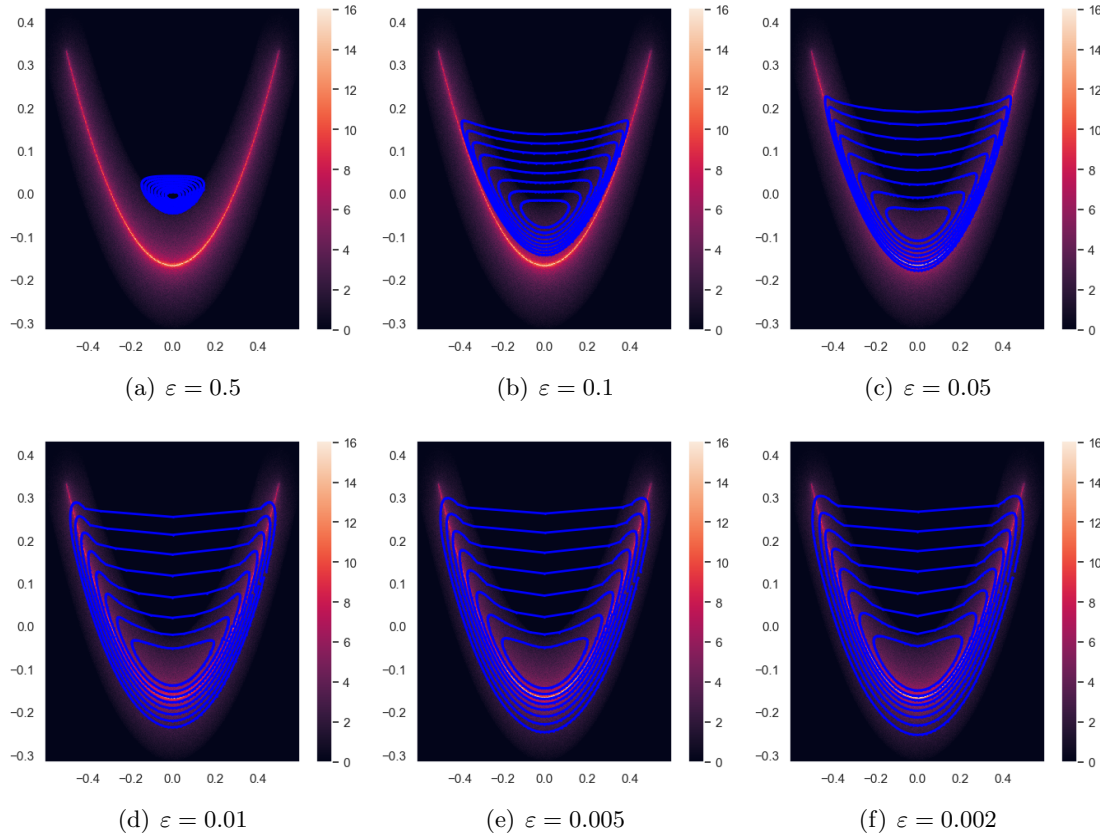
the cost at each iteration of either regularized or un-regularized SGD being  $\mathcal{O}(n)$ , these algorithms are much slower to converge than our approach.

## 6. CONCLUSION AND PERSPECTIVES

Throughout the paper, we advocated the use of the entropic map for MK quantiles' estimation. Indeed, it is a smooth approximation of an OT map and benefits from the crucial cyclical monotonicity together with computational benefits of EOT. Our new stochastic algorithm for the continuous OT problem showed potential improvement in terms of numerical complexity, because it is independent, at each iteration, from the size of the observed sample. Nonetheless, our implementation of the FFT may become intractable in high dimensions. Because of the known decay of Fourier coefficients, one can hope that more sophisticated FFTs could alleviate this, see *e.g.* [37], but this is beyond the scope of the present paper.

Our convergence study based on random iterative schemes extends results from [4] to the continuous setting instead of the semi-discrete setting. Minimax convergence rates of un-regularized estimators of OT maps have been obtained in recent works [23, 25]. Hence, it would be interesting to extend our analysis to the study of the rate of convergence of our regularized estimator. This is an interesting challenge that is left for future work.

As argued *e.g.* in [24], our assumption of finite second-order moment for  $\nu$  may be too restrictive for multivariate quantiles. In the seminal paper [24], using McCann's theorem [33], the definition of Monge-Kantorovich quantiles have been extended as a push-forward map between the reference and the target measures, that is also the gradient of a convex function. In order to get rid of this moment assumption using EOT, future work may consider the insightful results from [22], as their notion of *cyclically invariant* coupling



**Figure 6.** In all the figures, the image at the background represents a density histogram from the empirical measure  $\hat{\nu}_n = \frac{1}{n} \sum_{j=1}^n \delta_{Y_j}$  where  $Y_1, \dots, Y_n$  are sampled from the banana-shaped distribution with  $n = 10^7$ . The blue curves correspond to regularized MK quantile contours of levels  $r \in \{0.2, 0.3, \dots, 1\}$  for  $\varepsilon \in [0.002, 0.5]$ .

can always yield a mapping by barycentric projection, which coincides with  $Q_\varepsilon$  if the cost  $c$  belongs to  $L^1(\mu \otimes \nu)$ .

## APPENDIX A. PROOFS OF THE MAIN RESULTS

The proofs of Proposition 4.1, Proposition 4.2 and Proposition 4.3 are given in supplementary materials, see Appendix A, Appendix B and Appendix C. We shall now proceed to the proofs of the main results of the paper.

**A.1. Proof of Proposition 4.4.** For  $\theta \in \overline{\ell_1}(\Lambda)$  and  $t \in [0, 1]$ , we denote  $\theta_t = \theta^\varepsilon + t(\theta - \theta^\varepsilon)$  and we define the function  $\varphi(t) = H_\varepsilon(\theta_t)$ . Then, we deduce from a second order Taylor expansion of  $\varphi$  with integral remainder that

$$(A.1) \quad \varphi(1) = \varphi(0) + \varphi'(0) + \int_0^1 (1-t)\varphi''(t)dt.$$

However, we clearly have

$$(A.2) \quad \varphi'(t) = DH_\varepsilon(\theta_t)[\theta - \theta^\varepsilon] \quad \text{and} \quad \varphi''(t) = D^2H_\varepsilon(\theta_t)[\theta - \theta^\varepsilon, \theta - \theta^\varepsilon].$$

Consequently, as  $\varphi'(0) = DH_\varepsilon(\theta^\varepsilon)[\theta - \theta^\varepsilon] = 0$ , (A.1) can be rewritten as

$$(A.3) \quad H_\varepsilon(\theta) - H_\varepsilon(\theta^\varepsilon) = \int_0^1 (1-t) D^2 H_\varepsilon(\theta_t) [\theta - \theta^\varepsilon, \theta - \theta^\varepsilon] dt.$$

Therefore, (4.10) immediately follows from (4.9) and (A.3). It only remains to prove (4.11). Our strategy is to adapt to the setting of this paper the notion of self-concordance as introduced in [1, 2] and used in [4, 5] to study the statistical properties of stochastic optimal transport.

**Lemma A.1.** *For  $\theta \in \overline{\ell_1}(\Lambda)$  and for all  $0 < t < 1$ , denote  $\theta_t = \theta^\varepsilon + t(\theta - \theta^\varepsilon)$ . Then, the function  $\varphi(t) = H_\varepsilon(\theta_t)$  verifies the self-concordance property*

$$(A.4) \quad |\varphi'''(t)| \leq \frac{2}{\varepsilon} \|\theta - \theta^\varepsilon\|_{\ell_1} \varphi''(t).$$

*Proof.* For a fixed  $y \in \mathcal{Y}$ , let  $\phi(t) = h_\varepsilon(\theta_t, y)$ . Firstly, we show that  $\phi(t)$  verifies the self-concordance property. From the chain rule, we obtain that

$$\begin{aligned} \phi'(t) &= Dh_\varepsilon(\theta_t, y)[\theta - \theta^\varepsilon], \\ \phi''(t) &= D^2 h_\varepsilon(\theta_t, y)[\theta - \theta^\varepsilon, \theta - \theta^\varepsilon], \\ \phi'''(t) &= D^3 h_\varepsilon(\theta_t, y)[\theta - \theta^\varepsilon, \theta - \theta^\varepsilon, \theta - \theta^\varepsilon], \end{aligned}$$

where  $D^3 h_\varepsilon$  denotes the third order Fréchet derivative of  $h_\varepsilon(\cdot, y)$ . It follows from (4.1) that

$$(A.5) \quad \phi'(t) = \int_{\mathcal{X}} S(x) F_{\theta_t, y}(x) d\mu(x) \quad \text{where} \quad S(x) = \sum_{\lambda \in \Lambda} (\theta_\lambda - \theta_\lambda^\varepsilon) \phi_\lambda(x).$$

Similarly, (4.6) yields

$$(A.6) \quad \varepsilon \phi''(t) = \int_{\mathcal{X}} S(x)^2 F_{\theta_t, y}(x) d\mu(x) - \left( \int_{\mathcal{X}} S(x) F_{\theta_t, y}(x) d\mu(x) \right)^2.$$

Hereafter, denoting by  $Z_t$  the random variable with density  $F_{\theta_t, y}$  with respect to  $\mu$ , it appears that  $\varepsilon \phi''(t) = \mathbb{E}[S(Z_t)^2] - \mathbb{E}[S(Z_t)]^2 = \mathbb{E}[(S(Z_t) - \mathbb{E}[S(Z_t)])^2]$ , that is

$$(A.7) \quad \varepsilon \phi''(t) = \int_{\mathcal{X}} \left( S(x) - \int_{\mathcal{X}} S(z) F_{\theta_t, y}(z) d\mu(z) \right)^2 F_{\theta_t, y}(x) d\mu(x).$$

Furthermore, using (B.1) in the derivation of (A.6), we have that

$$\varepsilon \phi'''(t) = \int_{\mathcal{X}} S(x)^2 \frac{d}{dt} F_{\theta_t, y}(x) d\mu(x) - 2 \left( \int_{\mathcal{X}} S(x) \frac{d}{dt} F_{\theta_t, y}(x) d\mu(x) \right) \int_{\mathcal{X}} S(x) F_{\theta_t, y}(x) d\mu(x),$$

which yields

$$\begin{aligned} \varepsilon^2 \phi'''(t) &= \int S^3(x) F_{\theta_t, y}(x) d\mu(x) - \left( \int S^2(x) F_{\theta_t, y}(x) d\mu(x) \right) \left( \int S(x) F_{\theta_t, y}(x) d\mu(x) \right) \\ &\quad - 2 \int S(x) F_{\theta_t, y}(x) d\mu(x) \left[ \int S^2(x) F_{\theta_t, y}(x) d\mu(x) - \left( \int S(x) F_{\theta_t, y}(x) d\mu(x) \right)^2 \right]. \end{aligned}$$

Consequently,

$$\varepsilon^2 \phi'''(t) = m_3 - m_2 m_1 - 2 m_1 (m_2 - m_1^2) = m_3 - 3 m_2 m_1 + 2 m_1^3,$$

where  $m_i$  stands for the  $i$ -th moment of the distribution of  $S(Z_t)$ . Then, one recognizes the formula for the cumulant of order 3 of a random variable, and so the above equality can be factorized as

$$(A.8) \quad \varepsilon^2 \phi'''(t) = \mathbb{E}[(S(Z_t) - m_1)^3] = \int (S(x) - m_1)^3 F_{\theta_t, y}(x) d\mu(x).$$

Thanks to the connection between  $\varepsilon \phi''(t)$  and the variance term in (A.7), (A.8) leads to

$$|\varepsilon \phi'''(t)| \leq \sup_{x \in \mathcal{X}} |S(x) - m_1| \phi''(t).$$

It is easy to see that  $|S(x) - m_1| \leq |S(x)| + |m_1| \leq 2\|\theta - \theta^\varepsilon\|_{\ell_1}$ . Hence

$$(A.9) \quad |\phi'''(t)| \leq \frac{2}{\varepsilon} \|\theta - \theta^\varepsilon\|_{\ell_1} \phi''(t).$$

Finally, given that  $\varphi(t) = H_\varepsilon(\theta_t) = \int_{\mathcal{Y}} h_\varepsilon(\theta_t, y) d\nu(y) = \int_{\mathcal{Y}} \phi(t) d\nu(y)$ , (A.9) induces the self-concordance property of  $\varphi$ .  $\square$

We are now in a position to prove inequality (4.11). Denote  $\delta = 2\|\theta - \theta^\varepsilon\|_{\ell_1}/\varepsilon$ . It follows from inequality (A.4) that, for all  $0 < t < 1$ ,  $|\varphi'''(t)| \leq \delta \varphi''(t)$ , which leads to  $\frac{\varphi'''(t)}{\varphi''(t)} \geq -\delta$ . By integrating the above inequality between 0 and  $t$ , we obtain that  $\log \varphi''(t) - \log \varphi''(0) \geq -\delta t$ , which means that  $\frac{\varphi''(t)}{\varphi''(0)} \geq e^{-\delta t}$ . Integrating once again the previous inequality between 0 and 1, we obtain that

$$(A.10) \quad \varphi'(1) - \varphi'(0) \geq \left( \frac{1 - e^{-\delta}}{\delta} \right) \varphi''(0).$$

Finally, as  $\varphi'(1) = DH_\varepsilon(\theta)(\theta - \theta^\varepsilon)$ ,  $\varphi'(0) = 0$  and  $\varphi''(0) = D^2 H_\varepsilon(\theta^\varepsilon)[\theta - \theta^\varepsilon, \theta - \theta^\varepsilon]$ , inequality (4.11) holds, which completes the proof of Proposition 4.4.  $\square$

## A.2. A sufficient condition for Assumption 3.2.

**Lemma A.2.** *For any  $\tau \in \overline{\ell_1}(\Lambda)$ ,*

$$(A.11) \quad D^2 H_\varepsilon(\theta^\varepsilon)[\tau, \tau] \geq \frac{1}{\varepsilon} \left( 2 - \int_{\mathcal{Y}} \int_{\mathcal{X}} F_{\theta^\varepsilon, y}^2(x) d\mu(x) d\nu(y) \right) \|\tau\|_{\ell_2}^2.$$

*Proof.* We already saw from (4.8) that for any  $\tau \in \overline{\ell_1}(\Lambda)$ ,

$$(A.12) \quad \begin{aligned} D^2 H_\varepsilon(\theta^\varepsilon)[\tau, \tau] &= \frac{1}{\varepsilon} \sum_{\lambda' \in \Lambda} \sum_{\lambda \in \Lambda} \tau_{\lambda'} \overline{\tau_\lambda} \int_{\mathcal{Y}} \int_{\mathcal{X}} \phi_{\lambda'}(x) \overline{\phi_\lambda(x)} F_{\theta^\varepsilon, y}(x) d\mu(x) d\nu(y) \\ &- \frac{1}{\varepsilon} \int_{\mathcal{Y}} \left| \sum_{\lambda \in \Lambda} \tau_\lambda \int_{\mathcal{X}} \phi_\lambda(x) F_{\theta^\varepsilon, y}(x) d\mu(x) \right|^2 d\nu(y). \end{aligned}$$

Our proof consists in a study of the two terms in the right-hand side of (A.12). Since (4.4) only defines  $DH_\varepsilon(\theta^\varepsilon)_\lambda$  for all  $\lambda \neq 0$ , we deduce from (4.2) and (4.4) that, for all  $\lambda \neq \lambda'$ ,

$$\int_{\mathcal{X}} \phi_{\lambda'}(x) \overline{\phi_\lambda(x)} F_{\theta^\varepsilon, y}(x) d\mu(x) d\nu(y) = \int_{\mathcal{Y}} \int_{\mathcal{X}} \phi_{\lambda' - \lambda}(x) F_{\theta^\varepsilon, y}(x) d\mu(x) d\nu(y) = DH_\varepsilon(\theta^\varepsilon)_{\lambda' - \lambda}.$$

Moreover, as soon as  $\lambda = \lambda'$ ,

$$\int_{\mathcal{Y}} \int_{\mathcal{X}} \phi_\lambda(x) \overline{\phi_\lambda(x)} F_{\theta^\varepsilon, y}(x) d\mu(x) d\nu(y) = \int_{\mathcal{Y}} \int_{\mathcal{X}} F_{\theta^\varepsilon, y}(x) d\mu(x) d\nu(y) = 1.$$

Hence, from the optimality condition  $DH_\varepsilon(\theta^\varepsilon) = 0$ , we obtain that

$$\int_{\mathcal{Y}} \int_{\mathcal{X}} \phi_{\lambda'}(x) \overline{\phi_\lambda(x)} F_{\theta^\varepsilon, y}(x) d\mu(x) d\nu(y) = \delta_0(\lambda' - \lambda),$$

where  $\delta_0$  stands for the dirac function at 0. Therefore, it follows that

$$(A.13) \quad \frac{1}{\varepsilon} \sum_{\lambda' \in \Lambda} \sum_{\lambda \in \Lambda} \tau_{\lambda'} \overline{\tau_\lambda} \int_{\mathcal{Y}} \int_{\mathcal{X}} \phi_{\lambda'}(x) \overline{\phi_\lambda(x)} F_{\theta^\varepsilon, y}(x) d\mu(x) d\nu(y) = \frac{1}{\varepsilon} \|\tau\|_{\ell_2}^2.$$

From now on, our goal is to find an upper bound for the second term in the right-hand side of (A.12). By Cauchy-Schwarz's inequality, we have that

$$(A.14) \quad \left| \sum_{\lambda \in \Lambda} \tau_\lambda \int_{\mathcal{X}} \phi_\lambda(x) F_{\theta^\varepsilon, y}(x) d\mu(x) \right|^2 \leq \|\tau\|_{\ell_2(\Lambda)}^2 \|Dh_\varepsilon(\theta^\varepsilon, y)\|_{\ell_2(\Lambda)}^2.$$

Moreover, it follows from Parseval's identity, [40][Theorem 1.7] together with the fact that  $\int_{\mathcal{X}} F_{\theta^\varepsilon, y}(x) d\mu(x) = 1$ , that

$$(A.15) \quad \|Dh_\varepsilon(\theta^\varepsilon, y)\|_{\ell_2(\Lambda)}^2 = \int_{\mathcal{X}} F_{\theta^\varepsilon, y}^2(x) d\mu(x) - 1.$$

Hence, combining (A.14) and (A.15), we obtain that

$$(A.16) \quad \int_{\mathcal{Y}} \left| \sum_{\lambda \in \Lambda} \tau_\lambda \int_{\mathcal{X}} \phi_\lambda(x) F_{\theta^\varepsilon, y}(x) d\mu(x) \right|^2 d\nu(y) \leq \|\tau\|_{\ell_2}^2 \left( \int_{\mathcal{Y}} \int_{\mathcal{X}} F_{\theta^\varepsilon, y}^2(x) d\mu(x) d\nu(y) - 1 \right).$$

Finally, we deduce (A.11) from (A.12), (A.13) and (A.16).  $\square$

**A.3. Proof of Theorem 3.1.** We shall proceed to the almost sure convergence of the random sequence  $(\hat{\theta}_n)_n$ . Let  $(V_n)$  be the Lyapunov sequence defined, for all  $n \geq 1$ , by

$$V_n = \|\hat{\theta}_n - \theta^\varepsilon\|_{W^{-1}}^2.$$

Assumption 3.1 ensures that  $\|\theta^\varepsilon\|_{W^{-1}} < +\infty$ . Moreover, we clearly have from (2.5) that

$$W^{-1/2} \hat{\theta}_{n+1} = W^{-1/2} \hat{\theta}_n - \gamma_n W^{1/2} D_\theta h_\varepsilon(\hat{\theta}_n, Y_{n+1}),$$

where  $W^\alpha$  stands for the linear operator, for  $\alpha \in \{-1/2, 1/2\}$ , that maps  $v = (v_\lambda)_{\lambda \in \Lambda} \in \ell_\infty(\Lambda)$  to  $(w_\lambda^\alpha v_\lambda)_{\lambda \in \Lambda}$ . It follows from (2.8) that  $\|\hat{\theta}_n\|_{W^{-1}} = \|W^{-1/2} \hat{\theta}_n\|_{\ell_2}$ . Consequently,

$$\|\hat{\theta}_{n+1}\|_{W^{-1}} \leq \|\hat{\theta}_n\|_{W^{-1}} + \gamma_n \|W^{1/2} D_\theta h_\varepsilon(\hat{\theta}_n, Y_{n+1})\|_{\ell_2}.$$

Furthermore, we obtain from (4.3) that

$$\|W^{1/2} D_\theta h_\varepsilon(\hat{\theta}_n, Y_{n+1})\|_{\ell_2}^2 \leq \|w\|_{\ell_1} \sup_{\lambda \in \Lambda} \left| \frac{\partial h_\varepsilon(\theta, y)}{\partial \theta_\lambda} \right|^2 \leq \|w\|_{\ell_1} < \infty.$$

Therefore, thanks to the assumption that  $\|\hat{\theta}_0\|_{W^{-1}} < +\infty$ , we deduce by induction that  $\|\hat{\theta}_n\|_{W^{-1}} < +\infty$ , which means that the Lyapunov sequence  $(V_n)$  is well defined. From now on, it follows from (2.5) and (2.8) that for all  $n \geq 0$ ,

$$\begin{aligned} V_{n+1} &= \|\hat{\theta}_n - \theta^\varepsilon - \gamma_n W D_\theta h_\varepsilon(\hat{\theta}_n, Y_{n+1})\|_{W^{-1}}^2, \\ &= V_n - 2\gamma_n \langle \hat{\theta}_n - \theta^\varepsilon, D_\theta h_\varepsilon(\hat{\theta}_n, Y_{n+1}) \rangle + \gamma_n^2 \|D_\theta h_\varepsilon(\hat{\theta}_n, Y_{n+1})\|_W^2. \end{aligned}$$

Moreover, (4.3) implies that  $\|D_\theta h_\varepsilon(\hat{\theta}_n, Y_{n+1})\|_W^2 \leq \|w\|_{\ell_1}$  which ensures that for all  $n \geq 0$ ,

$$(A.17) \quad V_{n+1} \leq V_n - 2\gamma_n \langle \hat{\theta}_n - \theta^\varepsilon, D_\theta h_\varepsilon(\hat{\theta}_n, Y_{n+1}) \rangle + \gamma_n^2 \|w\|_{\ell_1}.$$

Denote by  $\mathcal{F}_n = \sigma(Y_1, \dots, Y_n)$  the  $\sigma$ -algebra generated by  $Y_1, \dots, Y_n$  drawn from  $\nu$ . From Proposition 4.1,  $\mathbb{E}[D_\theta h_\varepsilon(\hat{\theta}_n, Y_{n+1}) | \mathcal{F}_n] = DH_\varepsilon(\hat{\theta}_n)$ , which implies via (A.17) that for all  $n \geq 0$ ,

$$(A.18) \quad \mathbb{E}[V_{n+1} | \mathcal{F}_n] \leq V_n + A_n - B_n \quad \text{a.s.}$$

where  $(A_n)$  and  $(B_n)$  are the two positive sequences given, for all  $n \geq 0$ , by

$$A_n = \gamma_n^2 \|w\|_{\ell_1} \quad \text{and} \quad B_n = 2\gamma_n DH_\varepsilon(\hat{\theta}_n)[\hat{\theta}_n - \theta^\varepsilon].$$

Therefore, as  $\sum_{n=0}^{\infty} A_n < \infty$ , we deduce from the Robbins-Siegmund theorem [38] that the sequence  $(V_n)$  converges almost surely to a finite random variable  $V$  and that the series

$$(A.19) \quad \sum_{n=0}^{\infty} B_n = 2 \sum_{n=0}^{\infty} \gamma_n DH_\varepsilon(\hat{\theta}_n)[\hat{\theta}_n - \theta^\varepsilon] < \infty \quad \text{a.s.}$$

Hence, by combining (A.19) with the first condition in (2.6), it necessarily follows that

$$(A.20) \quad \lim_{n \rightarrow \infty} DH_\varepsilon(\hat{\theta}_n)[\hat{\theta}_n - \theta^\varepsilon] = 0 \quad \text{a.s.}$$

Hereafter, our goal is to prove that  $\|\hat{\theta}_n - \theta^\varepsilon\|_{\ell_2}$  goes to zero almost surely as  $n$  tends to infinity. From now on, let  $g$  be the function defined in (4.12). One can easily see that  $g$  is a continuous and strictly decreasing function. Moreover, using the Cauchy-Schwarz inequality, one has that  $\|\hat{\theta}_n - \theta^\varepsilon\|_{\ell_1}^2 \leq \|w\|_{\ell_1} V_n$ . Hence, it follows from inequality (4.11) that for all  $n \geq 0$ ,

$$(A.21) \quad DH_\varepsilon(\hat{\theta}_n)[\hat{\theta}_n - \theta^\varepsilon] \geq g\left(\frac{2}{\varepsilon} \|w\|_{\ell_1} V_n\right) D^2 H_\varepsilon(\theta^\varepsilon)[\hat{\theta}_n - \theta^\varepsilon, \hat{\theta}_n - \theta^\varepsilon].$$

Therefore, we obtain from Assumption 3.2 and inequality (A.21) that for all  $n \geq 0$ ,

$$(A.22) \quad DH_\varepsilon(\hat{\theta}_n)[\hat{\theta}_n - \theta^\varepsilon] \geq c_\varepsilon g\left(\frac{2}{\varepsilon} \|w\|_{\ell_1} V_n\right) \|\hat{\theta}_n - \theta^\varepsilon\|_{\ell_2}^2.$$

Since  $(V_n)$  converges a.s. to a finite random variable  $V$ , it follows by continuity of  $g$  that

$$(A.23) \quad \lim_{n \rightarrow \infty} g\left(\frac{2}{\varepsilon} \|w\|_{\ell_1} V_n\right) = g\left(\frac{2}{\varepsilon} \|w\|_{\ell_1} V\right) \quad \text{a.s.}$$

and the limit in the right-hand side of (A.23) is positive almost surely. Therefore, we conclude from (A.20), (A.22) and (A.23) that

$$\lim_{n \rightarrow \infty} \|\hat{\theta}_n - \theta^\varepsilon\|_{\ell_2} = 0 \quad \text{a.s.}$$

Finally, we deduce from Parseval's identity, [40][Theorem 1.7] that

$$\int_{\mathcal{X}} |\hat{u}_\varepsilon^n(x) - u_\varepsilon(x)|^2 d\mu(x) = \|\hat{\theta}_n - \theta^\varepsilon\|_{\ell_2}^2$$

which achieves the proof of Theorem 3.1. □



## APPENDIX A. PROOF OF PROPOSITION 4.1

We first state a result about Fréchet differentiation under Lebesgue integrals, that follows from [16, Lemma A.2], and which extends well-known results on the differentiation of integral functionals. For the proof of a similar result, we also refer to the unpublished note [27].

**Lemma A.1** (Leibniz's rules of Fréchet differentiation) *Let  $(\Theta, \|\cdot\|)$  be an infinite dimensional Banach space and  $\sigma$  a finite measure on a measurable space  $\mathbb{T}$ . Let  $\theta_0 \in \Theta$  and denote by  $B(\theta_0, R) \subset \Theta$  the ball of center  $\theta_0$  and radius  $R$ . Consider a function  $f : \Theta \times \mathbb{T} \rightarrow \mathbb{R}$  that is Fréchet differentiable at  $\theta_0$  (for every  $t \in \mathbb{T}$ ), and suppose that there exists  $K \in L^1(\sigma)$  such that, for all  $\theta_1, \theta_2 \in B(\theta_0, R)$  and all  $t \in \mathbb{T}$ ,*

$$|f(\theta_1, t) - f(\theta_2, t)| \leq K(t)\|\theta_1 - \theta_2\|.$$

*Then, the integral functional  $F : \Theta \rightarrow \mathbb{R}$  defined by  $F(\theta) = \int_{\mathbb{T}} f(\theta, t) d\sigma(t)$  is Fréchet differentiable at  $\theta_0$  and*

$$DF(\theta_0) = \int_{\mathbb{T}} D_{\theta} f(\theta_0, t) d\sigma(t),$$

where  $D_{\theta} f(\theta_0, t)$  denotes the Fréchet derivative of  $\theta \mapsto f(\theta, t)$  at  $\theta_0$ .

In what follows, we will apply Lemma A.1 with  $\Theta = \bar{\ell}_1(\Lambda)$ ,  $\mathbb{T} = \mathcal{X}$  and  $\sigma = \mu$  to obtain the expression of the Fréchet differential of  $H_{\varepsilon}$ . Let us first prove that, for every  $y \in \mathcal{Y}$ , the function  $h_{\varepsilon}(\cdot, y) : \bar{\ell}_1(\Lambda) \rightarrow \mathbb{R}$  defined in (2.2) is Fréchet differentiable. To this end, we introduce the function  $g_y(\cdot, x) : \bar{\ell}_1(\Lambda) \rightarrow \mathbb{R}$  defined as

$$(A.1) \quad g_y(\theta, x) = \frac{1}{\varepsilon} \left( \sum_{\lambda \in \Lambda} \theta_{\lambda} \phi_{\lambda}(x) - c(x, y) \right)$$

and  $G_y(\theta) = \int_{\mathcal{X}} \exp(g_y(\theta, x)) d\mu(x)$ . In this way, one has that  $h_{\varepsilon}(\theta, y) = \varepsilon \log G_y(\theta) + \varepsilon$ . For every  $x \in \mathcal{X}$ , the function  $\theta \mapsto \exp(g_y(\theta, x))$  is clearly Fréchet differentiable and, for  $\tau \in \bar{\ell}_1(\Lambda)$ ,

$$(A.2) \quad D_{\theta} \exp(g_y(\theta, x))[\tau] = \frac{1}{\varepsilon} \sum_{\lambda \in \Lambda} \exp(g_y(\theta, x)) \phi_{\lambda}(x) \tau_{\lambda}.$$

Moreover, it is a bounded linear operator from  $\bar{\ell}_1(\Lambda)$  to  $\mathbb{R}$ . In what follows, we identify this operator to the infinite-dimensional vector

$$D_{\theta} \exp(g_y(\theta, x)) = \frac{1}{\varepsilon} \exp(g_y(\theta, x)) \left( \overline{\phi_{\lambda}(x)} \right)_{\lambda \in \Lambda}.$$

From now on, let  $\theta_0 \in \bar{\ell}_1(\Lambda)$  and  $R > 0$ . Then, for any  $\theta_1, \theta_2 \in B(\theta_0, R)$ , the mean value theorem for functions defined on a Banach space implies that

$$(A.3) \quad |\exp(g_y(\theta_1, x)) - \exp(g_y(\theta_2, x))| \leq \sup_{\theta \in B(\theta_0, R)} \|D_{\theta} \exp(g_y(\theta, x))\|_{op} \|\theta_1 - \theta_2\|,$$

where the operator norm of  $D_{\theta} \exp(g_y(\theta, x))$  is defined as

$$\|D_{\theta} \exp(g_y(\theta, x))\|_{op} = \sup_{\|\tau\|_{\ell_1} \leq 1} |D_{\theta} \exp(g_y(\theta, x))[\tau]|.$$

Since

$$|D_{\theta} \exp(g_y(\theta, x))[\tau]| = \left| \frac{1}{\varepsilon} \sum_{\lambda \in \Lambda} \exp(g_y(\theta, x)) \phi_{\lambda}(x) \tau_{\lambda} \right| \leq \frac{1}{\varepsilon} \exp(g_y(\theta, x)) \sum_{\lambda \in \Lambda} |\tau_{\lambda}|,$$

one has that, for any  $\theta \in \bar{\ell}_1(\Lambda)$ ,

$$(A.4) \quad \|D_\theta \exp(g_y(\theta, x))\|_{op} \leq \frac{1}{\varepsilon} \exp(g_y(\theta, x)) \leq \frac{1}{\varepsilon} \exp\left(\frac{\sum_{\lambda \in \Lambda} |\theta_\lambda| + c(x, y)}{\varepsilon}\right)$$

Consequently, let

$$K_y(x) = \frac{1}{\varepsilon} \exp\left(\frac{c(x, y)}{\varepsilon}\right) \sup_{\theta \in B(\theta_0, R)} \exp\left(\frac{\|\theta\|_{\ell_1}}{\varepsilon}\right).$$

It follows from (A.3) and (A.4) that, for all  $\theta_1, \theta_2 \in B(\theta_0, R)$  and  $x \in \mathcal{X}$ ,

$$|\exp(g_y(\theta_1, x)) - \exp(g_y(\theta_2, x))| \leq K_y(x) \|\theta_1 - \theta_2\|.$$

Obviously, for all  $y \in \mathcal{Y}$ , the function  $K_y$  belongs to  $L^1(\mu)$ , and therefore, by Lemma A.1, we conclude that  $G_y$  and  $h_\varepsilon(\cdot, y)$  are Fréchet differentiable and that the linear operator  $D_\theta h_\varepsilon(\theta, y)$  is identified as an element of  $\bar{\ell}_\infty(\Lambda)$  given by

$$(A.5) \quad D_\theta h_\varepsilon(\theta, y) = \left( \frac{\int_{\mathcal{X}} \overline{\phi_\lambda(x)} \exp(g_y(\theta, x)) d\mu(x)}{\int_{\mathcal{X}} \exp(g_y(\theta, x)) d\mu(x)} \right)_{\lambda \in \Lambda}.$$

Similarly, to prove that the function  $H_\varepsilon(\theta) = \int_{\mathcal{Y}} h_\varepsilon(\theta, y) d\nu(y)$  is Fréchet differentiable, it is sufficient to bound the operator norm of  $D_\theta h_\varepsilon(\theta, y)$  for  $\theta \in B(\theta_0, R)$ . Recalling that

$$F_{\theta, y}(x) = \frac{\exp(g_y(\theta, x))}{\int_{\mathcal{X}} \exp(g_y(\theta, x)) d\mu(x)},$$

we remark that, for any  $\tau \in \bar{\ell}_1(\Lambda)$ ,

$$|D_\theta h_\varepsilon(\theta, y)[\tau]| = \left| \sum_{\lambda \in \Lambda} \int_{\mathcal{X}} F_{\theta, y}(x) \phi_\lambda(x) d\mu(x) \tau_\lambda \right| \leq \int_{\mathcal{X}} F_{\theta, y}(x) d\mu(x) \sum_{\lambda \in \Lambda} |\tau_\lambda| = \|\tau\|_{\ell_1}.$$

Therefore,  $\|D_\theta h_\varepsilon(\theta, y)\|_{op} \leq 1$  which proves inequality (4.3). It also means that  $D_\theta h_\varepsilon(\theta, y)$  can be identified as an element of  $\bar{\ell}_\infty(\Lambda)$ . Thus, arguing as previously, that is by combining the mean value theorem with Lemma A.1, we obtain that  $H_\varepsilon(\theta)$  is Fréchet differentiable with

$$DH_\varepsilon(\theta) = \int_{\mathcal{Y}} D_\theta h_\varepsilon(\theta, y) d\nu(y) = \left( \int_{\mathcal{Y}} \frac{\partial h_\varepsilon(\theta, y)}{\partial \theta_\lambda} d\nu(y) \right)_{\lambda \in \Lambda}$$

which can also be identified as an element of  $\bar{\ell}_\infty(\Lambda)$  such that  $\|DH_\varepsilon(\theta)\|_{op} = \|DH_\varepsilon(\theta)\|_{\ell_\infty}$  satisfies inequality (4.5) by combining inequality (4.3) together with the fact that  $\nu$  is a probability measure. This achieves the proof of Proposition 4.1.  $\square$

## APPENDIX B. PROOF OF PROPOSITION 4.2

First, let us recall that, for  $(\Theta, \|\cdot\|)$  a given Banach space, a function  $f : \Theta \rightarrow \mathbb{R}$  is twice Fréchet differentiable if  $Df$  is Fréchet differentiable. In this case, the second order Fréchet derivative of  $f$  at  $\theta_0$  is denoted by  $D^2 f(\theta_0)$  and it is identified as an element of  $L(\Theta \times \Theta, \mathbb{R})$  the set of continuous bilinear mapping from  $\Theta \times \Theta$  to  $\mathbb{R}$ . Moreover, the operator norm of  $D^2 f(\theta_0)$  is defined as

$$\|D^2 f(\theta_0)\|_{op} = \sup_{\|\theta\| \leq 1, \|\theta'\| \leq 1} |D^2 f(\theta_0)[\theta, \theta']|.$$

To derive the expression of the second order Fréchet derivative of the functions  $h_\varepsilon(\cdot, y)$  and  $H_\varepsilon$ , we use similar arguments to those in the proof of Proposition 4.1. First, recall from (A.5) that the Fréchet derivative  $D_\theta h_\varepsilon(\theta, y)$  is the linear operator defined as

$$D_\theta h_\varepsilon(\theta, y) = \left( \int F_{\theta, y}(x) \overline{\phi_\lambda(x)} d\mu(x) \right)_{\lambda \in \Lambda} = \int \psi_y(x, \theta) d\mu(x)$$

where  $\psi_y(x, \theta) : \bar{\ell}_1(\Lambda) \rightarrow \mathbb{R}$  is the linear operator

$$\psi_y(x, \theta)[\tau] = \sum_{\lambda \in \Lambda} F_{\theta, y}(x) \phi_\lambda(x) \tau_\lambda = \sum_{\lambda \in \Lambda} F_{\theta, y}(x) \overline{\phi_\lambda(x)} \overline{\tau_\lambda}.$$

As a standard strategy, we aim to derive this with respect to  $\theta$ . From (A.1), one has that

$$F_{\theta, y}(x) = \frac{\exp(g_y(\theta, x))}{G_y(\theta)}.$$

Therefore, using (A.2) combined with the differentiability of  $G_y(\theta)$ ,

$$(B.1) \quad D_\theta F_{\theta, y}(x)[\tau] = \frac{1}{\varepsilon} \left( \sum_{\lambda \in \Lambda} \tau_\lambda \phi_\lambda(x) F_{\theta, y}(x) - F_{\theta, y}(x) \int_{\mathcal{X}} \sum_{\lambda \in \Lambda} \tau_\lambda \phi_\lambda(z) F_{\theta, y}(z) d\mu(z) \right).$$

Thus, the mapping  $\theta \mapsto \psi_y(x, \theta)[\tau]$  is clearly Fréchet differentiable, and its Fréchet derivative can be identified as the following symmetric bilinear mapping from  $\bar{\ell}_1(\Lambda) \times \bar{\ell}_1(\Lambda)$  to  $\mathbb{R}$

$$\begin{aligned} D_\theta \psi_y(x, \theta)[\tau, \tau'] &= \frac{1}{\varepsilon} \left( \sum_{\lambda' \in \Lambda} \sum_{\lambda \in \Lambda} \tau'_{\lambda'} \overline{\tau_\lambda} \left( \phi_{\lambda'}(x) \overline{\phi_\lambda(x)} F_{\theta, y}(x) - \phi_{\lambda'}(x) F_{\theta, y}(x) \overline{\phi_\lambda(x)} F_{\theta, y}(x) \right) \right) \\ &= \frac{1}{\varepsilon} \left( \sum_{\lambda' \in \Lambda} \sum_{\lambda \in \Lambda} \tau'_{\lambda'} \overline{\tau_\lambda} \phi_{\lambda'}(x) \overline{\phi_\lambda(x)} F_{\theta, y}(x) - \sum_{\lambda \in \Lambda} \tau'_\lambda \phi_\lambda(x) F_{\theta, y}(x) \overline{\sum_{\lambda \in \Lambda} \tau_\lambda \phi_\lambda(x) F_{\theta, y}(x)} \right). \end{aligned}$$

We now compute an upper bound for the norm of this linear operator. One can observe that, for  $\tau = \tau'$ ,

$$(B.2) \quad |D_\theta \psi_y(x, \theta)[\tau, \tau]| \leq \frac{1}{\varepsilon} F_{\theta, y}(x) \|\tau\|_{\ell_1}^2,$$

thanks to the elementary fact that

$$(B.3) \quad \sum_{\lambda \in \Lambda} \tau_\lambda \phi_\lambda(x) F_{\theta, y}(x) \overline{\sum_{\lambda \in \Lambda} \tau_\lambda \phi_\lambda(x) F_{\theta, y}(x)} = \left| \sum_{\lambda \in \Lambda} \tau_\lambda \phi_\lambda(x) F_{\theta, y}(x) \right|^2 \geq 0.$$

Then, using the equality,

$$4D_\theta \psi_y(x, \theta)[\tau, \tau'] = D_\theta \psi_y(x, \theta)[\tau + \tau', \tau + \tau'] - D_\theta \psi_y(x, \theta)[\tau - \tau', \tau - \tau']$$

combined with the upper bound (B.2), we obtain that

$$4|D_\theta \psi_y(x, \theta)[\tau, \tau']| \leq \frac{1}{\varepsilon} F_{\theta, y}(x) (\|\tau + \tau'\|_{\ell_1}^2 + \|\tau - \tau'\|_{\ell_1}^2).$$

Therefore, we immediately obtain that

$$\sup_{\|\tau\|_{\ell_1} \leq 1, \|\tau'\|_{\ell_1} \leq 1} |D_\theta \psi_y(x, \theta)[\tau, \tau']| \leq \frac{2}{\varepsilon} F_{\theta, y}(x).$$

Consequently, we may proceed as in the proof of Proposition 4.1 to obtain that  $h_\varepsilon(\cdot, y)$  is twice Fréchet differentiable and that its second Fréchet derivative is the following symmetric bilinear mapping from  $\bar{\ell}_1(\Lambda) \times \bar{\ell}_1(\Lambda)$  to  $\mathbb{R}$

$$\begin{aligned} D_\theta^2 h_\varepsilon(\theta, y)[\tau, \tau'] &= \frac{1}{\varepsilon} \sum_{\lambda' \in \Lambda} \sum_{\lambda \in \Lambda} \tau_{\lambda'}' \overline{\tau_\lambda} \int_{\mathcal{X}} \phi_{\lambda'}(x) \overline{\phi_\lambda}(x) F_{\theta, y}(x) d\mu(x) \\ &\quad - \frac{1}{\varepsilon} \left( \sum_{\lambda \in \Lambda} \tau_\lambda' \int_{\mathcal{X}} \phi_\lambda(x) F_{\theta, y}(x) d\mu(x) \right) \overline{\left( \sum_{\lambda \in \Lambda} \tau_\lambda \int_{\mathcal{X}} \phi_\lambda(x) F_{\theta, y}(x) d\mu(x) \right)}. \end{aligned}$$

Note that, for  $\tau = \tau'$ , an application of Jensen's inequality with respect to the probability measure  $F_{\theta, y}(x) d\mu(x)$  implies that  $D_\theta^2 h_\varepsilon(\theta, y)[\tau, \tau] \geq 0$ . Moreover, it follows once again from (B.3) together with the elementary fact that  $\int_{\mathcal{X}} F_{\theta, y} d\mu = 1$ , that

$$(B.4) \quad D_\theta^2 h_\varepsilon(\theta, y)[\tau, \tau] \leq \frac{1}{\varepsilon} \|\tau\|_{\ell_1}^2.$$

Hereafter, we deduce from the equality

$$4D_\theta^2 h_\varepsilon(\theta, y)[\tau, \tau'] = D_\theta^2 h_\varepsilon(\theta, y)[\tau + \tau', \tau + \tau'] - D_\theta^2 h_\varepsilon(\theta, y)[\tau - \tau', \tau - \tau'],$$

the positivity of  $D_\theta^2 h_\varepsilon(\theta, y)[\tau - \tau', \tau - \tau']$  and inequality (B.4), that

$$4|D_\theta^2 h_\varepsilon(\theta, y)[\tau, \tau']| \leq D_\theta^2 h_\varepsilon(\theta, y)[\tau + \tau', \tau + \tau'] \leq \frac{1}{\varepsilon} \|\tau + \tau'\|_{\ell_1}^2.$$

It ensures that

$$\|D_\theta^2 h_\varepsilon(\theta, y)\|_{op} = \sup_{\|\tau\|_{\ell_1} \leq 1, \|\tau'\|_{\ell_1} \leq 1} |D_\theta^2 h_\varepsilon(\theta, y)[\tau, \tau']| \leq \frac{1}{\varepsilon},$$

which proves inequality (4.7).

Finally, combining the above upper bound on  $\|D_\theta^2 h_\varepsilon(\theta, y)\|_{op}$  and using again an adaptation of Lemma A.1 to obtain a Leibniz's formula for the second order Fréchet differentiation under the integral sign, one can prove that  $H_\varepsilon(\theta)$  is twice Fréchet differentiable by integrating  $D_\theta^2 h_\varepsilon(\theta, y)$  with respect to  $d\nu(y)$ , which implies that  $D^2 H_\varepsilon(\theta)$  is the linear operator defined by (4.8). Moreover, the upper bound (4.9) follows from inequality (4.7) and the fact that  $\nu$  is a probability measure, which completes the proof of Proposition 4.2.  $\square$

### APPENDIX C. PROOF OF PROPOSITION 4.3

For  $x \in \mathcal{X}$ ,  $y \in \mathcal{Y}$  and for  $(\theta^{(1)}, \theta^{(2)}) \in \bar{\ell}_1(\Lambda) \times \bar{\ell}_1(\Lambda)$ , denote

$$e_1(x, y) = \exp \left( \frac{\sum_{\lambda \in \Lambda} \theta_\lambda^{(1)} \phi_\lambda(x) - c(x, y)}{\epsilon} \right)$$

and

$$e_2(x, y) = \exp \left( \frac{\sum_{\lambda \in \Lambda} \theta_\lambda^{(2)} \phi_\lambda(x) - c(x, y)}{\epsilon} \right).$$

We have, for all  $0 < t < 1$ , and for a fixed  $y \in \mathcal{Y}$ , that

$$\begin{aligned} (C.1) \quad t h_\varepsilon(\theta^{(1)}, y) + (1-t) h_\varepsilon(\theta^{(2)}, y) &= \varepsilon t \log \int_{\mathcal{X}} e_1(x, y) d\mu(x) + (1-t) \log \int_{\mathcal{X}} e_2(x, y) d\mu(x) + \varepsilon \\ &= \varepsilon \log \left( \left( \int_{\mathcal{X}} e_1(x, y) d\mu(x) \right)^t \left( \int_{\mathcal{X}} e_2(x, y) d\mu(x) \right)^{1-t} \right) + \varepsilon. \end{aligned}$$

Hereafter, applying Hölder's inequality to the functions  $f(x) = e_1^t(x, y)$  and  $g(x) = e_2^{1-t}(x, y)$  with Hölder conjugates  $p = 1/t$  and  $q = 1/(1-t)$ , we obtain that

$$(C.2) \quad \int_{\mathcal{X}} f(x)g(x) d\mu(x) \leq \left( \int_{\mathcal{X}} e_1(x, y) d\mu(x) \right)^t \left( \int_{\mathcal{X}} e_2(x, y) d\mu(x) \right)^{1-t}.$$

However, one can observe that

$$\int_{\mathcal{X}} f(x)g(x) d\mu(x) = \int_{\mathcal{X}} \exp \left( \frac{t \sum_{\lambda \in \Lambda} \theta_{\lambda}^{(1)} \phi_{\lambda}(x) + (1-t) \sum_{\lambda \in \Lambda} \theta_{\lambda}^{(2)} \phi_{\lambda}(x) + c(x, y)}{\varepsilon} \right) d\mu(x),$$

which ensures that

$$\varepsilon \log \int_{\mathcal{X}} f(x)g(x) d\mu(x) + \varepsilon = h_{\varepsilon}(t\theta^{(1)} + (1-t)\theta^{(2)}, y).$$

Hence, combining the above equality with (C.1) and (C.2), we obtain that

$$h_{\varepsilon}(t\theta^{(1)} + (1-t)\theta^{(2)}, y) \leq th_{\varepsilon}(\theta^{(1)}, y) + (1-t)h_{\varepsilon}(\theta^{(2)}, y),$$

which proves the convexity of  $\theta \mapsto h_{\varepsilon}(\theta, y)$ . Since  $H_{\varepsilon}(\theta) = \mathbb{E}[h_{\varepsilon}(\theta, Y)]$ , we also obtain the convexity of the function  $H_{\varepsilon}$ . Furthermore, assume that Hölder's inequality (C.2) becomes an equality, which means that the functions  $f^p$  and  $g^q$  are linearly dependent in  $L^1(\mu)$ . This would mean that it exists  $\beta_y > 0$  such that  $e_1(x, y) = \beta_y e_2(x, y)$  for all  $x \in \mathcal{X}$ . Applying the logarithm, this equality is equivalent to

$$\frac{1}{\varepsilon} \sum_{\lambda \in \Lambda} \theta_{\lambda}^{(1)} \phi_{\lambda}(x) = \log \beta_y + \frac{1}{\varepsilon} \sum_{\lambda \in \Lambda} \theta_{\lambda}^{(2)} \phi_{\lambda}(x).$$

By integrating the above equality with respect to  $\mu$ , and from our normalization condition (2.1), the equality case in the Hölder inequality (C.2) implies that  $\log \beta_y = 0$  and thus  $\beta_y = 1$ . But then one has that  $e_1(x, y) = e_2(x, y)$ , implying that  $\sum_{\lambda \in \Lambda} (\theta_{\lambda}^{(1)} - \theta_{\lambda}^{(2)}) \phi_{\lambda}(x) = 0$  for all  $x \in \mathcal{X}$ . Hence, we necessarily have that  $\theta^{(1)} = \theta^{(2)}$  which yields a contradiction. Therefore, the function  $\theta \mapsto h_{\varepsilon}(\theta, y)$  is strictly convex. Since  $H_{\varepsilon}(\theta) = \mathbb{E}[h_{\varepsilon}(\theta, Y)]$  this also implies the strict convexity of  $H_{\varepsilon}$ , which achieves the proof of Proposition 4.3.  $\square$

## APPENDIX D. OPTIMAL TRANSPORT WITH PERIODICITY CONSTRAINTS

In this section, we focus our attention on conditions such that the dual potential (1.3) is periodic at the boundary of  $\mathcal{X}$ .

**D.1. The case of the standard quadratic cost.** Using classical results in the analysis of multiple Fourier series (see e.g. [40, Corollary 1.8]), assuming that the Fourier coefficients  $\theta^0 = (\theta_{\lambda}^0)_{\lambda \in \Lambda}$  of  $u_0$  form an absolutely convergent series implies that  $u_0$  can be extended as a continuous and  $\mathbb{Z}^d$ -periodic function on  $\mathbb{R}^d$ . Hence, under this assumption,  $u_0$  has to be a continuous function that is constant at the boundary of  $\mathcal{X} = [0, 1]^d$ . However, for the quadratic cost  $c(x, y) = \frac{1}{2} \|x - y\|^2$ , we are not aware of standard results on the regularity of optimal transport (through smoothness assumptions on  $\nu$ ) that would imply periodic properties of  $u_0$  and its derivatives at the boundary of  $\mathcal{X}$ .

**D.2. The quadratic cost on the torus.** Nevertheless, guaranteeing the periodicity of  $u_0$  and the summability of its Fourier coefficients is feasible by considering the setting  $\mathcal{X} = \mathcal{Y} = \mathbb{T}^d$ , where  $\mathbb{T}^d = \mathbb{R}^d / \mathbb{Z}^d$  is the  $d$ -dimensional torus, that is endowed with the usual distance

$$d_{\mathbb{T}^d}(x, y) = \min_{\lambda \in \mathbb{Z}^d} \|x - y + \lambda\|.$$

Hereafter, we identify the torus as the set of equivalence classes  $\{x + \lambda : \lambda \in \mathbb{Z}^d\}$  for  $x \in [0, 1]^d$ , and we use the notation  $[x] = x + \lambda_0$  where  $\lambda_0 \in \mathbb{Z}^d$  is such that  $\|x + \lambda\|$  is minimal for  $\lambda \in \mathbb{Z}^d$ . We also recall that a function  $u : \mathbb{T}^d \rightarrow \mathbb{R}$  can be identified as a  $\mathbb{Z}^d$ -periodic function on  $\mathbb{R}^d$ . Finally, one can observe that for a given  $y \in \mathbb{T}^d$ , the cost function  $c(x, y) = \frac{1}{2}d_{\mathbb{T}^d}^2(x, y)$  is almost everywhere differentiable, and its gradient is (see e.g. [39, Section 1.3.2])

$$\nabla_x c(x, y) = [x - y],$$

at every  $x \notin y + \{\partial\Omega + \mathbb{Z}^d\}$  where  $\partial\Omega$  denotes the boundary of  $\Omega = [-\frac{1}{2}, \frac{1}{2}]^d$ .

Assuming that the probability measure  $\nu$  is also supported on the  $d$ -dimensional torus  $\mathbb{T}^d$  allows to use existing results for optimal transport on the torus (see e.g. [15], [31, Section 2.2] and [39, Section 1.3.2]). Formally, taking  $\mathcal{X} = \mathcal{Y} = \mathbb{T}^d$  implies that  $\nu$  is considered as a periodic positive Radon measure on  $\mathbb{R}^d$  with  $\nu(\mathbb{T}^d) = 1$ , and that  $\mu$  is understood as the Lebesgue measure on  $\mathbb{R}^d$ . Note that this setting is not restrictive, as it allows to treat the example of an absolutely continuous measure  $\nu$  with support on  $[0, 1]^d$  whose density  $f_\nu$  takes a constant value on the boundary of  $[0, 1]^d$ , implying that  $f_\nu$  can be extended over  $\mathbb{R}^d$  as a  $\mathbb{Z}^d$ -periodic function.

Then, thanks to the identification of  $u : \mathbb{T}^d \rightarrow \mathbb{R}$  as a  $\mathbb{Z}^d$ -periodic function on  $\mathbb{R}^d$ , it follows that

$$\inf_{x \in \mathbb{T}^d} \left\{ \frac{1}{2}d_{\mathbb{T}^d}^2(x, y) - u(x) \right\} = \inf_{x \in \mathbb{R}^d} \left\{ \frac{1}{2}\|x - y\|^2 - u(x) \right\}.$$

Therefore, it is equivalent to define the conjugate of a function  $u : \mathbb{T}^d \rightarrow \mathbb{R}$  with respect to the cost  $c(x, y) = \frac{1}{2}d_{\mathbb{T}^d}^2(x, y)$  or to the quadratic cost  $c(x, y) = \frac{1}{2}\|x - y\|^2$  using the periodization of  $u$  over  $\mathbb{R}^d$ . Now, using results on optimal transport on  $\mathbb{T}^d$ , previously established in [15] or [31, Proposition 4], it follows that

- (i) there exists a unique optimal transport map  $Q : \mathbb{T}^d \rightarrow \mathbb{T}^d$  from  $\mu$  to  $\nu$  such that

$$Q = \operatorname{argmin}_{T : T\#\mu = \nu} \mathbb{E} (d_{\mathbb{T}^d}^2(X, T(X))),$$

- (ii)  $Q(x) = x - \nabla u_0(x)$  where  $u_0$  is a  $\mathbb{Z}^d$ -periodic function on  $\mathbb{R}^d$  that is a solution of the dual problem (1.3) with  $\mathcal{X} = \mathcal{Y} = \mathbb{T}^d$  and  $c(x, y) = \frac{1}{2}d_{\mathbb{T}^d}^2(x, y)$ ,
- (iii)  $\|Q(x) - x\|^2 = d_{\mathbb{T}^d}^2(x, Q(x))$  for almost every  $x \in \mathbb{R}^d$ .

Entropically regularized optimal transport on the torus has also been recently considered in [6] and [14, Section E]. One can thus also consider the dual formulation of entropic OT as in (1.6) with the cost  $c(x, y) = \frac{1}{2}d_{\mathbb{T}^d}^2(x, y)$ .

We conclude this section on optimal transport on the torus by a discussion on the regularity of the optimal dual functions in the un-regularized case  $\varepsilon = 0$ . For  $s \in \mathbb{N}$ , we denote by  $\mathcal{C}^s(\mathbb{T}^d)$ , the set of  $\mathbb{Z}^d$ -periodic functions  $f$  on  $\mathbb{R}^d$  having everywhere defined continuous partial derivatives. Then, the following regularity result holds as an immediate application of results from [15] and [31, Theorem 5].

**Lemma D.1.** *Let  $u_0$  be a solution of the dual problem (1.3) with  $\mathcal{X} = \mathcal{Y} = \mathbb{T}^d$  and  $c(x, y) = \frac{1}{2}d_{\mathbb{T}^d}^2(x, y)$ . Suppose that the probability distribution  $\nu$  is absolutely continuous*

with a density  $f_\nu$  that is lower and upper bounded by positive constants. Assume further that  $f_\nu \in \mathcal{C}^{s-1}(\mathbb{T}^d)$  for some  $s > 1$ . Then,  $u_0$  belongs to  $\mathcal{C}^{s+1}(\mathbb{T}^d)$ .

Consequently, under the assumptions of Lemma D.1, one has that if  $f_\nu \in \mathcal{C}^{s-1}(\mathbb{T}^d)$  for some  $s > d/2 - 1$ , then  $u_0$  belongs to  $\mathcal{C}^k(\mathbb{T}^d)$  with  $k > d/2$ . Therefore, by standard results for multiple Fourier series (see e.g. [40, Corollary 1.9]), one has that  $\sum_{\lambda \in \Lambda} |\theta_\lambda^0| < +\infty$ . Hence we can conclude that if the density of  $\nu$  is sufficiently smooth (and is upper and lower bounded by positive constants), then the Fourier series of  $u_0$  actually belongs to  $\ell_1(\Lambda)$ .

## REFERENCES

- [1] F. BACH, *Self-concordant analysis for logistic regression*, Electronic Journal of Statistics, 4 (2010), pp. 384 – 414.
- [2] F. R. BACH, *Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression*, Journal of Machine Learning Research, 15 (2014), pp. 595–627.
- [3] J. BEIRLANT, S. BUITENDAG, E. DEL BARRIO, M. HALLIN, AND F. KAMPER, *Center-outward quantiles and the measurement of multivariate risk*, Insurance: Mathematics and Economics, 95 (2020), pp. 79–100.
- [4] B. BERCU AND J. BIGOT, *Asymptotic distribution and convergence rates of stochastic algorithms for entropic optimal transportation between probability measures*, The Annals of Statistics, 49 (2021), pp. 968 – 987.
- [5] B. BERCU, J. BIGOT, S. GADAT, AND E. SIVIERO, *A stochastic Gauss-Newton algorithm for regularized semi-discrete optimal transport*, Information and Inference: A Journal of the IMA, (2022).
- [6] R. J. BERMAN, *The sinkhorn algorithm, parabolic optimal transport and geometric monge-ampère equations*, Numer. Math., 145 (2020), pp. 771–836.
- [7] Y. BRENIER, *Polar factorization and monotone rearrangement of vector-valued functions*, Comm. Pure Appl. Math., 44 (1991), pp. 375–417.
- [8] S. BUBECK, *Convex optimization: Algorithms and complexity*, Found. Trends Mach. Learn., 8 (2015).
- [9] C. BUNNE, A. KRAUSE, AND M. CUTURI, *Supervised training of conditional monge maps*, Advances in Neural Information Processing Systems, 35 (2022).
- [10] G. CARLIER, V. CHERNOZHUKOV, G. DE BIE, AND A. GALICHON, *Vector quantile regression and optimal transport, from theory to numerics*, Empirical Economics, 62 (2022), pp. 35–62.
- [11] P. CHAUDHURI, *On a geometric notion of quantiles for multivariate data*, Journal of the American Statistical Association, 91 (1996), pp. 862–872.
- [12] V. CHERNOZHUKOV, A. GALICHON, M. HALLIN, AND M. HENRY, *Monge–Kantorovich depth, quantiles, ranks and signs*, The Annals of Statistics, 45 (2017), pp. 223 – 256.
- [13] S. CHEWI AND A.-A. POOLADIAN, *An entropic generalization of caffarelli’s contraction theorem via covariance inequalities*, Comptes Rendus. Mathématique, 361 (2023), pp. 1471–1482.
- [14] L. CHIZAT, P. ROUSSILLON, F. LÉGER, F.-X. VIALARD, AND G. PEYRÉ, *Faster wasserstein distance estimation with the sinkhorn divergence*, in Proc. NeurIPS’20, 2020.
- [15] D. CORDERO-ERAUSQUIN, *Sur le transport de mesures périodiques*, Comptes Rendus de l’Académie des Sciences - Series I - Mathematics, 329 (1999), pp. 199–202.
- [16] R. CORREA, A. HANTOUTE, AND P. PÉREZ-AROS, *Subdifferential calculus rules for possibly nonconvex integral functions*, SIAM Journal on Control and Optimization, 58 (2020), pp. 462–484.
- [17] J. A. CUESTA AND C. MATRAN, *Notes on the Wasserstein metric in Hilbert spaces*, Ann. Probab., 17 (1989), pp. 1264–1276.
- [18] M. CUTURI, *Sinkhorn distances: Lightspeed computation of optimal transport*, in Advances in Neural Information Processing Systems, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds., vol. 26, Curran Associates, Inc., 2013.
- [19] N. DEB, P. GHOSAL, AND B. SEN, *Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections*, Advances in Neural Information Processing Systems, 34 (2021), pp. 29736–29753.
- [20] R. FLAMARY, N. COURTY, A. GRAMFORT, M. Z. ALAYA, A. BOISBUNON, S. CHAMBON, L. CHAPEL, A. CORENFLOS, K. FATRAS, N. FOURNIER, L. GAUTHERON, N. T. GAYRAUD, H. JANATI, A. RAKOTOMAMONJY, I. REDKO, A. ROLET, A. SCHUTZ, V. SEGUY, D. J. SUTHERLAND, R. TAVENARD, A. TONG, AND T. VAYER, *Pot: Python optimal transport*, Journal of Machine Learning Research, 22 (2021), pp. 1–8.



- [21] A. GENEVAY, M. CUTURI, G. PEYRÉ, AND F. BACH, *Stochastic Optimization for Large-scale Optimal Transport*, in NIPS 2016 - Thirtieth Annual Conference on Neural Information Processing System, NIPS, ed., 2016.
- [22] P. GHOSAL, M. NUTZ, AND E. BERNTON, *Stability of entropic optimal transport and schrödinger bridges*, Journal of Functional Analysis, 283 (2022), p. 109622.
- [23] P. GHOSAL AND B. SEN, *Multivariate ranks and quantiles using optimal transport: Consistency, rates and nonparametric testing*, The Annals of Statistics, 50 (2022), pp. 1012 – 1037.
- [24] M. HALLIN, E. DEL BARRIO, J. CUESTA-ALBERTOS, AND C. MATRÁN, *Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach*, The Annals of Statistics, 49 (2021), pp. 1139 – 1165.
- [25] J.-C. HÜTTER AND P. RIGOLLET, *Minimax estimation of smooth optimal transport maps*, The Annals of Statistics, 49 (2021), pp. 1166 – 1194.
- [26] J. JURGEN, *Postmodern analysis*, Universitext, Springer, Berlin, 2nd edition ed., 2003.
- [27] O. KAMMAR, *A note on fréchet differentiation under lebesgue integrals*. <http://denotational.co.uk/notes/kammar-a-note-on-frechet-differentiation-under-lebesgue-integrals.pdf>, 2016.
- [28] J. KITAGAWA, Q. MÉRIGOT, AND T. B., *Convergence of a newton algorithm for semi-discrete optimal transport*, Journal of the European Math Society, 21 (2019), pp. 2603–2651.
- [29] A. KOROTIN, V. EGIAZARIAN, A. ASADULAEV, A. SAFIN, AND E. BURNAEV, *Wasserstein-2 generative networks*, in International Conference on Learning Representations, 2021.
- [30] A. MAKKUVA, A. TAGHVAEI, S. OH, AND J. LEE, *Optimal transport mapping via input convex neural networks*, in International Conference on Machine Learning, PMLR, 2020, pp. 6672–6681.
- [31] T. MANOLE, S. BALAKRISHNAN, J. NILES-WEED, AND L. WASSERMAN, *Plugin estimation of smooth optimal transport maps*, 2022.
- [32] S. B. MASUD, M. WERENSKI, J. M. MURPHY, AND S. AERON, *Multivariate soft rank via entropic optimal transport: sample efficiency and generative modeling*, Journal of Machine Learning Research, 24 (2023), pp. 1–65.
- [33] R. J. MCCANN, *Existence and uniqueness of monotone measure-preserving maps*, Duke Mathematical Journal, 80 (1995), pp. 309 – 323.
- [34] B. MUZELLEC, A. VACHER, F. BACH, F.-X. VIALARD, AND A. RUDI, *Near-optimal estimation of smooth transport maps with kernel sums-of-squares*. arXiv, 2021.
- [35] G. PEYRÉ AND M. CUTURI, *Computational optimal transport: With applications to data science*, Foundations and Trends® in Machine Learning, 11 (2019), pp. 355–607.
- [36] A.-A. POOLADIAN AND J. NILES-WEED, *Entropic estimation of optimal transport maps*. arXiv, 2021.
- [37] D. POTTS AND M. SCHMISCHKE, *Approximation of high-dimensional periodic functions with fourier-based methods*, SIAM Journal on Numerical Analysis, 59 (2021), pp. 2393–2429.
- [38] H. ROBBINS AND S. MONRO, *A stochastic approximation method*, The Annals of Mathematical Statistics, 22 (1951), pp. 400–407.
- [39] F. SANTAMBROGIO, *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, Progress in Nonlinear Differential Equations and Their Applications, Springer International Publishing, 2015.
- [40] E. M. STEIN AND G. WEISS, *VII. Multiple Fourier Series*, Princeton University Press, 2016, pp. 245–286.
- [41] J. W. TUKEY, *Mathematics and the picturing of data*, Proceedings of the International Congress of Mathematicians (Vancouver, B. C., 1974), 2 (1975), pp. 523–531.
- [42] C. VILLANI, *Topics in optimal transportation*, vol. 58 of Graduate Studies in Mathematics, American Mathematical Society, 2003.