ESAIM: COCV 29 (2023) 33
https://doi.org/10.1051/cocv/2023019

ESAIM: Control, Optimisation and Calculus of Variations
www.esaim-cocv.org

REGULARIZATION FOR WASSERSTEIN DISTRIBUTIONALLY ROBUST OPTIMIZATION

Waïss Azizian[1], Franck Iutzeler[2] and Jérôme Malick[3,*]

Abstract. Optimal transport has recently proved to be a useful tool in various machine learning applications needing comparisons of probability measures. Among these, applications of distributionally robust optimization naturally involve Wasserstein distances in their models of uncertainty, capturing data shifts or worst-case scenarios. Inspired by the success of the regularization of Wasserstein distances in optimal transport, we study in this paper the regularization of Wasserstein distributionally robust optimization. First, we derive a general strong duality result of regularized Wasserstein distributionally robust problems. Second, we refine this duality result in the case of entropic regularization and provide an approximation result when the regularization parameters vanish.

Mathematics Subject Classification. 90C17, 90C25, 49N15, 49Q22.

Received January 13, 2022. Accepted March 22, 2023.

1. Introduction

Optimal transport (OT) has a long history and exciting recent developments, notably around applications in machine learning and data science; we refer to the monographs [37], [34], [30], and [25]. One of the key technical properties at the core of recent success of OT in these applications is the use of regularization, and specifically entropic regularization, opening the way to efficient computational schemes (see *e.g.*, [14]) to get theoretically-grounded approximations of the Wasserstein distances.

Distributionally robust optimization (DRO) has recently been formulated using OT metrics and has proven to be useful in machine learning (see *e.g.*, [22]). But regularization has still to be studied and used in this context. In this paper, we propose a study of regularization in Wasserstein distributionally robust optimization (WDRO), inspired from several recent developments in OT, namely [18], [11], [17], and [27].

1.1. Distributionally robust optimization with Wasserstein neighborhoods

DRO is a popular approach in optimization under uncertainty. We briefly present here the ideas and the notation that we will use; we refer to the celebrated paper [15] and the survey paper [22] for more details and for applications in machine learning.

Keywords and phrases: Distributionally robust optimization, optimal transport, duality, entropic regularization.

[1] DI, ENS, Univ. PSL, 75005, Paris, France and Univ. Grenoble Alpes, 38000 Grenoble, France.
[2] Univ. Grenoble Alpes, 38000 Grenoble, France.
[3] CNRS & LJK, 38000, Grenoble, France.

* Corresponding author: jerome.malick@univ-grenoble-alpes.fr

© The authors. Published by EDP Sciences, SMAI 2023

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Standard approaches in stochastic optimization consider minimizing the expectation of a random loss with respect to some input distribution or available data points: For an objective $f_\theta : \Xi \to \mathbb{R}$ defined on a sample space Ξ and depending on parameters $\theta \in \Theta$, this consists in considering

$$\min_{\theta\in\Theta} \mathbb{E}_{\xi\sim \mathrm{P}}[f_\theta(\xi)].$$

Here P is a fixed probability distribution on Ξ; in practice, it is typically an empirical distribution $\mathrm{P} = \frac{1}{n}\sum_{i=1}^n \delta_{\xi_i}$ stemming from data samples $(\xi_i)_{i=1}^n$.

A DRO counterpart of this problem is to minimize the expectation of the loss with respect to a set of probability distributions close to P. More precisely, we choose a neighborhood $\mathcal{U}(\mathrm{P})$ of P (called the ambiguity set or the distributional uncertainty region) in the set of probabilities measures on Ξ, denoted by $\mathcal{P}(\Xi)$ and consider the worst possible expectation of the objective in this neighborhood. The resulting problem is thus of the form

$$\min_{\theta\in\Theta} F(\theta) \qquad \text{where} \quad F(\theta) \coloneqq \sup_{\mathrm{Q}\in\mathcal{U}(\mathrm{P})} \mathbb{E}_{\xi\sim\mathrm{Q}}[f_\theta(\xi)]. \tag{1.1}$$

A natural way to define the ambiguity set $\mathcal{U}(\mathrm{P})$ is to consider a ball centered at P with radius $\rho > 0$ controlling the required level of robustness. When using the Wasserstein distance to define the ball, this gives so-called Wasserstein DRO problems (WDRO).

For a cost function $c : \Xi \times \Xi \to \mathbb{R}_+$, the Wasserstein distance between two probability distributions $\mathrm{P}, \mathrm{Q} \in \mathcal{P}(\Xi)$ is defined as

$$W_c(\mathrm{P},\mathrm{Q}) = \inf\left\{\mathbb{E}_{(\xi,\zeta)\sim\pi}\, c(\xi,\zeta) : \pi \in \mathcal{P}(\Xi\times\Xi), \pi_1 = \mathrm{P}, \pi_2 = \mathrm{Q}\right\},$$

where π_1 and π_2 denote the first and second marginals of the coupling probability, or transport plan, π defined on $\Xi \times \Xi$. A WDRO problem thus has the form (1.1) with the ambiguity set

$$\mathcal{U}(\mathrm{P}) = \left\{\mathrm{Q} \in \mathcal{P}(\Xi) : W_c(\mathrm{P},\mathrm{Q}) \le \rho\right\}. \tag{1.2}$$

When the the objective f_θ exhibits a simple structure, this problem reformulates as a tractable convex optimization problem; see *e.g.*, [22]. This is exploited in several applications, for instance: logistic regression (see *e.g.*, [40]), support vector machines (see *e.g.*, [35]), or ℓ^1-regression (see *e.g.*, [12]). Another argument supporting a WDRO approach for machine learning applications is that it provides generalization guarantees, see *e.g.*, [2, 15].

1.2. Contributions, related works, and outline

In this paper we study regularization in the context of Wasserstein distributionally robust optimization. First, we propose a unified framework for double regularization of the WDRO objective function (both in the objective and in the constraint). We then provide a strong duality result with general convex regularization functions. This result can be seen as the analogue for WDRO of the general result of [27] for OT. Second, we refine our analysis in the case of the entropic regularization and obtain an explicit expression for the dual problem. Furthermore, we provide approximation guarantees when the regularization parameters are driven to 0, adapting the reasoning from [11]. These results can be seen as analogues for WDRO of results of [17] for the entropic-regularized OT. Note however, that the reasoning and results used OT do not directly apply in the context of WDRO. Indeed, in OT, the entropic regularization depends on (the product of) the two marginal distributions whereas, in WDRO, the second marginal is not fixed but rather an optimization variable, which makes the analysis different and more involved. This distinction is further discussed in the beginning of Section 3.

Up to our knowledge, regularization in the context of WDRO has not been investigated on its own, as we do in this paper.[1] Nevertheless, let us mention the two recent papers related to our developments: [3] and [39]. In [3], an entropic smoothing of a specific WDRO dual function is introduced and used for computational purposes. Such dual smoothing implicitly corresponds to a regularization of the associated primal problem, but this link is not formally made. In contrast, the preprint [39] (which appeared while we were preparing this manuscript) shares similar spirit as our work. The entropic regularization of WDRO is proposed and analyzed, with a special focus on computational aspects. This is complementary to our work which provides a theoretical study of general regularizations as well as approximation guarantees for the entropic regularization. We will come back more precisely on the connections between our results and those of [3] and [39] in Remark 3.2.

The outline of this paper is the following. The introduction ends below with the definition of the framework of this paper. In Section 2, we provide a duality result for a general double regularization, together with an illustration in the case when the transport cost is used a regularization function. In Section 3, we focus our analysis to the entropic regularization to get refined expressions and an explicit control of the quality of the approximation of the underlying WDRO problem.

1.3. Set-up, notation, and assumptions

The framework of this paper is the following. With Ξ a subset of $\mathbb{R}^d$, P a reference probability distribution over Ξ, and $f\colon \Xi \to \mathbb{R}$ the underlying objective function (we drop the dependence in θ to simplify the notation), we consider the sup problem in the objective function of (1.1) with the Wasserstein ball of radius ρ as an ambiguity set (1.2). Our objective thus writes:

$$\sup\{\mathbb{E}_{\xi\sim\mathrm{Q}} f(\xi) : \mathrm{Q} \in \mathcal{P}(\Xi),\, W_c(\mathrm{P},\mathrm{Q}) \le \rho\}.$$

We reformulate the above problem, in a concise way, using only couplings as

$$\sup_{\pi\in\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi):\ \mathbb{E}_\pi c\le\rho} \mathbb{E}_{\pi_2} f \tag{WDRO}$$

where $\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)$ is the set of probability distributions on $\Xi\times\Xi$ having P as a first marginal

$$\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi) \coloneqq \{\pi \in \mathcal{P}(\Xi\times\Xi) : \pi_1 = \mathrm{P}\}.$$

When the space Ξ is compact, we have that the topological dual of $\mathcal{C}(\Xi\times\Xi)$, the set of continuous functions on $\Xi\times\Xi$, is exactly $\mathcal{M}(\Xi\times\Xi)$, the set of finite signed measures over $\Xi\times\Xi$ by the Riesz representation [32], Theorem 2.14. We denote by $\langle\cdot,\cdot\rangle$ the duality pairing between $\mathcal{C}(\Xi\times\Xi)$ and $\mathcal{M}(\Xi\times\Xi)$:

$$\begin{cases} \mathcal{M}(\Xi\times\Xi)\times\mathcal{C}(\Xi\times\Xi) & \longrightarrow \mathbb{R} \\ (\pi,\varphi) & \longmapsto \langle\pi,\varphi\rangle \coloneqq \int \varphi \,\mathrm{d}\,\pi\,. \end{cases}$$

When establishing general duality results, we will also make a constant use of the convex conjugate of a function $F\colon \mathcal{C}(\Xi\times\Xi) \to \mathbb{R}\cup\{+\infty\}$ defined as

$$F^*\colon \begin{cases} \mathcal{M}(\Xi\times\Xi) & \longrightarrow \mathbb{R}\cup\{+\infty\} \\ \pi & \longmapsto \sup_{\varphi\in\mathcal{C}(\Xi)} \langle\pi,\varphi\rangle - F(\varphi), \end{cases}$$

[1] Let us mention that studying "regularization for WDRO" as we propose here should not be confused with studying "the regularizing effect of WDRO on learning problems", which is a separate field of study (see *e.g.*, [4, 35]).

as well as the preconjugate of a function $G\colon \mathcal{M}(\Xi\times\Xi)\to\mathbb{R}\cup\{+\infty\}$ defined as

$$G_*\colon \begin{cases}\mathcal{C}(\Xi\times\Xi) & \longrightarrow \mathbb{R}\cup\{+\infty\}\\ \varphi & \longmapsto \sup_{\pi\in\mathcal{M}(\Xi)} \langle\pi,\varphi\rangle - G(\pi).\end{cases}$$

In presence of convexity, these two operations are dual one another; see *e.g.*, [13], Remark 5.2. More precisely, we have that $(F^*)_* = F$ when F is lower semi-continuous (l.s.c.), convex, and proper, and $(G_*)^* = G$ when G is weakly-$\star$ l.s.c., convex, and proper. Furthermore, the following duality result will be instrumental in our developments; it is a reformulation [8], Theorem 3.2.6, adapted to our purposes.

Lemma 1.1 (General Fenchel duality)**.** *Consider a compact subset $\mathcal{X}\subset\mathbb{R}^d$, a function $h\in\mathcal{C}(\mathcal{X})$, and two functions $F,G\colon\mathcal{M}(\mathcal{X})\to\mathbb{R}\cup\{+\infty\}$ convex weakly-$\star$ l.s.c. proper. If there exists $\varphi\in\operatorname{dom}F_*\cap(h-\operatorname{dom}G_*)$ such that F_* is continuous at φ, then*

$$\sup_{\pi\in\mathcal{M}(\mathcal{X})} \langle\pi,h\rangle - F(\pi) - G(\pi) = \inf_{\varphi,\psi\in\mathcal{C}(\mathcal{X}):\varphi+\psi=h} F_*(\varphi) + G_*(\psi).$$

Proof. First, the right-hand side (RHS) can be rewritten as

$$\inf_{\varphi,\psi\in\mathcal{C}(\mathcal{X}):\varphi+\psi=h} F_*(\varphi) + G_*(\psi) = \inf_{\varphi\in\mathcal{C}(\mathcal{X})} F_*(\varphi) + G_*(h-\varphi).$$

Then, since F is convex weakly-$\star$ l.s.c., we get that $F=(F_*)^*$ and that F_* is proper, convex, and l.s.c.; see [13], Remark 5.2. The same holds for G.

We can thus apply the duality result from [8], Theorem 3.2.6 with F_* and $G_*(h-\cdot)$ as primal functions that are proper convex functions, and $\mathcal{C}(\mathcal{X})$ as primal space. Indeed, the regularity assumption of the lemma exactly gives the regularity condition (RC_1^{id}) of this result. Hence,

$$\begin{aligned}\inf_{\varphi\in\mathcal{C}(\mathcal{X})} F_*(\varphi) + G_*(h-\varphi) &= \sup_{\pi\in\mathcal{M}(\mathcal{X})} -(F_*)^*(-\pi) - (G_*(h-\cdot))^*(\pi)\\ &= \sup_{\pi\in\mathcal{M}(\mathcal{X})} -F(-\pi) - G(-\pi) - \langle\pi,h\rangle.\end{aligned}$$

Carrying out the change of variable $\pi\leftarrow-\pi$ then concludes the proof. □

2. Regularization of the WDRO objective function

In this section, we study (WDRO) objectives with additional regularization functions both in the constraints $\mathbb{E}_\pi c\le\rho$ and in the objective $\mathbb{E}_{\pi_2} f$. For two arbitrary convex functions $R,S:\mathcal{M}(\Xi\times\Xi)\to\mathbb{R}\cup\{+\infty\}$, the regularized objective we consider is

$$\sup_{\pi\in\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi):\ \mathbb{E}_\pi c + S(\pi)\le\rho} \mathbb{E}_{\pi_2} f - R(\pi). \tag{R-WDRO}$$

We give in Section 2.1 the expression of the dual of this problem, under some compactness assumptions. We specialize in Section 2.2 the obtained expression for two specific regularizations, namely the cost regularization and the entropic one. Finally, we present in Section 2.3 a general framework, beyond any compactness assumption, where the dual expression still holds.

2.1. Duality for regularized WDRO

In this section, we give the expression of the dual of (R-WDRO). This result is the analogue for WDRO of the main result of [27] which gives the dual of the OT problem regularized by a general convex function. We get this dual expression under the following assumptions, providing a nice duality between objects.

Assumption 2.1.

(i) $\Xi \subset \mathbb{R}^d$ is convex and compact;

(ii) $f\colon \Xi \to \mathbb{R}$ and $c\colon \Xi \times \Xi \to \mathbb{R}_+$ are both continuous;

(iii) For all ξ in Ξ, $c(\xi, \xi) = 0$.

This formulation of the dual problem can also be seen as a generalization of the existing one for the non-regularized case ($R = S = 0$), see *e.g.*, [5, 16]. Note however that the duality results in these papers rely on weaker assumptions; in particular, Ξ is not assumed to be compact and f, c are only upper- and lower-semicontinuous respectively. In the following result, these assumptions are needed to handle general regularizations. We discuss how to alleviate them for particular regularization functions in the next section.

Theorem 2.2 (Strong duality for doubly-regularized WDRO)**.** *Let Assumption 2.1 hold and take two convex, proper, and weakly-$\star$ l.s.c. functions $R, S\colon \mathcal{M}(\Xi \times \Xi) \to \mathbb{R} \cup \{+\infty\}$, such that $R + S$ is also proper. If the primal problem (R-WDRO) is strictly feasible (i.e., if there exists $\pi \in \mathcal{P}_{\mathrm{P}}(\Xi \times \Xi) \cap \operatorname{dom} R$ such that $\mathbb{E}_\pi c + S(\pi) < \rho$), then we have*

$$\operatorname{val}(\text{R-WDRO}) \;=\; \inf_{\lambda \geq 0} \inf_{\varphi \in \mathcal{C}(\Xi \times \Xi)} \lambda\rho + \mathbb{E}_{\xi \sim \mathrm{P}}\left[\sup_{\zeta \in \Xi} f(\zeta) - \lambda c(\xi, \zeta) - \varphi(\xi, \zeta)\right] + (R + \lambda S)_*(\varphi), \tag{2.1}$$

and there exists a primal optimal solution $\pi^\star \in \mathcal{P}_{\mathrm{P}}(\Xi \times \Xi)$ and an optimal dual parameter $\lambda^\star \geq 0$.

We prove this theorem by carefully combining two standard duality results (the Lagrangian duality theorem in Banach spaces [29] and the Fenchel duality theorem recalled in Lemma 1.1) with a powerful theorem for exchanging minimization/integration [31]. The latter is rarely considered in similar contexts; see [36], Theorem 5 for an exception.

Proof. The first step consists in applying the Lagrangian duality theorem of [29], Theorem 3.68; let us check its assumptions. First note that Slater's condition holds by assumption, so we only need to check that the primal problem has at least a solution. This is the case by the following arguments: (i) the problem is feasible by assumption; (ii) $\mathcal{P}(\Xi \times \Xi)$ is weakly-$\star$ sequentially compact (see *e.g.*, [10], Cor. 3.30); (iii) the constraint set $\{\pi \in \mathcal{P}_{\mathrm{P}}(\Xi \times \Xi) : \mathbb{E}_\pi c + S(\pi) \leq \rho\}$ is weakly-$\star$ closed (since S is weakly-$\star$ l.s.c. and the constraint $\pi_1 = \mathrm{P}$ is weakly-$\star$ closed); and (iv) the objective $\pi \mapsto \mathbb{E}_{(\xi,\zeta)\sim\pi}[f(\zeta) - \lambda c(\xi, \zeta)] - R(\pi)$ is weakly-$\star$ upper semi-continuous (u.s.c.) by assumption. As a result, we have Lagrangian duality and existence of a dual solution

$$\operatorname{val}(\text{R-WDRO}) = \inf_{\lambda \geq 0} \sup_{\pi \in \mathcal{P}(\Xi \times \Xi) : \pi_1 = \mathrm{P}} \mathbb{E}_{(\xi,\zeta)\sim\pi}[f(\zeta) - \lambda c(\xi, \zeta)] - (R + \lambda S)(\pi) + \lambda\rho \tag{2.2}$$

The next step is to write the inner sup as an inf. For concision, let us introduce $T_\lambda := R + \lambda S$ and $F_\lambda(\xi, \zeta) := f(\zeta) - \lambda c(\xi, \zeta)$. The sup over π, for a fixed $\lambda \geq 0$, thus writes

$$\sup_{\pi \in \mathcal{P}(\Xi \times \Xi) : \pi_1 = \mathrm{P}} \mathbb{E}_{(\xi,\zeta)\sim\pi}[f(\zeta) - \lambda c(\xi, \zeta)] - T_\lambda(\pi) = \sup_{\pi \in \mathcal{M}(\Xi \times \Xi)} \langle \pi, F_\lambda \rangle - \iota_{\mathcal{P}_{\mathrm{P}}(\Xi \times \Xi)}(\pi) - T_\lambda(\pi)$$

where $\iota_{\mathcal{P}_{\mathrm{P}}(\Xi \times \Xi)}$ is the indicator function in the sense of convex analysis, *i.e.*, for $\pi \in \mathcal{M}(\Xi \times \Xi)$, $\iota_{\mathcal{P}_{\mathrm{P}}(\Xi \times \Xi)}(\pi) = 0$ if $\pi \in \mathcal{P}(\Xi \times \Xi)$ and $\pi_1 = \mathrm{P}$, and $+\infty$ otherwise.

Now, we want to apply the duality result of Lemma 1.1 with $F \leftarrow \iota_{\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)}$, $G \leftarrow\ T_\lambda$, and $h \leftarrow F_\lambda$. To do so, we need to derive the (pre)conjugates of $\iota_{\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)}$ and T_λ. By the disintegration theorem, any coupling $\pi(\mathrm{d}\xi, \mathrm{d}\zeta)$ can be written as $\mathrm{P}(\mathrm{d}\xi)\,\mathrm{Q}(\mathrm{d}\zeta|\xi)$ with Q a conditional probability on Ξ. Therefore,

$$\begin{aligned}(\iota_{\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)})_*(\varphi) &= \sup\left\{\mathbb{E}_{(\xi,\zeta)\sim\pi}\varphi(\xi,\zeta) : \pi \in \mathcal{P}(\Xi\times\Xi),\, \pi_1 = \mathrm{P}\right\}\\ &= \sup\left\{\mathbb{E}_{\xi\sim\mathrm{P}}\mathbb{E}_{\zeta\sim\mathrm{Q}(\cdot|\xi)}\varphi(\xi,\zeta) : \mathrm{Q}(\cdot|\cdot) \text{ conditional probability on } \Xi\right\}\\ &\leq \mathbb{E}_{\xi\sim\mathrm{P}}\left[\sup_{\zeta\in\Xi}\varphi(\xi,\zeta)\right].\end{aligned}$$

We now proceed to the reverse inequality, noting that a measurable map $\zeta\colon \Xi \to \Xi$ induces a conditional probability $\mathrm{Q}(\cdot|\xi) = \delta_{\zeta(\xi)}$. Hence,

$$\begin{aligned}(\iota_{\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)})_*(\varphi) &= \sup\left\{\mathbb{E}_{\xi\sim\mathrm{P}}\mathbb{E}_{\zeta\sim\mathrm{Q}(\cdot|\xi)}\varphi(\xi,\zeta) : \mathrm{Q}(\cdot|\cdot) \text{ conditional probability on } \Xi\right\}\\ &\geq \sup\left\{\mathbb{E}_{\xi\sim\mathrm{P}}\varphi(\xi,\zeta(\xi)) : \zeta\colon \Xi\to\Xi \text{ measurable}\right\}.\end{aligned}$$

The sup is finite since Ξ is compact and φ continuous. Let us define $\mathsf{N}\colon \Xi\times\mathbb{R}^d \longrightarrow \mathbb{R}\cup\{+\infty\}$ defined as $\mathsf{N}(\xi,\zeta) = -\varphi(\xi,\zeta)$ if $\zeta\in\Xi$ and $+\infty$ otherwise. Since φ is continuous and Ξ is closed, N is jointly l.s.c., and, as a consequence, it is a normal integrand by [31], Example 14.31. So we can apply [31], Theorem 14.60 to get that

$$\inf\left\{\mathbb{E}_{\xi\sim\mathrm{P}} - \varphi(\xi,\zeta(\xi)) : \zeta\colon \Xi\to\Xi \text{ measurable}\right\} = \mathbb{E}_{\xi\sim\mathrm{P}}\left[\inf_{\zeta\in\Xi} -\varphi(\xi,\zeta)\right].$$

Inverting the signs, we have showed both upper and lower-inequalities, which means that

$$(\iota_{\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)})_*(\varphi) = \mathbb{E}_{\xi\sim\mathrm{P}}\left[\sup_{\zeta\in\Xi}\varphi(\xi,\zeta)\right] \qquad \text{for any } \varphi\in\mathcal{C}(\Xi\times\Xi),$$

and thus $\mathrm{dom}(\iota_{\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)})_* = \mathcal{C}(\Xi\times\Xi)$. Also, $\varphi\in\mathcal{C}(\Xi\times\Xi)\mapsto\sup_{\zeta\in\Xi}\varphi(\cdot,\zeta)\in\mathcal{C}(\Xi)$ is 1-Lipschitz w.r.t. the norm of the uniform convergence so $(\iota_{\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)})_*$ is continuous on its domain.

Finally, since T_λ is convex, proper and weakly-$\star$ l.s.c., $T_{\lambda*}$ is proper and therefore $\mathrm{dom}\,T_* \neq \emptyset$. Thus, Lemma 1.1 can be used and gives that

$$\sup_{\pi\in\mathcal{P}(\Xi\times\Xi):\pi_1=\mathrm{P}} \mathbb{E}_{(\xi,\zeta)\sim\pi}[f(\zeta) - \lambda c(\xi,\zeta)] - T(\pi) = \inf_{\varphi\in\mathcal{C}(\Xi\times\Xi)} \mathbb{E}_{\xi\sim\mathrm{P}}\sup_{\zeta\in\Xi} f(\zeta) - \lambda c(\xi,\zeta) - \varphi(\xi,\zeta) + T_{\lambda*}(\varphi),$$

which, combined with (2.2), leads to the claimed result. □

2.2. Examples of regularized WDRO

As an illustration of the duality result, let us consider two cases: first, when the regularization is the transport cost itself and, second, when it is a ϕ-divergence.

When the transport cost itself is used a regularization, the expression of the dual simplifies as follow. This expression will be used in the analysis of the next section.

Corollary 2.3 (Duality for cost-regularized WDRO)**.** *Let Assumption 2.1 hold and take $\varepsilon, \delta > 0$. Then we have*

$$\sup_{\pi\in\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)\,:\,\mathbb{E}_\pi c+\delta\,\mathbb{E}_\pi c\leq\rho} \mathbb{E}_{\pi_2} f - \varepsilon\,\mathbb{E}_\pi c \;=\; \inf_{\lambda\geq 0} \lambda\rho + \mathbb{E}_{\xi\sim\mathrm{P}}\sup_{\zeta\in\Xi}\left\{f(\zeta) - \big(\varepsilon + \big(1+\delta\big)\lambda\big)c(\xi,\zeta)\right\}.$$

Proof. Define $R : \pi \in \mathcal{M}(\Xi \times \Xi) \mapsto \varepsilon\langle \pi, c\rangle$, $S : \pi \in \mathcal{M}(\Xi \times \Xi) \mapsto \delta\langle \pi, c\rangle$ which are convex, proper and weakly-$\star$ continuous by construction. The preconjugate of their sum is $(R + \lambda S)_* = \iota_{\{(\varepsilon+\lambda\delta)c\}}$. Moreover, the primal is strictly feasible thanks to the transport plan $\pi(\mathrm{d}\xi, \mathrm{d}\zeta) = \mathrm{P}(\mathrm{d}\xi)\,\delta_\xi(\mathrm{d}\zeta)$. Thus, we can apply Theorem 2.2 to get the expression. □

When ϕ-divergences are used as regularizations, the expression of the dual problem of (2.1) simplifies as follows. We will come back in more details in Section 3 on the KL-divergence, which is a popular ϕ-divergence.

Corollary 2.4 (Duality for ϕ-divergence-regularized WDRO)**.** *Let Assumption 2.1 hold and take $\varepsilon, \delta \geq 0$. Consider $\pi_0 \in \mathcal{P}(\Xi \times \Xi)$, a convex l.s.c. function $\phi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ such that $\phi(1) = 0$ and $\phi'(\pm\infty) = \pm\infty$ and define the associated divergence for all $\pi \in \mathcal{P}(\Xi \times \Xi)$ as*

$$D_\phi(\pi\,|\pi_0) := \begin{cases} \int_{\Xi^2} \phi\left(\frac{\mathrm{d}\,\pi}{\mathrm{d}\pi_0}\right)\mathrm{d}\pi_0 & \text{if } \pi \text{ is absolutely continuous w.r.t. } \pi_0 \\ +\infty & \text{otherwise.} \end{cases}$$

Then, with $R \leftarrow \varepsilon D_\phi(\,\cdot\,|\pi_0)$ and $S \leftarrow \delta D_\phi(\,\cdot\,|\pi_0)$, if the primal (R-WDRO) is strictly feasible, its value is equal to

$$\inf_{\lambda \geq 0} \inf_{\psi \in \mathcal{C}(\Xi\times\Xi)} \lambda\rho + \mathbb{E}_{\xi\sim\mathrm{P}}\left[\sup_{\zeta\in\Xi} f(\zeta) - \lambda c(\xi,\zeta) - \psi(\xi,\zeta)\right] + (\varepsilon + \lambda\delta)\int_{\Xi^2} \phi^*\left(\frac{\psi(\xi,\zeta)}{\varepsilon+\lambda\delta}\right)\mathrm{d}\pi_0(\xi,\zeta).$$

Proof. R and S are convex and proper by construction. Furthermore, D_ϕ admits the following variational formula, see *e.g.*, [1], Proposition 4.2.8 applied with the pair $(L^\infty(\pi_0), \mathcal{M}(\Xi \times \Xi))$,

$$\forall\, \pi \in \mathcal{M}(\Xi \times \Xi),\ D_\phi(\pi\,|\pi_0) = \sup\left\{\langle \pi, \psi\rangle - \int_{\Xi^2} \phi^* \circ \psi\ \mathrm{d}\pi_0 : \psi \in L^\infty(\pi_0)\right\}.$$

By Lusin's theorem, see *e.g.*, [32], Theorem 2.24, the maximization over $L^\infty(\pi_0)$ can be replaced by maximization over $\mathcal{C}(\Xi \times \Xi)$, which guarantees that D_ϕ is indeed weak-$\star$ l.s.c.. Then, we can apply Theorem 2.2, using that the preconjugate of D_ϕ is exactly $\psi \mapsto \int_{\Xi^2} \phi^* \circ \psi\ \mathrm{d}\pi_0$ by [1], Proposition 4.2.6. □

2.3. Duality under weaker assumptions

In this section, we present a generalization of Theorem 2.2 with weaker assumptions. The main change is that the compactness of Ξ is no longer required at the expense of a growth condition on f. Specifically, Assumption 2.1 is replaced by the following set of assumptions.

Assumption 2.5.

(i) $\Xi \subset \mathbb{R}^d$ is a closed set;

(ii) $f : \Xi \to \mathbb{R}$ is measurable w.r.t. the Borel σ-algebra on Ξ;

(iii) $c : \Xi \times \Xi \to \mathbb{R}_+$ is continuous and, for all ξ in Ξ, $c(\xi, \xi) = 0$;

(iv) There is some $\xi_0 \in \Xi$ such that f satisfies the growth condition[2]:

$$\exists\, \alpha > 0,\ \text{ s.t.}\, \forall \xi, \zeta \in \Xi, \quad |f(\zeta)| \leq \alpha(1 + c(\xi, \xi_0) + c(\xi, \zeta)),$$

and the integrability condition $\mathbb{E}_{\xi\sim\mathrm{P}}[c(\xi, \xi_0)] < +\infty$ holds.

[2] In the case when c is a power of a distance $c = d^p$, the growth condition is equivalent to $|f(\xi)| \leq \alpha(1 + d(\xi, \xi_0)^p)$, which is standard in unregularized WDRO, see *e.g.*, [6, 16]

For the sake of readability, we only state here an informal version of our result, where we omit some technical assumptions on the regularizers and the construction of the spaces involved. The formal statement (Thm. A.8), along with its proof, is provided in Appendix A.

Theorem 2.6 (Strong duality for general doubly-regularized WDRO, informal)**.** *Let Assumption 2.5 hold and take two convex, proper extended-valued functions R, S which satisfy some regularity conditions. If the primal problem (R-WDRO) is strictly feasible, then we have*

$$\operatorname{val}(\text{R-WDRO}) = \inf_{\lambda\geq 0}\inf_{\varphi\in\mathcal{X}} \lambda\rho + \mathbb{E}_{\xi\sim \mathrm{P}}\left[\operatorname*{ess\,sup}_{\zeta\in\Xi} f(\zeta) - \lambda c(\xi,\zeta) - \varphi(\xi,\zeta)\right] + (R|_{\mathcal{X}^*} + \lambda S|_{\mathcal{X}^*})_*(\varphi), \tag{2.3}$$

where $\mathcal{X}$ is a Banach function space built from c, R and S, and the essential supremum is taken w.r.t. the Lebesgue measure on Ξ.

The gist of the proof consists in carefully crafting $\mathcal{X}$ from c, R and S and then taking advantage of the duality structure of the pair $(\mathcal{X}, \mathcal{X}^*)$. Up to the function spaces involved, the dual expression of (2.3) corresponds to the one of (2.1). Also, the regularity conditions on R and S encompass the examples of Section 2.2, *i.e.*, the cost regularization and ϕ-divergences. Note also that the continuity assumption on the cost function c can be removed in the case of ϕ-divergences; see Appendix A.

3. Entropic regularization

In this section, we specialize and refine the study of the previous section in the case entropic regularization, *i.e.*, when the Kullback-Leibler (KL) divergence is used a regularizing function. This regularization is defined for two signed measures with finite variations μ, ν as

$$\mathrm{KL}(\mu|\nu) = \begin{cases} \int \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\,\mathrm{d}\mu & \text{if } \mu \text{ and } \nu \text{ are non-negative and } \mu \ll \nu \\ +\infty & \text{otherwise} \end{cases}.$$

This kind of regularization is very popular in computational OT, where it enables the derivation of useful approximations of the Wasserstein distance, as for example the so-called Sinkhorn distance:

$$\inf_{\pi\in\mathcal{P}(\Xi\times\Xi):\pi_1=\mathrm{P},\pi_2=\mathrm{Q}} \mathbb{E}_\pi c + \varepsilon\,\mathrm{KL}(\pi\,|\,\mathrm{P}\otimes\mathrm{Q}) \qquad \text{for } \mathrm{P},\mathrm{Q}\in\mathcal{P}(\Xi) \text{ given.} \tag{3.1}$$

When Ξ is compact, the dual of this problem is given in [18], Proposition 2.1

$$\sup_{f\in\mathcal{C}(\Xi)} \mathbb{E}_{\mathrm{P}} f - \varepsilon\,\mathbb{E}_{\xi\sim\mathrm{P}} \log\Big(\mathbb{E}_{\zeta\sim\mathrm{Q}} e^{\frac{f(\xi)-c(\xi,\zeta)}{\varepsilon}}\Big). \tag{3.2}$$

In Section 3.1, we establish a similar result for WDRO. But let us point out here a technical difficulty arising from the WDRO framework compared to OT. The KL divergence (3.1) is taken w.r.t. the measure $\mathrm{P}\otimes\mathrm{Q}$, which does not restrict the set of feasible transport plans (if $\pi\in\mathcal{P}(\Xi\times\Xi)$ satisfies $\pi_1 = \mathrm{P}$ and $\pi_2 = \mathrm{Q}$, π is indeed absolutely continuous w.r.t. $\mathrm{P}\otimes\mathrm{Q}$). In WDRO however, only one marginal of the transport plans π is fixed and the support of the optimal coupling can be arbitrary, so that the same regularization as in (3.1) cannot directly be used.

We thus propose to regularize (WDRO) with KL using a base coupling π_0 with first marginal $(\pi_0)_1 = \mathrm{P}$ and consider

$$\sup_{\pi\in\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi):\ \mathbb{E}_\pi c+\delta\,\mathrm{KL}(\pi\,|\pi_0)\leq\rho} \mathbb{E}_{\pi_2} f - \varepsilon\,\mathrm{KL}(\pi\,|\pi_0). \tag{E-WDRO}$$

The choice of π_0 restricts the set of transport plans to those that are absolutely continuous w.r.t. π_0. We see in Section 3.2 that this restriction has a very limited impact since a natural choice of π_0 still provides a good approximations of the original problem (WDRO) when the regularization parameters ε, δ vanish.

3.1. Duality for the entropy-regularized problem

We derive here a duality theorem for (E-WDRO) involving the KL regularization with an arbitrary base coupling $\pi_0 \in \mathcal{P}_{\mathrm{P}}(\Xi \times \Xi)$. The result naturally involves the same features as in (3.2).

Theorem 3.1 (Strong duality for entropy-regularized problems)**.** *Let Assumption 2.1 hold, take $\varepsilon, \delta > 0$, and fix an arbitrary $\pi_0 \in \mathcal{P}_{\mathrm{P}}(\Xi \times \Xi)$. If the primal problem (E-WDRO) is strictly feasible (i.e., if there exists π such that $\mathbb{E}_\pi c + \delta \,\mathrm{KL}(\pi \,|\pi_0) < \rho$), then*

$$\mathrm{val}\,(\text{E-WDRO}) \;=\; \inf_{\lambda \geq 0} \lambda\rho + (\varepsilon + \lambda\delta)\mathbb{E}_{\xi\sim \mathrm{P}} \log\left(\mathbb{E}_{\zeta\sim\pi_0(\cdot|\xi)} e^{\frac{f(\zeta)-\lambda c(\xi,\zeta)}{\varepsilon+\lambda\delta}}\right), \tag{3.3}$$

and there exists a primal optimal solution $\pi^\star \in \mathcal{P}_{\mathrm{P}}(\Xi \times \Xi)$ and an optimal dual parameter $\lambda^\star \geq 0$.

The interest of using the entropy as a regularization appears when comparing (3.3) to the general dual (2.1). We see that there is no inf on φ and moreover the inner sup is replaced by a smoothed approximation (of the log-sum-exp type).

Remark 3.2 (Two related results and weaker assumptions)**.** The expression (3.3) already appeared, for special cases, in recent articles/preprints. First, the dual problem in (3.3) is proposed in [3] as a smoothing technique (independently from duality and KL-regularization) in a specific context of semi-supervised learning over a finite set Ξ. Second, a very similar duality result (in the case of the KL-regularization in constraints only) appears in [39], Theorem 1. We note that this duality result holds under weaker assumptions: their proof is based on arguments specific to the KL divergence, whereas the one of Theorem 3.1 is based on Corollary 2.4 for ϕ-divergences and thus on Theorem 2.2 for generic regularizations.

It is possible to weaken the assumptions of Theorem 3.1 by using the general Theorem 2.6 instead of Theorem 2.2, which essentially replaces Assumption 2.1 by Assumption 2.5. We could also weaken the assumptions by using an alternative proof from duality formulas in variational inference; see [23], Theorem 2.1 which is itself inspired by [9], Lemma 1. We do not provide the details of these two refinements here. Indeed, obtaining duality under weaker assumptions is not what matters most here since the approximation result of the next section requires a combination of compactness and continuity. □

Proof. We start by applying Corollary 2.4 with $\phi : x \in \mathbb{R}_+ \mapsto x\log x - x + 1$ (whose conjugate is $\phi^*(y) = e^y - 1$), to get

$$\mathrm{val}\,(\text{E-WDRO}) = \inf_{\lambda\geq 0} \lambda\rho + \underbrace{\inf_{\varphi\in\mathcal{C}(\Xi\times\Xi)} \mathbb{E}_{\xi\sim\mathrm{P}}\left[\sup_{\zeta\in\Xi} f(\zeta) - \lambda c(\xi,\zeta) - \varphi(\xi,\zeta)\right] + (\varepsilon+\lambda\delta)\langle \pi_0, e^{\frac{\varphi}{\varepsilon+\lambda\delta}} - 1\rangle}_{(a)}.$$

For a fixed $\lambda \geq 0$, we can simplify the expression of the term (a) above by introducing $\tau \coloneqq \varepsilon + \lambda\delta$, defining $F_\lambda : (\xi,\zeta) \mapsto f(\zeta) - \lambda c(\xi,\zeta) \in \mathcal{C}(\Xi\times\Xi)$, and carrying out the change of variable $\varphi \leftarrow F_\lambda - \varphi$. We thus obtain

$$\begin{aligned}(a) &= \inf_{\varphi\in\mathcal{C}(\Xi\times\Xi)} \mathbb{E}_{\xi\sim\mathrm{P}}\left[\sup_{\zeta\in\Xi} f(\zeta) - \lambda c(\xi,\zeta) - \varphi(\xi,\zeta)\right] + \tau\,\mathbb{E}_{\pi_0}\left[e^{\frac{\varphi}{\tau}} - 1\right] \\ &= \inf_{\varphi\in\mathcal{C}(\Xi\times\Xi)} \mathbb{E}_{\xi\sim\mathrm{P}}\left[\sup_{\zeta\in\Xi} \varphi(\xi,\zeta)\right] + \tau\,\mathbb{E}_{\pi_0}\left[e^{\frac{F_\lambda-\varphi}{\tau}} - 1\right].\end{aligned} \tag{3.4}$$

The rest of the proof is devoted to the reformulation of the term (a) expressed as (3.4). In order to get rid of the supremum in this expression, we restrict the minimization to $\mathcal{C}(\Xi)$ instead of $\mathcal{C}(\Xi \times \Xi)$, *i.e.*, we consider functions $\varphi \in \mathcal{C}(\Xi \times \Xi)$ of the form $\varphi(\xi, \zeta) = g(\xi)$ with $g \in \mathcal{C}(\xi)$:

$$\text{val}\,(3.4) \leq \inf_{g \in \mathcal{C}(\Xi)} \mathbb{E}_{\xi \sim \mathrm{P}}\left[g(\xi)\right] + \tau \mathbb{E}_{(\xi,\zeta)\sim\pi_0}\left[e^{\frac{F_\lambda(\xi,\zeta) - g(\xi)}{\tau}} - 1\right]. \tag{3.5}$$

Let us prove that the inequality above is actually an equality. Fix a bivariate function $\varphi \in \mathcal{C}(\Xi \times \Xi)$ and consider the univariate function $g^\varphi := \sup_{\zeta \in \Xi} \varphi(\cdot, \zeta)$. Then, since $-\varphi(\xi, \zeta) \geq -g^\varphi(\xi)$, we have

$$\mathbb{E}_{\xi \sim \mathrm{P}}\left[\sup_{\zeta \in \Xi} \varphi(\xi, \zeta)\right] + \tau \mathbb{E}_{\pi_0}\left[e^{\frac{F_\lambda - \varphi}{\tau}} - 1\right] \geq \mathbb{E}_{\xi \sim \mathrm{P}}\left[g^\varphi(\xi)\right] + \tau \mathbb{E}_{(\xi,\zeta)\sim\pi_0}\left[e^{\frac{F_\lambda(\xi,\zeta) - g^\varphi(\xi)}{\tau}} - 1\right]. \tag{3.6}$$

However, g^φ is not continuous in general but only l.s.c., hence we cannot lower bound the RHS of (3.6) by the RHS of (3.5). To remedy this issue, we approximate g^φ with continuous functions defined for $k \geq 1$ by $g_k^\varphi(\xi) := \inf_{\zeta \in \Xi} g^\varphi(\zeta) + k\|\xi - \zeta\|$. Since g^φ is l.s.c., the functions (g_k^φ) converge pointwise to g^φ when k goes to $+\infty$, see *e.g.*, [31], Example 9.11. Moreover, these functions are uniformly bounded by $\sup_{\xi \in \Xi} |g^\varphi(\xi)|$, thus Lebesgue's dominated convergence implies that

$$\begin{aligned}
\mathbb{E}_{\xi \sim \mathrm{P}}\left[\sup_{\zeta \in \Xi} \varphi(\xi, \zeta)\right] + \tau \mathbb{E}_{\pi_0}\left[e^{\frac{F_\lambda - \varphi}{\tau}} - 1\right] &\geq \mathbb{E}_{\xi \sim \mathrm{P}}\left[g^\varphi(\xi)\right] + \tau \mathbb{E}_{(\xi,\zeta)\sim\pi_0}\left[e^{\frac{F_\lambda(\xi,\zeta) - g^\varphi(\xi)}{\tau}} - 1\right] \\
&= \lim_{k \to +\infty} \mathbb{E}_{\xi \sim \mathrm{P}}\left[g_k^\varphi(\xi)\right] + \tau \mathbb{E}_{(\xi,\zeta)\sim\pi_0}\left[e^{\frac{F_\lambda(\xi,\zeta) - g_k^\varphi(\xi)}{\tau}} - 1\right] \\
&\geq \inf_{g \in \mathcal{C}(\Xi)} \mathbb{E}_{\xi \sim \mathrm{P}}\left[g(\xi)\right] + \tau \mathbb{E}_{(\xi,\zeta)\sim\pi_0}\left[e^{\frac{F_\lambda(\xi,\zeta) - g(\xi)}{\tau}} - 1\right].
\end{aligned} \tag{3.7}$$

Combining (3.5) and (3.7) gives that

$$(a) = \inf_{g \in \mathcal{C}(\Xi)} \mathbb{E}_{\xi \sim \mathrm{P}}\left[g(\xi)\right] + \tau \mathbb{E}_{(\xi,\zeta)\sim\pi_0}\left[e^{\frac{F_\lambda(\xi,\zeta) - g(\xi)}{\tau}} - 1\right] \tag{3.8}$$

The final step of the proof consists in solving the above minimum over g. Indeed, the objective

$$g \mapsto \mathbb{E}_{\xi \sim \mathrm{P}}\left[g(\xi)\right] + \tau \mathbb{E}_{(\xi,\zeta)\sim\pi_0}\left[e^{\frac{F_\lambda(\xi,\zeta) - g(\xi)}{\tau}} - 1\right]$$

is convex and differentiable on $\mathcal{C}(\Xi)$. As a consequence, the critical points are minimizers. The gradient at a continuous function $g \in \mathcal{C}(\Xi)$ lives in $\mathcal{M}(\Xi)$ and is given by

$$\mathrm{P}(\mathrm{d}\xi) - \left(\int_\Xi e^{\frac{F_\lambda(\xi,\zeta) - g(\xi)}{\tau}} \pi_0(\mathrm{d}\zeta|\xi)\right) \mathrm{P}(\mathrm{d}\xi). \tag{3.9}$$

Thus the continuous function

$$g^\star : \xi \mapsto \tau \log\left(\int_\Xi e^{\frac{F_\lambda(\xi,\zeta)}{\tau}} \pi_0(\mathrm{d}\zeta|\xi)\right)$$

is a solution of (3.8). We get

$$(a) \;=\; \mathbb{E}_{\xi\sim \mathrm{P}}\left[\tau \log\left(\int_\Xi e^{\frac{F_\lambda(\xi,\zeta)}{\tau}}\pi_0(\mathrm{d}\zeta|\xi)\right)\right] + \underbrace{\tau\mathbb{E}_{\xi\sim \mathrm{P}}\left[\int_\Xi e^{\frac{F_\lambda(\xi,\zeta)-g^\star(\xi)}{\tau}}\pi_0(\mathrm{d}\zeta|\xi) - 1\right]}_{=0 \quad \text{(by nullity of the gradient in (3.9) for } g^\star)}$$

which in turn gives the desired expression (3.3). □

Remark 3.3 (Optimal transport plan)**.** In the setting of Theorem 3.1, the primal optimal solution $\pi^\star \in \mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)$ can be built explicitly from $\lambda^\star$ by taking for any $\xi\in\Xi$

$$\pi^\star(\mathrm{d}\zeta|\xi) \propto e^{\frac{f(\zeta)-\lambda^\star c(\xi,\zeta)}{\varepsilon+\lambda\delta}}\pi_0(\mathrm{d}\zeta|\xi).$$

Indeed, since the dual function

$$\lambda \mapsto \lambda\rho + (\varepsilon+\lambda\delta)\mathbb{E}_{\xi\sim\mathrm{P}}\log\left(\mathbb{E}_{\zeta\sim\pi_0(\cdot|\xi)}e^{\frac{f(\zeta)-\lambda c(\xi,\zeta)}{\varepsilon+\lambda\delta}}\right)$$

is differentiable at $\lambda^\star$, the optimality condition for (3.3) implies that $\mathbb{E}_{\pi^\star}[c] + \delta\,\mathrm{KL}(\pi^\star|\pi_0) \le \rho$ (with equality if and only if $\lambda^\star > 0$), *i.e.*, that $\pi^\star$ is feasible in (E-WDRO). Then, one can check that

$$\mathbb{E}_{\pi_2^*}[f] - \varepsilon\,\mathrm{KL}(\pi^\star|\pi_0) = \lambda^\star\rho + (\varepsilon+\lambda^\star\delta)\mathbb{E}_{\xi\sim\mathrm{P}}\log\left(\mathbb{E}_{\zeta\sim\pi_0(\cdot|\xi)}e^{\frac{f(\zeta)-\lambda^\star c(\xi,\zeta)}{\varepsilon+\lambda^\star\delta}}\right)$$

so that, by strong duality (Thm. 3.1), $\pi^\star$ is optimal for (E-WDRO).

3.2. Approximation error of entropy-regularized problems

In this section, we study the behavior of the approximation error of (E-WDRO) as the regularization parameters vanish to 0. To quantify this approximation, we specify the cost c and the reference measure π_0. Specifically, we consider that the cost c is a norm to some power $p \ge 1$

$$c(\xi,\zeta) = \|\xi-\zeta\|^p \tag{3.10}$$

and that the reference measure $\pi_0 \in \mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)$ is taken, for some $\sigma > 0$, as

$$\pi_0(\mathrm{d}\xi,\mathrm{d}\zeta) \;\propto\; \mathrm{P}(\mathrm{d}\xi)\,\mathbb{1}_{\zeta\in\Xi}\,e^{-\frac{c(\xi,\zeta)}{2^{p-1}\sigma}}\mathrm{d}\zeta. \tag{3.11}$$

For example, when $c(\xi,\zeta) = \|\xi-\zeta\|_2^p$ with $p\in\{1,2\}$, $\pi_0(\cdot|\xi)$ is a Laplace or a Gaussian distribution (which is easy to sample from for any $\xi\in\Xi$). We also slightly strengthen Assumption 2.1 by assuming that Ξ is a convex body and that the functions are Lipschitz continuous.

Theorem 3.4 (Approximation for entropic regularization)**.** *Let the following conditions hold:*

(i) *the objective $f\colon \Xi\to\mathbb{R}$ and the cost $c\colon \Xi\times\Xi\to\mathbb{R}_+$ are Lipschitz continuous;*
(ii) *the cost c and the coupling π_0 are taken as (3.10) and (3.11) with $\sigma>0$ such that $\mathbb{E}_{\pi_0}c < \rho$;*
(iii) *the set $\Xi\subset\mathbb{R}^d$ is compact, convex, with nonempty interior.*

Then, as the regularization parameters $\varepsilon,\delta>0$ go to zero, we have

$$0 \le \mathrm{val}\,(\mathrm{WDRO}) - \mathrm{val}\,(\text{E-WDRO}) \le \mathcal{O}\left(d\,(\varepsilon+\overline{\lambda}\delta)\log\frac{1}{\varepsilon+\overline{\lambda}\delta}\right)$$

where $\overline{\lambda} \coloneqq \frac{2\sup_{\Xi}|f|}{\rho - \mathbb{E}_{\pi_0} c}$ *is an explicit dual bound.*

This result is the analogue for (WDRO) of quantitative bounds for (3.1) established in [17] in the context of OT. The core of the proof consists in introducing a block approximation of the optimal transport plan, following [11]. In our situation, the second marginal of the transport plan π is not fixed, and it is the variable of the optimization problem. Moreover we have an additional variable λ to take into account. So we have to be careful to modify the approximation scheme of [11, 17] and we introduce an auxiliary regularized problem, that can be better handled that the entropy-regularized problem. Thus the proof of Theorem 3.4 requires several original steps, as described in the next section.

3.3. Proof of the approximation theorem

This section is devoted to the proof of Theorem 3.4. In fact, we state and prove a slightly more detailed result, formalized in the next theorem, and we show afterwards how Theorem 3.4 can be derived as a consequence. The following theorem can indeed be seen as a global version of Theorem 3.4, with explicit constants and slightly more general assumptions. To simplify the reading, we denote by the optimal solution of the *entropy*-regularized problem.

$$F_\rho^{\varepsilon,\delta}(f) = \sup_{\pi \in \mathcal{P}_{\mathrm{P}}(\Xi\times\Xi):\ \mathbb{E}_\pi c + \delta \,\mathrm{KL}(\pi\,|\pi_0) \le \rho} \mathbb{E}_{\pi_2} f - \varepsilon\,\mathrm{KL}(\pi\,|\pi_0).$$

Theorem 3.5 (Extended approximation theorem)**.** *Take a radius $\rho > 0$, regularization parameters $\varepsilon, \delta > 0$, and suppose that the following conditions hold:*

(i) *The objective $f\colon \Xi \to \mathbb{R}$ and the cost $c\colon \Xi \times \Xi \to \mathbb{R}_+$ are Lipschitz continuous, and that their respective Lipschitz constants satisfy $\varepsilon \le \mathrm{Lip}(f)$ and $\delta \le \mathrm{Lip}(c)$;*

(ii) *The cost c and the coupling π_0 are taken as (3.10) and (3.11) with $\sigma > 0$ such that $\mathbb{E}_{\pi_0} c < \rho$;*

(iii) *The set $\Xi \subset \mathbb{R}^d$ is compact, convex, and satisfies (for $\overline{\mathbb{B}}(\xi, \Delta)$ the ball for $\|\cdot\|$)*

$$V \coloneqq \inf_{\xi\in\Xi, 0<\Delta\le d} \frac{\mathrm{vol}\big(\Xi \cap \overline{\mathbb{B}}(\xi,\Delta)\big)}{\Delta^d} > 0. \tag{3.12}$$

Then, we have,

$$F^{0,0}_{\frac{\rho}{1+\delta/\sigma}}(f) - (\varepsilon + \overline{\lambda}\delta)\left(d + d\log\left(\frac{L}{(\varepsilon+\overline{\lambda}\delta)d}\right) + C + \frac{1}{\sigma}\left(\frac{(\varepsilon+\overline{\lambda}\delta)d}{L}\right)^p\right) - \frac{\varepsilon\rho}{\sigma+\delta}$$
$$\le F_\rho^{\varepsilon,\delta}(f) \le F_\rho^{0,0}(f).$$

where $\overline{\lambda} = \frac{2\sup_{\Xi}|f|}{\rho - \mathbb{E}_{\pi_0} c}$, $L = \mathrm{Lip}(f) + \overline{\lambda}\,\mathrm{Lip}(c)$, *and* $C = \min\left\{\log\frac{\mathrm{vol}(\Xi)}{V}, \log\frac{I_\sigma}{V}\right\}$ *with* $I_\sigma = \sigma^{\frac{d}{p}} \int_{\mathbb{R}^d} e^{-\frac{\|\zeta\|^p}{2p-1}}\,\mathrm{d}\zeta$.

The proof of this result requires a few preliminary steps. First, we provide in Lemma 3.6 a simple approximation result for the *cost*-regularized problem

$$G_\rho^{\varepsilon,\delta}(f) = \sup_{\pi \in \mathcal{P}_{\mathrm{P}}(\Xi\times\Xi):\ \mathbb{E}_\pi c + \delta\mathbb{E}_\pi c \le \rho} \mathbb{E}_{\pi_2} f - \varepsilon\mathbb{E}_\pi c. \tag{3.13}$$

Next, we bound in Lemma 3.7 the dual optimal solution of the *entropy*-regularized problem $F_\rho^{\varepsilon,\delta}(f)$. Finally, for a fixed dual variable, we compare in Lemma 3.8 the values of the Lagrangians of the *entropy*-regularized problem and the *cost*-regularized one, which is the most technical part of the proof. After these three lemmas, we prove the approximation result in the extended version (Thm. 3.5), and show how the initial version (Thm. 3.4) can be derived from it.

Lemma 3.6 (Approximation for cost-regularization)**.** *Let Assumption 2.1 hold and take $\varepsilon, \delta > 0$. Then, the following bound hold*

$$G^{0,0}_{\frac{\rho}{1+\delta}}(f) - \frac{\varepsilon\rho}{1+\delta} \;\leq\; G^{\varepsilon,\delta}_{\rho}(f) \;\leq\; G^{0,0}_{\rho}(f).$$

Proof. Since the cost function is non-negative, we directly have $G^{\varepsilon,\delta}_{\rho}(f) \leq G^{0,0}_{\rho}(f)$. From Corollary 2.3, we write the cost-regularized function $G^{\varepsilon,\delta}_{\rho}(f)$ of (3.13) as follows

$$\begin{aligned}
G^{\varepsilon,\delta}_{\rho}(f) &= \inf_{\lambda\geq 0} \lambda\rho + \mathbb{E}_{\xi\sim \mathrm{P}}\left[\sup_{\zeta\in\Xi} f(\zeta) - (\varepsilon + (1+\delta)\lambda)c(\xi,\zeta)\right] \\
&= \inf_{\lambda'\geq \varepsilon} \frac{\lambda'-\varepsilon}{1+\delta}\rho + \mathbb{E}_{\xi\sim \mathrm{P}}\left[\sup_{\zeta\in\Xi} f(\zeta) - \lambda' c(\xi,\zeta)\right] \qquad (\text{with } \lambda' = \varepsilon + (1+\delta)\lambda) \\
&\geq \inf_{\lambda'\geq 0} \frac{\lambda'-\varepsilon}{1+\delta}\rho + \mathbb{E}_{\xi\sim \mathrm{P}}\left[\sup_{\zeta\in\Xi} f(\zeta) - \lambda' c(\xi,\zeta)\right] \\
&\geq \frac{-\varepsilon\rho}{1+\delta} + \inf_{\lambda'\geq 0} \lambda' \frac{\rho}{1+\delta} + \mathbb{E}_{\xi\sim \mathrm{P}}\left[\sup_{\zeta\in\Xi} f(\zeta) - \lambda' c(\xi,\zeta)\right] \\
&= \sup_{\pi\in\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi):\mathbb{E}_\pi c\leq \frac{\rho}{1+\delta}} \mathbb{E}_{\pi_2} f - \frac{\varepsilon\rho}{1+\delta} \;=\; G^{0,0}_{\frac{\rho}{1+\delta}}(f) - \frac{\varepsilon\rho}{1+\delta},
\end{aligned}$$

where the last line follows again from Corollary 2.3 with $\varepsilon = \delta = 0$. □

Lemma 3.7 (Upper-bound on dual solutions)**.** *Under the assumptions of Theorem 3.1, the optimal solution $\lambda^\star$ of the dual problem of $F^{\varepsilon,\delta}_{\rho}(f)$ is bounded as follows*

$$\lambda^\star \leq \overline{\lambda} = \frac{2\sup_\Xi |f|}{\rho - \mathbb{E}_{\pi_0} c}. \tag{3.14}$$

Proof. Theorem 3.1 gives the existence of $\lambda^\star$, which, by definition, minimizes

$$g\colon \lambda \mapsto \lambda\rho + (\varepsilon + \lambda\delta)\mathbb{E}_{\xi\sim \mathrm{P}} \log\left(\mathbb{E}_{\zeta\sim\pi_0(\cdot|\xi)} e^{\frac{f(\zeta) - \lambda c(\xi,\zeta)}{\varepsilon + \lambda\delta}}\right).$$

On the one hand, $g(\lambda^\star)$ is upper bounded as

$$g(\lambda^\star) \leq g(0) = \varepsilon\, \mathbb{E}_{\xi\sim \mathrm{P}} \log\left(\mathbb{E}_{\zeta\sim\pi_0(\cdot|\xi)} e^{\frac{f(\zeta)}{\varepsilon}}\right) \leq \sup_\Xi |f|.$$

On the other hand, thanks to Jensen's inequality, $g(\lambda^\star)$ is lower-bounded as

$$g(\lambda^\star) \geq \lambda^\star\rho + \mathbb{E}_{(\xi,\zeta)\sim\pi_0}[f(\zeta) - \lambda^\star c(\xi,\zeta)] \geq \lambda^\star(\rho - \mathbb{E}_{\pi_0} c) - \sup_\Xi |f|.$$

Combining the two inequalities gives (3.14). □

Lemma 3.8 (Approximation bound for the Lagrangians)**.** *Under the assumptions of Theorem 3.5, consider*

$$F^{\varepsilon,\delta}_{\rho}(\lambda, f) = \sup_{\pi\in\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)} \mathbb{E}_{(\xi,\zeta)\sim\pi}[f(\zeta) - \lambda c(\xi,\zeta)] - (\varepsilon + \lambda\delta)\,\mathrm{KL}(\pi\,|\pi_0),$$

$$\textit{and} \quad G_\rho^{\frac{\varepsilon}{\sigma},\frac{\delta}{\sigma}}(\lambda, f) = \sup_{\pi \in \mathcal{P}_{\mathrm{P}}(\Xi \times \Xi)} \mathbb{E}_{(\xi,\zeta)\sim\pi}[f(\zeta) - \lambda c(\xi,\zeta)] - \Big(\frac{\varepsilon + \lambda\delta}{\sigma}\Big) c(\xi,\zeta).$$

Then we have, for a fixed $\Delta \in (0, d]$ *and with* $I_\sigma(\xi) \coloneqq \int_\Xi e^{-\frac{c(\xi,\zeta)}{2^{p-1}\sigma}} \mathrm{d}\zeta$,

$$G_\rho^{\frac{\varepsilon}{\sigma},\frac{\delta}{\sigma}}(\lambda, f) \leq F_\rho^{\varepsilon,\delta}(\lambda, f) + (\mathrm{Lip}(f) + \lambda\,\mathrm{Lip}(c))\Delta + (\varepsilon + \lambda\delta)\bigg(\frac{\Delta^p}{\sigma} - \log\big(V\Delta^d\big) + \mathbb{E}_{\xi\sim\mathrm{P}} \log I_\sigma(\xi)\bigg).$$

Proof. We start by reformulating $G_\rho^{\frac{\varepsilon}{\sigma},\frac{\delta}{\sigma}}(\lambda, f)$. By continuity and compactness, the function $\zeta \mapsto f(\zeta) - (\lambda + (\varepsilon + \lambda\delta)/\sigma)c(\xi,\zeta)$ has a maximizer on Ξ for every ξ. By [31], Theorem 14.37, we get that there exists a measurable map $\zeta^\star \colon \Xi \to \Xi$ such that $\zeta^\star(\xi) \in \arg\max_{\zeta\in\Xi} f(\zeta) - (\lambda + (\varepsilon + \lambda\delta)/\sigma)c(\xi,\zeta)$ for any $\xi \in \Xi$. Then,

$$\pi^\star(\mathrm{d}\xi, \mathrm{d}\zeta) \coloneqq \mathrm{P}(\mathrm{d}\xi)\, \delta_{\zeta^\star(\xi)}(\mathrm{d}\zeta)$$

is an optimal solution, and therefore

$$G_\rho^{\frac{\varepsilon}{\sigma},\frac{\delta}{\sigma}}(\lambda, f) = \mathbb{E}_{\pi^\star} F_\lambda - \frac{\varepsilon + \lambda\delta}{\sigma} \mathbb{E}_{\pi^\star} c. \tag{3.15}$$

Now define $\overline{\mathbb{B}}^\Delta(\xi) \coloneqq \overline{\mathbb{B}}(\zeta^\star(\xi), \Delta)$ and $\pi^\Delta \in \mathcal{P}_{\mathrm{P}}(\Xi \times \Xi)$ such that

$$\pi^\Delta \propto \mathbb{1}_{\zeta\in\overline{\mathbb{B}}^\Delta(\xi)}\, \pi_0(\mathrm{d}\xi, \mathrm{d}\zeta).$$

Note first that we have

$$\mathbb{E}_{\pi^\star} F_\lambda - \mathbb{E}_{\pi^\Delta} F_\lambda = \mathbb{E}_{\xi\sim\mathrm{P}} \mathbb{E}_{\zeta\sim\pi^\Delta(\cdot|\xi)}[F_\lambda(\xi, \zeta^\star(\xi)) - F_\lambda(\xi,\zeta)] \;\leq\; (\mathrm{Lip}(f) + \lambda\,\mathrm{Lip}(c))\Delta, \tag{3.16}$$

since F_λ is $\big(\mathrm{Lip}(f) + \lambda\,\mathrm{Lip}(c)\big)$-Lipschitz continuous and the support of $\pi^\Delta(\cdot|\xi)$ is $\overline{\mathbb{B}}^\Delta(\xi)$. Now, we proceed to bound $\mathrm{KL}(\pi^\Delta\,|\pi_0)$ by first noticing that

$$\begin{aligned}
\mathrm{KL}(\pi^\Delta\,|\pi_0) &= \mathbb{E}_{\xi\sim\mathrm{P}} \mathbb{E}_{\zeta\sim\pi^\Delta(\cdot|\xi)} \log\bigg(\frac{\mathrm{d}\,\pi^\Delta(\zeta|\xi)}{\mathrm{d}\pi_0(\zeta|\xi)}\bigg) \\
&= -\mathbb{E}_{\xi\sim\mathrm{P}} \log\bigg(\int_{\Xi\cap\overline{\mathbb{B}}^\Delta(\xi)} e^{-\frac{c(\xi,\zeta)}{2^{p-1}\sigma}} \mathrm{d}\zeta\bigg) + \mathbb{E}_{\xi\sim\mathrm{P}} \log\bigg(\int_\Xi e^{-\frac{c(\xi,\zeta)}{2^{p-1}\sigma}} \mathrm{d}\zeta\bigg)
\end{aligned} \tag{3.17}$$

We focus on lower-bounding $\int_{\Xi\cap\overline{\mathbb{B}}^\Delta(\xi)} e^{-\frac{c(\xi,\zeta)}{2^{p-1}\sigma}} \mathrm{d}\zeta$. First, note that by the triangular inequality (and since $p \geq 1$), for any $\xi, \zeta \in \Xi$,

$$\frac{c(\xi,\zeta)}{2^p} \leq \frac{c(\xi,\zeta^\star(\xi)) + c(\zeta^\star(\xi),\zeta)}{2}.$$

If, in addition, ζ is in $\overline{\mathbb{B}}^\Delta(\xi)$, this bound becomes $\frac{c(\xi,\zeta)}{2^{p-1}} \leq c(\xi,\zeta^\star(\xi)) + \Delta^p$. Hence, we have

$$\int_{\Xi\cap\overline{\mathbb{B}}^\Delta(\xi)} e^{-\frac{c(\xi,\zeta)}{2^{p-1}\sigma}} \mathrm{d}\zeta \geq e^{-\frac{c(\xi,\zeta^\star(\xi))+\Delta^p}{\sigma}} \mathrm{vol}(\Xi \cap \overline{\mathbb{B}}^\Delta(\xi)) \geq e^{-\frac{c(\xi,\zeta^\star(\xi))+\Delta^p}{\sigma}} V\Delta^d,$$

where V is defined in (3.12). Plugging the above lower and upper bounds into (3.17) yields

$$\begin{aligned}\mathrm{KL}(\pi^{\Delta}\,|\pi_0) &\leq \frac{\mathbb{E}_{\xi\sim \mathrm{P}} c(\xi,\zeta^\star(\xi)) + \Delta^p}{\sigma} - \log\big(V\Delta^d\big) + \mathbb{E}_{\xi\sim\mathrm{P}} \log I_\sigma(\xi)\\ &= \frac{\mathbb{E}_{\pi^\star} c + \Delta^p}{\sigma} - \log\big(V\Delta^d\big) + \mathbb{E}_{\xi\sim\mathrm{P}} \log I_\sigma(\xi). \end{aligned} \tag{3.18}$$

Finally, putting (3.15), (3.16), and (3.18) together gives,

$$\begin{aligned} &G_\rho^{\frac{\varepsilon}{\sigma},\frac{\delta}{\sigma}}(\lambda, f) = \mathbb{E}_{\pi^\star} F_\lambda - \frac{\varepsilon+\lambda\delta}{\sigma}\mathbb{E}_{\pi^\star} c\\ &= \mathbb{E}_{\pi^\Delta} F_\lambda - (\varepsilon+\lambda\delta)\,\mathrm{KL}(\pi^\Delta\,|\pi_0) + (\mathbb{E}_{\pi^\star} F_\lambda - \mathbb{E}_{\pi^\Delta} F_\lambda) + \left((\varepsilon+\lambda\delta)\,\mathrm{KL}(\pi^\Delta\,|\pi_0) - \frac{\varepsilon+\lambda\delta}{\sigma}\mathbb{E}_{\pi^\star} c\right)\\ &\leq F_\rho^{\varepsilon,\delta}(\lambda, f) + (\mathrm{Lip}(f) + \lambda\,\mathrm{Lip}(c))\Delta + (\varepsilon+\lambda\delta)\left(\frac{\Delta^p}{\sigma} - \log\big(V\Delta^d\big) + \mathbb{E}_{\xi\sim\mathrm{P}}\log I_\sigma(\xi)\right)\end{aligned}$$

which is the claimed inequality. □

We have now all the ingredients to establish the extended version of the approximation result.

Proof of Theorem 3.5. First, notice that by Lemma 3.6, we have that

$$F^{0,0}_{\frac{\rho}{1+\delta/\sigma}}(f) - \frac{\varepsilon\rho}{\sigma+\delta} = G^{0,0}_{\frac{\rho}{1+\delta/\sigma}}(f) - \frac{\varepsilon\rho}{\sigma+\delta} \leq G_\rho^{\frac{\varepsilon}{\sigma},\frac{\delta}{\sigma}}(f).$$

Thus, using the bound at λ fixed, given by Lemma 3.8 and the upper-bound (3.14), we get

$$\begin{aligned} G_\rho^{\frac{\varepsilon}{\sigma},\frac{\delta}{\sigma}}(f) \;&\leq\; \inf_{0\leq\lambda\leq\overline{\lambda}} G_\rho^{\frac{\varepsilon}{\sigma},\frac{\delta}{\sigma}}(\lambda, f)\\ &\leq \inf_{0\leq\lambda\leq\overline{\lambda}} F_\rho^{\varepsilon,\delta}(\lambda, f) + (\mathrm{Lip}(f) + \lambda\,\mathrm{Lip}(c))\Delta + (\varepsilon+\lambda\delta)\left(\frac{\Delta^p}{\sigma} - \log\big(V\Delta^d\big) + \mathbb{E}_{\xi\sim\mathrm{P}}\log I_\sigma(\xi)\right)\\ &\leq F_\rho^{\varepsilon,\delta}(f) + \big(\mathrm{Lip}(f) + \overline{\lambda}\,\mathrm{Lip}(c)\big)\,\Delta + (\varepsilon+\overline{\lambda}\delta)\left(\frac{\Delta^p}{\sigma} - \log\big(V\Delta^d\big) + \mathbb{E}_{\xi\sim\mathrm{P}}\log I_\sigma(\xi)\right).\end{aligned}$$

Minimizing the above bound over $\Delta > 0$, we get that the optimal value $\Delta^\star$ is of the form $\Delta^\star = \frac{(\varepsilon+\overline{\lambda}\delta)d}{L} + \mathcal{O}((\varepsilon+\overline{\lambda}\delta)^{p+1})$, with $L = \mathrm{Lip}(f) + \overline{\lambda}\,\mathrm{Lip}(c)$, as introduced in the statement of the theorem. As a consequence, we set $\Delta = \frac{(\varepsilon+\overline{\lambda}\delta)d}{L}$ in the bound above, which becomes

$$G_\rho^{\frac{\varepsilon}{\sigma},\frac{\delta}{\sigma}}(f) \leq F_\rho^{\varepsilon,\delta}(f) + (\varepsilon+\overline{\lambda}\delta)\left(d + d\log\left(\frac{L}{(\varepsilon+\overline{\lambda}\delta)d}\right) + \mathbb{E}_{\xi\sim\mathrm{P}}\log I_\sigma(\xi) + \log\frac{1}{V} + \frac{1}{\sigma}\left(\frac{(\varepsilon+\overline{\lambda}\delta)d}{L}\right)^p\right).$$

There is only left to bound the term in $\mathbb{E}_{\xi\sim\mathrm{P}}\log I_\sigma(\xi)$. On one hand, we have $e^{-\frac{c(\xi,\zeta)}{2^{p-1}\sigma}} \leq 1$ so that

$$\int_\Xi e^{-\frac{c(\xi,\zeta)}{2^{p-1}\sigma}}\mathrm{d}\zeta \leq \mathrm{vol}(\Xi).$$

On the other hand, we also have (using the change of variable $\zeta' = \sigma^{-\frac{1}{p}}(\zeta - \xi)$ to get I_σ)

$$\int_\Xi e^{-\frac{c(\xi,\zeta)}{2^{p-1}\sigma}} \mathrm{d}\zeta \leq \int_{\mathbb{R}^d} e^{-\frac{\|\xi-\zeta\|^p}{2^{p-1}\sigma}} \mathrm{d}\zeta = I_\sigma$$

This makes the constant C appear in the bound and thus ends the proof. □

We finish by explaining how the main theorem, Theorem 3.4, stems from Theorem 3.5. On the left-hand side (LHS) of the inequality in Theorem 3.5, the unregularized objective has radius $\frac{\rho}{1+\delta/\sigma}$, instead of simply ρ in $F_\rho^{0,0}(f) = \mathrm{val}\,(\mathrm{WDRO})$. Thus, we compare in the next lemma the optimal values for these two parameters.

Lemma 3.9 (Comparing optimal values)**.** *Under the assumptions of Theorem 3.5,*

$$F_\rho^{0,0}(f) \leq F^{0,0}_{\frac{\rho}{1+\delta/\sigma}}(f) + \mathcal{O}(\delta).$$

Proof. We first apply [34], Theorem 5.27 to get a constant-speed geodesic for the p-Wasserstein distance connecting P and Q, which is $W_c(\mathrm{P},\mathrm{Q})^{\frac{1}{p}}$ with our notation. This means that there exists a family of probability distributions $(\mathrm{Q}_t)_{t\in[0,1]}$ such that $\mathrm{Q}_0 = \mathrm{Q}$, $\mathrm{Q}_1 = \mathrm{P}$ and, for any $t \in [0,1]$,

$$W_c(\mathrm{P},\mathrm{Q}_t)^{\frac{1}{p}} = (1-t)W_c(\mathrm{P},\mathrm{Q})^{\frac{1}{p}} \qquad \text{and} \qquad W_c(\mathrm{Q}_t,\mathrm{Q})^{\frac{1}{p}} = tW_c(\mathrm{P},\mathrm{Q})^{\frac{1}{p}}.$$

We apply these equations with Q such that $W_c(\mathrm{P},\mathrm{Q}) \leq \rho$ and $t = 1 - (1+\delta/\sigma)^{-\frac{1}{p}}$ to obtain

$$W_c(\mathrm{P},\mathrm{Q}_t) \leq \frac{\rho}{1+\delta/\sigma} \qquad \text{and} \qquad W_c(\mathrm{Q},\mathrm{Q}_t) \leq t^p\rho = \mathcal{O}(\delta).$$

Note that the first inequality above yields

$$\mathbb{E}_{\mathrm{Q}_t} f \leq \sup_{\pi\in\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi):\mathbb{E}_\pi c\leq\frac{\rho}{1+\delta/\sigma}} \mathbb{E}_{\pi_2} f = F^{0,0}_{\frac{\rho}{1+\delta/\sigma}}(f), \tag{3.19}$$

We now use the Kantorovich-Rubinstein inequality (*e.g.*, [37], Theorem 1.14) to write

$$\mathbb{E}_{\mathrm{Q}} f - \mathbb{E}_{\mathrm{Q}_t} f \;\leq\; W_1(\mathrm{Q},\mathrm{Q}_t)\,\mathrm{Lip}(f) \;\leq\; W_c(\mathrm{Q},\mathrm{Q}_t)^{\frac{1}{p}}\,\mathrm{Lip}(f),$$

where for the second inequality we used that the p-Wasserstein distance is always greater than or equal to the 1-Wasserstein distance (*e.g.*, [37], Section 7.1.2). Together with (3.19), this yields:

$$\mathbb{E}_{\mathrm{Q}} f \leq \mathbb{E}_{\mathrm{Q}_t} f + \mathcal{O}(\delta) \leq F^{0,0}_{\frac{\rho}{1+\delta/\sigma}}(f) + \mathcal{O}(\delta).$$

Taking the supremum over all Q such that $W_c(\mathrm{P},\mathrm{Q}) \leq \rho$ allows us to conclude. □

With the help of the previous lemma, the proof of Theorem 3.4 comes easily from Theorem 3.5.

Proof of Theorem 3.4. We start with checking that the fact that Ξ is a compact convex body (condition (iv) in Thm. 3.4) implies that $V > 0$ (condition (iv) in Thm. 3.5). Introduce, for $\xi \in \Xi$, the function

$$\nu_\xi : \Delta \mapsto \frac{\mathrm{vol}\big(\Xi \cap \overline{\mathbb{B}}(\xi,\Delta)\big)}{\Delta^d} = \mathrm{vol}\bigg(\frac{1}{\Delta}(\Xi-\xi)\cap\overline{\mathbb{B}}(0,1)\bigg).$$

Since Ξ is convex, we easily get that ν_ξ is non-increasing. Thus we can lower-bound the constant V as follows, using $\mathrm{diam}(\Xi) := \sup_{\xi,\zeta\in\Xi}\|\xi-\zeta\|$ the diameter of Ξ.

$$\begin{aligned} V &= \inf_{\xi\in\Xi, 0<\Delta\le d} \nu_\xi(\Delta) \;\ge\; \inf_{\xi\in\Xi, 0<\Delta\le \max(d,\mathrm{diam}(\Xi))} \nu_\xi(\Delta) \\ &= \inf_{\xi\in\Xi} \frac{\mathrm{vol}\big(\Xi\cap\overline{\mathbb{B}}(\xi,\max(d,\mathrm{diam}(\Xi)))\big)}{(\max(d,\mathrm{diam}(\Xi))^d} \;\ge\; \frac{\mathrm{vol}(\Xi)}{(\max(d,\mathrm{diam}(\Xi))^d} > 0. \end{aligned}$$

So we get $V>0$ which is condition (iv) in Theorem 3.5. Thus we can apply Theorem 3.5 and Lemma 3.9. Noting that $F_\rho^{\varepsilon,\delta}(f) = \mathrm{val}\,(\text{E-WDRO})$ and $F_\rho^{0,0}(f) = \mathrm{val}\,(\text{WDRO})$, and combining the obtained bound with Lemma 3.9 gives the result. □

4. Conclusion, perspectives

Inspired by the success of regularization in OT, we proposed and studied a regularization scheme for WDRO problems. We derived the expression of the dual objective function in the general case as well as a refined one in the particular setting of the entropic regularization. In addition, we showed that the difference between the original WDRO problem and the entropic one is properly controlled by the regularization parameters.

Since regularization in OT has shown attractive computational advantages and statistical benefits, an exciting research direction is to investigate whether similar gains hold for regularization in WDRO. It is also worth studying possible extensions of the results of this paper, including the approximation errors with general regularizations, beyond the entropic case.

Appendix A. Duality under weaker assumptions

In Theorem 2.2, we derive the expression of the dual of the regularized WDRO problem, with arbitrary regularizations. In this situation, the primal is not completely explicit, and then compactness and continuity are crucial, as we leverage the duality between continuous functions and measures on a compact space.

In this appendix, we provide a generalization of the duality result that does away with the compactness and the continuity assumptions. This is achieved by carefully designing the spaces in which duality arguments are made. The general duality result was informally given in Section 2.3. Here, we provide the mathematical developments, as follows: Appendix A.1 presents the setting; Appendix A.2 provides the theorem; Appendix A.3 states some basic results used in the the proof of Appendix A.4; Appendix A.5 finishes with two examples.

A.1 Setting: assumptions, recalls, and construction of the spaces

In the appendix, we consider the following set of blanket assumptions.

Assumption A.1.

(i) $\Xi\subset\mathbb{R}^d$ is a closed set.

(ii) $f\colon\Xi\to\mathbb{R}$ is measurable w.r.t. the Borel σ-algebra on Ξ;

(iii) $c:\Xi\times\Xi\to\mathbb{R}_+$ is measurable as well and, for all ξ in Ξ, $c(\xi,\xi)=0$;

(iv) There is some $\xi_0\in\Xi$ such that the integrability condition $\mathbb{E}_{\xi\sim\mathrm{P}}[c(\xi,\xi_0)]<+\infty$ holds.

For convenience, we denote by c_0 the function $(\xi,\zeta)\mapsto c(\xi,\xi_0)$. We work in this appendix with the probability space $(\Xi^2,\mathcal{F}^{\otimes^2},\pi_0)$ where π_0 is a measure satisfying the following assumption.

Assumption A.2. $\pi_0\in\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)$ is a reference measure such that $1+c_0+c$ is integrable w.r.t. π_0.

As a consequence of (ii), we consider $L^0(\pi_0)$ the set of measurable functions on Ξ^2 (where functions which agree almost everywhere w.r.t. π_0 are identified), equipped with the topology which metrizes the convergence in probability, see *e.g.*, [20], Chapter 2, Section 2 or [7], Section 4.7.60. For simplicity, we identify $L^0(\pi_0)$, and

more generally the spaces $L^p(\pi_0)$, as subsets of $\mathcal{M}(\Xi \times \Xi)$ through the map $g \mapsto g\,\mathrm{d}\pi_0$. In turn, $\mathcal{M}(\Xi \times \Xi)$ is itself seen as a subspace of $(L^b(\Xi^2))^*$, the topological dual of the space of bounded (everywhere) measurable functions.

A.1.1 Recalls in modular spaces

The upcoming results are built over modular spaces [26], *i.e.*, subsets of $L^1(\pi_0)$ defined for some $H\colon L^0(\pi_0) \to \mathbb{R} \cup \{+\infty\}$ as

$$L^H \coloneqq \left\{g \in L^1(\pi_0) : \exists \beta > 0, H(g/\beta) < +\infty\right\}.$$

Under suitable assumptions, the following lemma states that L^H equipped with the norm $\|g\|_H \coloneqq \inf\{\beta > 0 : H(g/\beta) \leq 1\}$ is a Banach space.

Lemma A.3. *Consider $H : L^0(\pi_0) \to \mathbb{R}_+ \cup \{+\infty\}$ convex, symmetric, with $H(0) = 0$. If H is $L^0(\pi_0)$-l.s.c. and $\{g \in L^0(\pi_0) : H(g) \leq 1\}$ is bounded in $L^0(\pi_0)$, then $(L^H, \|\cdot\|_H)$ is a Banach space.*

Proof. The proof follows the second paragraph of the proof of [28], Theorem 22, which consists in applying [28], Remark 9. For this, one needs to show that the topology on L^H is no weaker than the $L^0(\pi_0)$ topology. First note that the $L^0(\pi_0)$-l.s.c. of H implies the $L^0(\pi_0)$-l.s.c. of $\|\cdot\|_{L^H}$. Indeed, take a sequence g_t of L^H which converges to g: for any $\alpha > \|g_t\|_{L^H}$, it holds that $H(g_t/\alpha) \leq 1$.

Now, take $\mathcal{U} \subset L^H$ an $L^0(\pi_0)$-neighborhood of the origin. By the boundedness assumption, there is $\alpha > 0$ such that $\{g \in L^H : \|g\|_H \leq 1\} = \{g \in L^0(\pi_0) : H(g) \leq 1\} \subset \alpha\mathcal{U}$ so that $\mathcal{U} \cap L^H$ is a neighborhood of 0 in L^H. Applying [28], Remark 9 yields the result. □

We also consider a special subspace of L^H, named the Orlicz heart, denoted by $L^H_\heartsuit$ and defined by

$$L^H_\heartsuit \coloneqq \left\{g \in L^H : \forall \beta > 0, H(g/\beta) < +\infty\right\}.$$

Our motivation for introducing such spaces was guided by [1], which explores the link between ϕ-divergences and associated Orlicz spaces. Note though that our construction is more general and slightly different than theirs.

A.1.2 Construction of the function space for duality

We first introduce two modular spaces, L^{H_c} and L^{H_R}, used later to build our function space $\mathcal{X}$.

Lemma A.4. *Define $H_c : L^0(\pi_0) \to \mathbb{R} \cup \{+\infty\}$ as*

$$H_c \coloneqq \iota_{\left\{g \in L^0(\pi_0) : \left\|\frac{g}{1+c_0+c}\right\|_{L^\infty(\pi_0)} \leq 1\right\}}.$$

Then $(L^{H_c}, \|\cdot\|_{H_c})$ is a Banach space.

Proof. We check that H_c satisfies the assumptions of Lemma A.3, and in particular that it is $L^0(\pi_0)$-l.s.c. and that $\left\{g \in L^0(\pi_0) : H_c(g) \leq 1\right\}$ is bounded in probability, which comes down to showing that $\left\{g \in L^0(\pi_0) : \|g/(1 + c_0 + c)\|_{L^\infty(\pi_0)} \leq 1\right\}$ is closed in $L^0(\pi_0)$ and bounded in $L^0(\pi_0)$.

To show that it is closed, consider a sequence of functions g_t for $t = 1, 2, \ldots$ belonging to this set which converges in probability to some $g \in L^0(\pi_0)$. Then, for any $\eta > 0$,

$$\pi_0\left(\frac{|g|}{1+c_0+c} > 1 + \eta\right) \leq \pi_0\left(\frac{|g - g_t|}{1+c_0+c} > \eta\right)\pi_0\left(\frac{|g_t|}{1+c_0+c} > 1\right)$$

$$\leq \pi_0(|g - g_t| > \eta) \xrightarrow{t\to+\infty} 0.$$

Hence, by continuity of probability,

$$\pi_0\left(\frac{|g|}{1+c_0+c} > 1\right) = \lim_{\eta\to 0} \pi_0\left(\frac{|g|}{1+c_0+c} > 1+\eta\right) = 0,$$

which shows that $\|g/(1+c_0+c)\|_{L^\infty(\Xi)} \leq 1$.

We now show that $\left\{g \in L^0(\pi_0) : \|g/(1+c_0+c)\|_{L^\infty(\Xi)} \leq 1\right\}$ is bounded in $L^0(\pi_0)$ *i.e.*,

$$\forall \eta > 0,\, \exists M > 0,\, \forall g \in L^0(\pi_0) \quad \text{s.t.} \quad H(g) \leq 1,\, \pi_0(|g| \geq M) \leq \eta.$$

To do so, let us take $g \in L^0(\pi_0)$ such that $\|g/(1+c_0+c)\|_{L^\infty(\pi_0)} \leq 1$. Then, for any $M > 0$, by Markov's inequality

$$\pi_0(|g| \geq M) \leq \frac{\mathbb{E}_{\pi_0}[|g|]}{M} \leq \frac{\mathbb{E}_{\pi_0}[1+c_0+c]}{M},$$

and by assumption, $\mathbb{E}_{\pi_0}[1+c_0+c]$ is finite, so that the RHS is arbitrarily small as M grows. □

Now, let us define a modular space, starting from a "regularization" R (a more formal link will be made in Appendix A.5).

Lemma A.5. *Consider $R : (L^b(\Xi^2))^* \to \mathbb{R}_+$ convex such that*

1. $\operatorname{dom} R$ *is included in the cone of non-negative linear forms;*
2. $\inf R = 0 = R(\pi_R)$ *for some $\pi_R \in L^\infty(\pi_0)$ such that both $(1+\alpha)\,\pi_R$ and $\pi_R + \alpha$ are in $\operatorname{dom} R$ for some $\alpha > 0$;*

and define $H_R : L^0(\pi_0) \to \mathbb{R} \cup \{+\infty\}$ by

$$H_R(g) \coloneqq \sup_{\pi \in L^0(\pi_0)} \int_{\Xi^2} |g|\,\pi\,\mathrm{d}\pi_0 - R(\pi + \pi_R).$$

Then $(L^{H_R}, \|\cdot\|_{H_R})$ is a Banach space. As a consequence, $(L^{H_R}_\heartsuit, \|\cdot\|_{H_R})$ is a Banach space too.

Proof. Note first that $H_R(g) = \sup_{\pi \in L^0(\pi_0) : \pi \geq 0} \int_{\Xi^2} |g|\,\pi\,\mathrm{d}\pi_0 - \int_{\Xi^2} |g|\,\pi_R\,\mathrm{d}\pi_0 - R(\pi)$ so that it is well-defined for any $g \in L^1(\pi_0)$ since the integral $\int_{\Xi^2} |g|\,\pi\,\mathrm{d}\pi_0$ always has a value in $\mathbb{R} \cup \{+\infty\}$ and because $g\,\pi_R$ is still in $L^1(\pi_0)$. Moreover H_R is convex as a supremum of convex functions.

Let us show that H_R satisfies the conditions of Lemma A.3. First, H_R is non-negative since

$$H_R(g) \geq \langle 0, |g| \rangle - R(0 + \pi_R) = 0 \qquad \text{for } g \in L^0(\pi_0).$$

Then, H_R is $L^0(\pi_0)$-l.s.c. since the functions $g \mapsto \int |g|(\pi - \pi_R)\mathrm{d}\pi_0 - R(\pi)$ are $L^0(\pi_0)$-l.s.c. for any $\pi \in L^0(\pi_0)$ non-negative by Fatou's lemma. Finally, it remains to check the boundedness condition, *i.e.*, that $\left\{g \in L^0(\pi_0) : H_R(g) \leq 1\right\}$ is bounded in $L^0(\pi_0)$. But, since $\pi_R + \alpha \in \operatorname{dom} R$ for some $\alpha > 0$, for any $g \in L^0(\pi_0)$, $H_R(g) \geq \alpha \|g\|_{L^1(\pi_0)}$ so that $\left\{g \in L^0(\pi_0) : H_R(g) \leq 1\right\}$ is included in a $L^1(\pi_0)$-ball and is *a fortiori* bounded in probability by Markov inequality.

For $L^{H_R}_\heartsuit$ to be Banach space too, it suffices to check that it is closed in L^{H_R}, following the arguments of the proof of [28], Theorem 22. Indeed, take a sequence $(g_t)_t$ from $L^{H_R}_\heartsuit$ that converges to $g \in L^H$ for the norm $\|\cdot\|_{L^{H_R}}$. For any $\beta > 0$, $H_R((g - g_t)/\beta) \leq 1$ for t large enough so the convexity of H_R gives us that,

$$H_R(g/(2\beta)) \leq \frac{1}{2} H_R(g_t/\beta) + \frac{1}{2} < +\infty,$$

so that $g \in L^{H_R}_\heartsuit$, and this ends the proof. □

As a consequence of these two lemmas, we can build the Banach space $\mathcal{X}$ that will be used in the duality proofs, see [24], page ix for instance.

Corollary A.6. *Define $\mathcal{X} = L^{H_c} + L^{H_R}_\heartsuit$ with the norm*

$$\|g\|_{\mathcal{X}} = \inf\Big\{ \|g_c\|_{L^{H_c}} + \|g_R\|_{L^{H_R}} : g_c \in L^{H_c},\, g_R \in L^{H_R}_\heartsuit \text{ s.t. } g_c + g_R = g \Big\}.$$

Then $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ is a Banach space, whose dual can be identified with

$$\mathcal{X}^* = (L^{H_c})^* \cap (L^{H_R}_\heartsuit)^*.$$

Moreover, we will need the following criterion under which a given measure belongs to $\mathcal{X}$.

Lemma A.7. *Under the assumptions of Lemma A.5, for $\pi \in \mathcal{P}_{\mathrm{P}}(\Xi \times \Xi) \cap L^0(\pi_0)$, if both $\mathbb{E}_\pi[c]$ and $R(\pi)$ are finite, then π belongs to $\mathcal{X}^*$.*

Proof. We first show that π is in $(L^{H_c})^*$. To do so, we use that by [28], Theorem 22, $L^{H_c^*}$ is a subspace of $(L^{H_c})^*$ where H_c^* is defined for all $\pi \in (L^{H_c})^*$ as $H_c^*(\pi) = \sup_{g \in L^\infty(\pi_0)} \langle \pi, g \rangle - H_c(g)$. By definition of H_c, we have that $H_c^*(\pi) = \langle \pi, 1 + c_0 + c \rangle = 1 + \mathbb{E}_{\xi \sim \mathrm{P}}[c(\xi, \xi_0)] + \mathbb{E}_\pi[c]$, which is finite by assumption. We thus have that $\pi \in (L^{H_c})^*$.

Next, we show that π defines a continuous linear form on $(L^{H_R}_\heartsuit, \|\cdot\|_{L^{H_R}})$. To do so, let us take $g \in L^{H_R}_\heartsuit$ so that by definition we can take $\beta > 0$ so that $H_R(g/\beta) < +\infty$. By the assumption of Lemma A.5, there is some $\alpha > 0$ such that $(1 + \alpha)\, \pi_\varepsilon \in \operatorname{dom} R$ so that one can choose $\lambda \in (0, 1)$ satisfying $(1 + \alpha)(1 - \lambda) = 1$. Then, by definition of H_R and convexity of R, one has that for all $\pi \in L^0(\pi_0)$, the following inequality holds

$$\frac{\lambda}{\beta} \int |g|\, \pi \, \mathrm{d}\pi_0 \leq H_R(g/\beta) + \lambda R(\pi) + (1 - \lambda) R((1 + \alpha)\, \pi_R). \tag{A.1}$$

Since the RHS is finite from the assumptions as soon as $R(\pi)$ is finite, we have that $\int g\, \pi \, \mathrm{d}\pi_0$ is well-defined for $g \in L^{H_R}_\heartsuit$. All that is left to show is that this linear form is indeed continuous. Take $g_t \in L^{H_R}_\heartsuit$ which goes to zero as t goes to $+\infty$. This means that there exists a sequence $\beta_t > 0$ which goes to zero such that $H_R(g_t/\beta_t) \leq 1$ for all $t = 1, 2, \ldots$. Applying (A.1) with $g \leftarrow g_t$, $\beta \leftarrow \beta_t$, and π such that $R(\pi)$ is finite yields

$$\frac{\lambda}{\beta_t} \int |g_t|\, \pi \, \mathrm{d}\pi_0 \leq 1 + \lambda R(\pi) + (1 - \lambda) R((1 + \alpha)\, \pi_R) < +\infty$$

which implies that $\int |g_t|\, \pi \, \mathrm{d}\pi_0$ goes to 0 as t goes to $+\infty$. □

A.2 Statement of the theorem

We can now state the formal version of Theorem 2.6.

Theorem A.8. *Take $S \coloneqq \frac{\delta}{\varepsilon} R$. Let Assumption A.1 hold and assume that*

1. *f satisfies the growth condition :*

$$\exists g \in L^\infty(\pi_0),\, \forall \beta > 0,\, (\xi,\zeta) \mapsto \beta\left(f(\zeta) - (1 + c(\xi,\xi_0) + c(\xi,\zeta))g(\xi,\zeta)\right) \in \operatorname{dom} H_R,$$

   or equivalently, that the function $\widetilde{f}$, defined as $\widetilde{f} : (\xi,\zeta) \mapsto f(\zeta)$, lies in $\mathcal{X}$;
2. *R convex satisfies the assumptions of Lemma A.5 and $R|_{\mathcal{X}^*}$ is $\sigma(\mathcal{X}^*,\mathcal{X})$-l.s.c.;*
3. *$\pi_0 \in \mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)$ satisfies Assumption A.2;*
4. *there exists $\pi_S \in \mathcal{P}_{\mathrm{P}}(\Xi\times\Xi) \cap L^0(\pi_0)$ such that $\mathbb{E}_{\pi_S}[c] + S(\pi_S) < \rho$;*
5. *one of the two following regularity conditions hold:*
   (a) *R is $\sigma(\mathcal{X}^*, \mathcal{C}(\Xi\times\Xi) \cap L^{H_c})$-u.s.c. on $\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)$, f is l.s.c., $f_- \in L^{H_c}$, $\overline{\operatorname{int}\Xi} = \Xi$ and for all $\xi \in \Xi$ and $\pi_0(\cdot|\xi)$ has a positive density w.r.t. the Lebesgue measure on Ξ.*
   (b) *$\operatorname{dom} R \subset L^0(\pi_0)$ and, for any non-increasing non-negative sequence $g_t \in \operatorname{dom} H_R$ such that $g_t \to 0$ as $t \to \infty$, we have $H_R(g_t) \to 0$.*

Then, the problem

$$\sup\left\{\mathbb{E}_{\pi_2}[f] - R(\pi) : \pi \in \mathcal{P}_{\mathrm{P}}(\Xi\times\Xi), \mathbb{E}_\pi[c] + S(\pi) \le \rho\right\}$$

coincides with its dual

$$\inf\left\{\lambda\rho + \mathbb{E}_{\xi\sim\mathrm{P}}\left[\operatorname*{ess\,sup}_{\pi_0(\cdot|\xi)}\left\{f - \lambda c(\xi,\cdot) - g(\xi,\cdot)\right\}\right] + \left(R|_{\mathcal{X}^*} + \lambda S|_{\mathcal{X}^*}\right)_*(g) : g \in \mathcal{X},\, \lambda \ge 0\right\}.$$

We state the theorem and provide its proof in the case of when R and S are proportional. The case with different R and S would follow that the same lines, at the price of slightly more complicated assumptions and a slightly different treatment of subcases in the proof. To avoid these extra-technicalities, we stick here with the proportional case, which already presents all the intrinsic reasoning and the mathematical tools. Before going to the proof of this result in Appendix A.4, we introduce in the next section some general duality lemmas, tailored for our context.

A.3 Basic duality lemmas

Lemma A.9. *Consider $\mathcal{X}$ a Banach space and $g\colon \mathcal{X}^* \to \mathbb{R}\cup\{+\infty\}$ convex, proper and weakly-$\star$ l.s.c.. Then $(g_*)^* = g$.*

Proof. Equipped with its weak-$\star$ topology, $\mathcal{X}^*$ is a locally convex topological vector space whose topological dual is $\mathcal{X}$, see [33], Theorem 3.10, Section 3.14. As a consequence, the usual theorem of the biconjugate (*e.g.*, [41], Theorem 2.3.3) ensures that $(g_*)^* = g$. □

Lemma A.10. *Given $\mathcal{X}$ a Banach space, $g,h,s,l\colon \mathcal{X}^* \to \mathbb{R}\cup\{+\infty\}$ convex, proper and weakly-$\star$ l.s.c., consider the two problems,*

$$\min_{x\in\mathcal{X},\,\lambda\ge 0} (g+\lambda s)_*(x) + (h+\lambda l)_*(-x) \tag{P}$$

$$\max_{x^*\in\mathcal{X}^*} -g(x^*) - h(x^*)\ \ s.t.\ s(x^*) + l(x^*) \le 0. \tag{D}$$

If the following qualification condition holds

$$\{x \in \mathcal{X} : \exists \lambda \ge 0,\, x \in \operatorname{dom}(g+\lambda s)_* + \operatorname{dom}(h+\lambda l)_*\} = \mathcal{X}$$

then, $(P) = (D)$.

Proof. Let us define the perturbation function $\Phi : (\mathcal{X} \times \mathbb{R}) \times \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ by

$$\Phi((x, \lambda), x') = (g + \lambda s)_*(x) + (h + \lambda l)_*(x' - x) + \iota_{\mathbb{R}_+}(\lambda).$$

Φ is l.s.c. and convex since both $(x, \lambda), x' \mapsto (g + \lambda s)_*(x)$ and $(x, \lambda), x' \mapsto (h + \lambda l)_*(x' - x)$ are jointly convex and l.s.c. as supremums of continuous linear functions. Let us now compute the conjugate of Φ, for $((x^*, \lambda^*), y) \in (\mathcal{X}^* \times \mathbb{R}) \times \mathcal{X}^*$

$$\begin{aligned}
\Phi^*((x^*, \lambda^*), y) &= \sup_{x, x' \in \mathcal{X},\, \lambda \geq 0} \{\langle x^*, x\rangle + \langle y, x'\rangle + \lambda^* \lambda - (g + \lambda s)_*(x) - (h + \lambda l)_*(x' - x)\} \\
&= \sup_{x, x' \in \mathcal{X},\, \lambda \geq 0} \{\langle x^*, x\rangle + \langle y, x' + x\rangle + \lambda^* \lambda - (g + \lambda s)_*(x) - (h + \lambda l)_*(x')\} \\
&= \sup_{\lambda \geq 0} \left\{ \lambda^* \lambda + (g + \lambda s)(x^* + y) + \sup_{x \in \mathcal{X}} \{\langle x^*, x\rangle - (h + \lambda l)_*(x')\} \right\} \\
&= \sup_{\lambda \geq 0} \{\lambda^* \lambda + (g + \lambda s)(x^* + y) + (h + \lambda l)(y)\} \\
&= g(x^* + y) + h(y) + \iota_{\lambda^* + s(x^* + y) + l(y) \leq 0},
\end{aligned}$$

where we applied Lemma A.9 to $g + \lambda s$ and $h + \lambda l$. Now, note that (P) is equal to

$$\min_{(x, \lambda) \in (\mathcal{X}, \mathbb{R})} \Phi((x, \lambda), 0) \qquad \text{and (D) to} \qquad \max_{x^* \in \mathcal{X}^*} -\Phi^*(0, x^*).$$

Now we leverage the duality theorem [41], Theorem 2.7.1 with qualification condition (vii), see [41], page 15 for the relevant definition, to show that the values of those problems are equal. This qualification condition requires that the set

$$\operatorname{cone}\{x' \in \mathcal{X} : \exists (x, \lambda) \in \mathcal{X} \times \mathbb{R},\, ((x, \lambda), x') \in \operatorname{dom} \Phi\}$$

be a closed linear subset of $\mathcal{X}$. But this set can be rewritten as

$$\operatorname{cone}\{x' \in \mathcal{X} : \exists \lambda \geq 0,\, x' \in \operatorname{dom}(g + \lambda s)_* + \operatorname{dom}(h + \lambda l)_*\}$$

and our qualification assumption ensures that it is equal to the whole space $\mathcal{X}$. □

Lemma A.11. *Consider a vector space* $\mathcal{Y}$, $g : \mathcal{Y} \to \{-\infty\} \cup \mathbb{R}$ *concave,* $s : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ *convex and assume that there exists* $y_s \in \operatorname{dom} g$ *such that* $s(y_s) < 0$. *Then,*

$$\sup\{g(y) : y \in \mathcal{Y},\, s(y) \leq 0\} = \sup\{g(y) : y \in \mathcal{Y},\, s(y) < 0\}.$$

Proof. The inequality $(\geq)$ always holds. To prove the reverse, take $y \in \operatorname{dom} g$ such that $s(y) \leq 0$, consider $y_t \coloneqq t y_s + (1 - t) y$ for $t \in [0, 1]$. By convexity of s and definition of y_s, we have that $s(y_t) < 0$ for all $t \in (0, 1]$. Moreover, since y_s is in $\operatorname{dom} g$ and g is concave, it holds that

$$g(y) \leq \liminf_{t \to 0} \frac{g(y_t) - t g(y_s)}{1 - t} = \liminf_{t \to 0} g(y_t) \leq \sup\{g(y) : y \in \mathcal{Y},\, s(y) < 0\},$$

which shows the required inequality. □

A.4 Proof of Theorem A.8

The proof of Theorem A.8 is divided two main lemmas, Lemma A.14 and Lemma A.16, which, put together, exactly give Theorem A.8.

Lemma A.12. *Assume that $\mathcal{X} \subset L^1(\pi_0)$ is a Banach space then the support function of the set $\mathcal{P}_{\mathrm{P}}(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*$, seen as a subspace of $\mathcal{X}^*$, is given for all $g \in \mathcal{X}$ by*

$$\sigma_{\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)\cap L^0(\pi_0)\cap\mathcal{X}^*}(g) \coloneqq \sup_{\pi\in\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)\cap L^0(\pi_0)\cap\mathcal{X}^*} \langle \pi, g\rangle = \mathbb{E}_{\xi\sim\mathrm{P}}\left[\operatorname*{ess\,sup}_{\pi_0(\cdot|\xi)} g(\xi,\cdot)\right]$$

Remark A.13. We observe that $\mathbb{E}_{\xi\sim\mathrm{P}}\left[\operatorname{ess\,sup}_{\pi_0(\cdot|\xi)} g(\xi,\cdot)\right]$ is well-defined for $g \in \mathcal{X}$. First, the map $\xi \mapsto \operatorname{ess\,sup}_{\pi_0(\cdot|\xi)} g(\xi,\cdot)$ is measurable since, for any $\alpha \in \mathbb{R}$, by definition of the essential supremum,

$$\begin{aligned}\left\{\xi \in \Xi : \operatorname*{ess\,sup}_{\pi_0(\cdot|\xi)} g(\xi,\cdot) > \alpha\right\} &= \{\xi \in \Xi : \pi_0(\{\zeta \in \Xi : g(\xi,\zeta) > \alpha\}|\xi) > 0\} \\ &= \left\{\xi \in \Xi : \int_\Xi \mathbb{1}_{\{(\xi,\zeta)\in\Xi^2 : g(\xi,\zeta)>\alpha\}}(\xi,\zeta)\mathrm{d}\pi_0(\zeta|\xi) > 0\right\},\end{aligned}$$

where the resulting integral is a measurable function of ξ, see *e.g.*, [19], Lemma 3.2 (i). Then, if $g \in \mathcal{X}$, then for π_0-almost every (ξ,ζ), $\operatorname{ess\,sup}_{\pi_0(\cdot|\xi)} g(\xi,\cdot) \geq g(\xi,\zeta)$ so that by integrating w.r.t. π_0, one gets that $\mathbb{E}_{\xi\sim\mathrm{P}}\left[\operatorname{ess\,sup}_{\pi_0(\cdot|\xi)} g(\xi,\cdot)\right] \geq \mathbb{E}_{\pi_0}[g] > -\infty$ since $g \in \mathcal{X} \subset L^1(\pi_0)$.

Proof. Fix some $g \in \mathcal{X}$. First, let us show the inequality $(\leq)$. To do so, take $\pi \in \mathcal{P}_{\mathrm{P}}(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*$. By disintegration of measure, it holds that

$$\mathbb{E}_\pi[g] = \mathbb{E}_{\xi\sim\mathrm{P}}\left[\mathbb{E}_{\zeta\sim\pi(\cdot|\xi)}[g(\xi,\zeta)]\right] \leq \mathbb{E}_{\xi\sim\mathrm{P}}\left[\operatorname*{ess\,sup}_{\pi_0(\cdot|\xi)} g(\xi,\cdot)\right],$$

since $\pi(\cdot|\xi)$ is a probability distribution which is absolutely continuous w.r.t. $\pi_0(\cdot|\xi)$ for any $\xi \in \Xi$.

To prove the reverse inequality, we distinguish three cases. For convenience, denote by h the measurable function $\xi \mapsto \operatorname{ess\,sup}_{\pi_0(\cdot|\xi)} g(\xi,\cdot)$. Furthermore, note that the fact that $\mathcal{X}$ is a subset of $L^1(\pi_0)$ implies that $\mathcal{X}^*$ contains $L^\infty(\pi_0)$, so the indicator functions of measurable sets, in particular.

1. If $(\mathrm{P}-a.s.\ g(\xi,\cdot) \in L^\infty(\pi_0(\cdot|\xi)))$ is false, *i.e.*, if h is not finite $\mathrm{P}-a.s.$. This means that, for any $n \geq 1$, the sets $A_n(\xi) \coloneqq \{\zeta \in \Xi : g(\xi,\zeta) \geq n\}$ satisfy $\mathrm{P}-a.s.$, $\pi_0(A_n(\xi)|\xi) > 0$. Therefore, one can define $\pi_n \in \mathcal{P}_{\mathrm{P}}(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*$ as the measurable map $(\xi,\zeta) \mapsto \frac{\mathbb{1}_{A_n(\xi)}(\zeta)}{\pi_0(A_n(\xi)|\xi)}$ and it satisfies $\langle \pi_n, g\rangle \geq n$. This shows that

$$\sup_{\pi\in\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)\cap L^0(\pi_0)\cap\mathcal{X}^*} \langle \pi, g\rangle = +\infty = \mathbb{E}_{\xi\sim\mathrm{P}}[h(\xi)].$$

2. If $\mathrm{P}-a.s.\ g(\xi,\cdot) \in L^\infty(\pi_0(\cdot|\xi))$, *i.e.*, if h is finite $\mathrm{P}-a.s.$ *and* $h \in L^1(\mathrm{P})$. As a consequence of the definition of h, for any $n \geq 1$, the sets $A_n(\xi) \coloneqq \{\zeta \in \Xi : g(\xi,\zeta) \geq h(\xi) - 1/n\}$ satisfy $\mathrm{P}\,a.s.$, $\pi_0(A_n(\xi)|\xi) > 0$. Again, one defines $\pi_n \in \mathcal{P}_{\mathrm{P}}(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*$ as the map $(\xi,\zeta) \mapsto \frac{\mathbb{1}_{A_n(\xi)}(\zeta)}{\pi_0(A_n(\xi)|\xi)}$ so that it satisfies $\langle \pi_n, g\rangle \geq \mathbb{E}_{\mathrm{P}}[h] - 1/n$ for all $n \geq 1$ and thus

$$\sup_{\pi\in\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)\cap L^0(\pi_0)\cap\mathcal{X}^*} \langle \pi, g\rangle \geq \mathbb{E}_{\xi\sim\mathrm{P}}[h(\xi)].$$

3. Finally, if $\mathrm{P}-a.s.\ g(\xi,\cdot)\in L^\infty(\pi_0(\cdot|\xi))$, *i.e.*, if h is finite $\mathrm{P}-a.s.$ *but* $h\notin L^1(\mathrm{P})$. By Remark A.13, this means that $\mathbb{E}_{\mathrm{P}}[h]=+\infty$. By mimicking the proof of the previous point but with $h_n(\xi,\zeta)\coloneqq h(\xi)\,\mathbb{1}_{h(\xi)\le n}+g(\xi,\zeta)\,\mathbb{1}_{h(\xi)>n}$ instead of h, we obtain that, for any $n\ge 1$,

$$\sup_{\pi\in\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)\cap L^0(\pi_0)\cap\mathcal{X}^*}\langle\pi,g\rangle\ge\mathbb{E}_{\xi\sim\pi_0}[h_n(\xi,\zeta)].$$

Now, by applying the monotone convergence theorem to h_n-g, which is non-negative $\pi_0-a.s.$, and with $h\in L^1(\pi_0)$ by assumption, we get that

$$\sup_{\pi\in\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)\cap L^0(\pi_0)\cap\mathcal{X}^*}\langle\pi,g\rangle\ge\mathbb{E}_{\xi\sim\mathrm{P}}[h(\xi)]=+\infty,$$

which concludes the proof. □

Lemma A.14. *Under the assumptions of Theorem A.8, the problem*

$$\sup\left\{\mathbb{E}_{\pi_2}[f]-R(\pi):\pi\in\overline{\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)\cap L^0(\pi_0)\cap\mathcal{X}^*}^{*},\langle\pi,c\rangle+S(\pi)\le\rho\right\}$$

coincides with its dual

$$\inf\left\{\lambda\rho+\mathbb{E}_{\xi\sim\mathrm{P}}\left[\operatorname*{ess\,sup}_{\pi_0(\cdot|\xi)}\{f-\lambda c(\xi,\cdot)-g(\xi,\cdot)\}\right]+\left(R|_{\mathcal{X}^*}+\lambda S|_{\mathcal{X}^*}\right)_*(g):g\in\mathcal{X},\,\lambda\ge 0\right\}.$$

Proof. We apply Lemma A.10 with $\mathcal{X}$ and $g\leftarrow R|_{\mathcal{X}^*}$, $s\leftarrow S|_{\mathcal{X}^*}$ which are convex, proper and $\sigma(\mathcal{X}^*,\mathcal{X})$-l.s.c.. Indeed, $\sigma(\mathcal{X}^*,L^{H_c})$ is weaker than $\sigma(\mathcal{X}^*,\mathcal{X})$, and $h\leftarrow\left\langle\cdot,-\widetilde{f}\right\rangle+\iota_{\overline{\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)\cap L^0(\pi_0)\cap\mathcal{X}^*}^{*}}$, $l\leftarrow\langle\cdot,c\rangle-\rho$ are convex, proper and weakly-$\star$ l.s.c. by construction. For $\lambda\ge 0$, we have

$$\begin{aligned}\forall g\in\mathcal{X},\quad(h+\lambda l)_*(g)&=\lambda\rho+\sup_{\pi\in\overline{\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)\cap L^0(\pi_0)\cap\mathcal{X}^*}^{*}}\langle\pi,g\rangle-\left\langle\pi,\lambda c-\widetilde{f}\right\rangle\\&=\lambda\rho+\sigma_{\overline{\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)\cap L^0(\pi_0)\cap\mathcal{X}^*}^{*}}\left(\widetilde{f}-\lambda c+g\right)\\&=\lambda\rho+\sigma_{\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)\cap L^0(\pi_0)\cap\mathcal{X}^*}\left(\widetilde{f}-\lambda c+g\right)\\&=\lambda\rho+\mathbb{E}_{\xi\sim\mathrm{P}}\left[\operatorname*{ess\,sup}_{\pi_0(\cdot|\xi)}\{f-\lambda c(\xi,\cdot)-g(\xi,\cdot)\}\right]\end{aligned}$$

where the last equality comes from Lemma A.12. The qualification condition of Lemma A.10 writes,

$$\left\{g\in\mathcal{X}:\exists\lambda\ge 0,\,g\in\operatorname{dom}(R+\lambda S)_*+\operatorname{dom}\sigma_{\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)\cap L^0(\pi_0)\cap\mathcal{X}^*}-\left(\widetilde{f}-\lambda c\right)\right\}=\mathcal{X}.\tag{A.2}$$

Take any $g\in\mathcal{X}$ and let us show that it actually belongs to this set. Since $\widetilde{f}$ also belongs to $\mathcal{X}$ by assumption, there exists $h\in L^{H_R}_{\heartsuit}$ some constant $\alpha>0$ such that, π_0 everywhere

$$g(\xi,\zeta)+\widetilde{f}(\xi,\zeta)-h(\xi,\zeta)\le\alpha(1+c_0(\xi,\zeta)+c(\xi,\zeta)).$$

Hence, with $\lambda\coloneqq\alpha$,

$$g(\xi,\zeta)+\widetilde{f}(\xi,\zeta)-h(\xi,\zeta)-\alpha c(\xi,\zeta)\le\alpha(1+c(\xi,\xi_0)).$$

But the RHS belongs to $\operatorname{dom} \sigma_{\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)\cap L^{H^*}}$ since $\mathbb{E}_{\xi\sim\mathrm{P}}[c(\xi,\xi_0)] < +\infty$. Then the LHS is inside the domain, as well. Moreover, h belongs to $L^{H_R}_{\heartsuit}$ so that it lies inside $\operatorname{dom}(R+\lambda S)_* = \frac{1}{1+\lambda\delta/\varepsilon}\operatorname{dom} R$. Hence, we have shown that g belongs to the LHS of (A.2), which shows the equality. □

Lemma A.15. *Assume that c is continuous on Ξ^2. For any $\pi \in \mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)$ such that $\mathbb{E}_\pi[c]$ is finite, there exists $\pi_t \in \mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)$ for $t = 1, 2, \dots$ such that*

1. *π_t is absolutely continuous w.r.t. π_0 for $t = 1, 2, \dots$,*
2. *$(\pi_t)_t$ converges weakly to π, i.e., w.r.t. the topology $\sigma(\mathcal{M}(\Xi\times\Xi), \mathcal{C}_b(\Xi\times\Xi))$,*
3. *$\mathbb{E}_{\pi_t}[c] \to \mathbb{E}_\pi[c]$ as $t\to\infty$.*

In particular, for any $g \in L^0(\pi_0)$ l.s.c. such that $g_- \in L^{H_c}$,

$$\mathbb{E}_\pi[g] \le \liminf_{t\to\infty} \mathbb{E}_{\pi_t}[g],$$

and, as a consequence, $(\pi)_t$ also converges to π for the topology $\sigma((L^H_c)^*, \mathcal{C}(\Xi\times\Xi)\cap L^{H_c})$.

Proof. Step 1: building a partition of Ξ^2. For convenience, denote by $\|\cdot\|$ a norm on $\mathbb{R}^d$, which contains Ξ and by $\overline{\mathbb{B}}(0,r)$ the closed ball of radius r around 0 in $\mathbb{R}^d$. Fix $\Delta > 0$. By absolute continuity of c on compact sets, for any $n, m \ge 1$, there exists $\eta_{n,m} \in (0,\Delta]$ such that, for any $\xi \in \Xi\cap\overline{\mathbb{B}}(0,n)$, ζ, $\zeta' \in \Xi\cap\overline{\mathbb{B}}(0,m)$,

$$\|\zeta-\zeta'\| \le \eta_{n,m} \implies |c(\xi,\zeta') - c(\xi,\zeta)| \le \Delta.$$

Fix n and m greater or equal than 1. We consider a finite family of disjoint open sets $\mathcal{U}^m_i$ for $i \in I^m$ which are included in $\mathbb{B}(0,m)\setminus\overline{\mathbb{B}}(0,m-1)$, whose diameters are at most $\eta_{n,m}$ and which satisfy $\overline{\bigcup_{i\in I^m}\mathcal{U}^m_i} \supset \mathbb{B}(0,m)\setminus\overline{\mathbb{B}}(0,m-1)$. Define $J^m = \{i \in I^m : \mathcal{U}^m_i \cap \operatorname{int}\Xi \neq \emptyset\}$. We first show

$$\overline{\bigcup_{i\in J^m} \mathcal{U}^m_i \cap \operatorname{int}\Xi} \supset \big(\mathbb{B}(0,m)\setminus\overline{\mathbb{B}}(0,m-1)\big)\cap\Xi \tag{A.3}$$

by considering $\zeta \in (\mathbb{B}(0,m)\setminus\overline{\mathbb{B}}(0,m-1))\cap\Xi$ and $\mathcal{U}$ neighborhood of ζ. We thus have to prove that $\bigcup_{i\in J^m}\mathcal{U}^m_i \cap \operatorname{int}\Xi\cap\mathcal{U}$ is not empty. Since $\mathbb{B}(0,m)\setminus\overline{\mathbb{B}}(0,m-1)$ is an open set that contains ζ, it suffices to show this for some $\mathcal{U} \subset (\mathbb{B}(0,m)\setminus\overline{\mathbb{B}}(0,m-1))$. This means that, in particular, $\mathcal{U} \subset \overline{\bigcup_{i\in I^m}\mathcal{U}^m_i}$. Because ζ belongs to Ξ with $\Xi = \overline{\operatorname{int}\Xi}$, it holds that $\mathcal{U}\cap\operatorname{int}\Xi \neq \emptyset$. Therefore, we have that $\mathcal{U}\cap\operatorname{int}\Xi\cap\overline{\bigcup_{i\in I^m}\mathcal{U}^m_i} \neq \emptyset$. Since $\mathcal{U}\cap\operatorname{int}\Xi$ and $\bigcup_{i\in I^m}\mathcal{U}^m_i$ are both open, this implies that $\mathcal{U}\cap\operatorname{int}\Xi\cap\bigcup_{i\in I^m}\mathcal{U}^m_i \neq \emptyset$, which concludes the proof of (A.3). Moreover, since J^m is finite, $\overline{\bigcup_{i\in J^m}\mathcal{U}^m_i\cap\operatorname{int}\Xi} = \bigcup_{i\in J^m}\overline{\mathcal{U}^m_i\cap\operatorname{int}\Xi}$ so that $\bigcup_{i\in J^m}\overline{\mathcal{U}^m_i\cap\operatorname{int}\Xi} \supset \big(\mathbb{B}(0,m)\setminus\overline{\mathbb{B}}(0,m-1)\big)\cap\Xi$ holds as well.

Still keeping $n \ge 1$ fixed, we have built an (at most) countable number of disjoint open sets $(\mathcal{V}_i)_{i\le M_n} \coloneqq (\mathcal{U}^m_i\cap\operatorname{int}\Xi)_{i\in J^m,\, m\ge1}$, with $M_n \in \mathbb{N}\cup\{+\infty\}$ such that for any $i \le M_n$, there is some $m \ge 1$ such that $\mathcal{V}_i \subset \mathbb{B}(0,m)\setminus\overline{\mathbb{B}}(0,m-1)\cap\operatorname{int}\Xi$ and such that the diameter of $\mathcal{V}_i$ is at most $\eta_{n,m}$. Moreover, $\bigcup_{i\in I}\overline{\mathcal{V}_i}$ is the whole of Ξ. We build a partition of Ξ by recursively defining

$$\mathcal{W}^n_i = \overline{\mathcal{V}_i}\setminus\bigcup_{j\le i-1}\mathcal{W}^n_j = \mathcal{V}_i \cup \left(\partial\mathcal{V}_i \setminus \bigcup_{j\le i-1}\mathcal{W}^n_j\right) \text{ for } i \le M_n.$$

Therefore, we have built an (at most) countable number of disjoint measurable sets $\mathcal{W}^n_i$ for $i \le M_n$, whose union is Ξ and such that for any $i \le M_n$, there is some $m \ge 1$ such that $\mathcal{W}^n_i \subset \overline{\mathbb{B}}(0,m)\cap\Xi$. Moreover, the diameter of each $\mathcal{W}^n_i$ is at most $\eta_{n,m}$ and each of these sets has non-empty interior.

The full partition of Ξ^2 that we finally consider is made of the sets $((\mathbb{B}(0,n)\setminus\mathbb{B}(0,n-1))\cap\Xi)\times\mathcal{W}_i^n$ for $n\geq 1$ and $i\leq M_n$. To summarize, we have built, for $\Delta>0$, a partition of the form $(A_i^\Delta\times B_j^\Delta)_{i\in I,\, j\in J(i)}$ which is a (at most) countable and measurable partition of Ξ^2 satisfying:

1. $(A_i^\Delta)_{i\in I}$ is a partition of Ξ and, for any $i\in I$, $(B_j^\Delta)_{j\in J(i)}$ is a partition of Ξ.
2. For $i\in I$, $j\in J(i)$, $\xi\in A_i^\Delta$, and $\zeta,\zeta'\in B_j^\Delta$, we have $\|\zeta-\zeta'\|\leq\Delta$ and $|c(\xi,\zeta')-c(\xi,\zeta)|\ \leq\Delta$.
3. For any $i\in I$, $j\in J(i)$, B_j^Δ has non-empty interior.

Step 2: construction of the sequence of measures. For $\Delta>0$, we define the measure $\pi^\Delta\in\mathcal{P}_{\mathrm{P}}(\Xi\times\Xi)$ for any $\xi\in\Xi$ and any measurable $B\subset\Xi$ by

$$\pi^\Delta(B|\xi)\coloneqq\sum_{i\in I}\mathbb{1}_{\xi\in A_i^\Delta}\sum_{j\in J(i)}\frac{\pi(B_j^\Delta|\xi)}{\pi_0(B_j^\Delta|\xi)}\pi_0(B\cap B_j^\Delta|\xi),$$

which is well-defined because of the B_j^Δ having non-empty interiors. Moreover, by construction, it is absolutely continuous w.r.t. π_0, so that π^Δ satisfies (1).

For $i\in I$ and $j\in J(i)$, choose $\zeta_{i,j}^\Delta\in B_j^\Delta$. By the Portmanteau theorem, see [21], Theorem 13.16, it suffices to show that $\int g\,\mathrm{d}\,\pi^\Delta$ converges to $\int g\,\mathrm{d}\,\pi$ as $\Delta\to 0$ for any bounded Lipschitz function g to prove the weak convergence of the sequence of measures (π^Δ) to π. Thus, take a bounded 1-Lipschitz function $g:\Xi^2\to\mathbb{R}$, then

$$\begin{aligned}
\left|\int g\,\mathrm{d}\,\pi^\Delta-\int g\,\mathrm{d}\,\pi\right|\leq&\mathbb{E}_{\xi\sim\mathrm{P}}\left[\sum_{i\in I}\mathbb{1}_{\xi\in A_i^\Delta}\sum_{j\in J(i)}\left|\frac{\pi(B_j^\Delta|\xi)}{\pi_0(B_j^\Delta|\xi)}\int_{B_j^\Delta}g(\xi,\zeta)\pi_0(\mathrm{d}\zeta|\xi)-\int_{B_j^\Delta}g(\xi,\zeta)\,\pi(\mathrm{d}\zeta|\xi)\right|\right]\\
\leq&\mathbb{E}_{\xi\sim\mathrm{P}}\left[\sum_{i\in I}\mathbb{1}_{\xi\in A_i^\Delta}\sum_{j\in J(i)}\pi(B_j^\Delta|\xi)\left(\frac{1}{\pi_0(B_j^\Delta|\xi)}\int_{B_j^\Delta}\left|g(\xi,\zeta)-g(\xi,\zeta_{i,j}^\Delta)\right|\pi_0(\mathrm{d}\zeta|\xi)\right)\right]\\
&+\mathbb{E}_{\xi\sim\mathrm{P}}\left[\sum_{i\in I}\mathbb{1}_{\xi\in A_i^\Delta}\sum_{j\in J(i)}\pi(B_j^\Delta|\xi)\left(\frac{1}{\pi(B_j^\Delta|\xi)}\int_{B_j^\Delta}\left|g(\xi,\zeta)-g(\xi,\zeta_{i,j}^\Delta)\right|\,\pi(\mathrm{d}\zeta|\xi)\right)\right]\\
\leq&\mathbb{E}_{\xi\sim\mathrm{P}}\left[\sum_{i\in I}\mathbb{1}_{\xi\in A_i^\Delta}\sum_{j\in J(i)}\pi(B_j^\Delta|\xi)\left(\frac{1}{\pi_0(B_j^\Delta|\xi)}\int_{B_j^\Delta}\|\zeta-\zeta_{i,j}^\Delta\|\pi_0(\mathrm{d}\zeta|\xi)\right)\right]\\
&+\mathbb{E}_{\xi\sim\mathrm{P}}\left[\sum_{i\in I}\mathbb{1}_{\xi\in A_i^\Delta}\sum_{j\in J(i)}\pi(B_j^\Delta|\xi)\left(\frac{1}{\pi(B_j^\Delta|\xi)}\int_{B_j^\Delta}\|\zeta-\zeta_{i,j}^\Delta\|\,\pi(\mathrm{d}\zeta|\xi)\right)\right]\\
\leq&2\mathbb{E}_{\xi\sim\mathrm{P}}\left[\sum_{i\in I}\mathbb{1}_{\xi\in A_i^\Delta}\sum_{j\in J(i)}\pi(B_j^\Delta|\xi)\Delta\right]=2\Delta,
\end{aligned}$$

which vanishes as Δ goes to 0. This proves point (2) of the result.

Using the same decomposition of the integral, we get the same inequality for c,

$$\left|\int c\,\mathrm{d}\,\pi^\Delta-\int c\,\mathrm{d}\,\pi\right|\leq\Delta,$$

which shows the convergence of the integral $\mathbb{E}_{\pi^\Delta}[c]$ to $\mathbb{E}_\pi[c]$, which is point (3) of the result.

Finally, take $g \in L^0(\pi_0)$ l.s.c. and such that $g_- \in L^{H_c}$, this means that there exists $\alpha > 0$ such that, π_0 *a.s.*, $g \geq -\alpha(1 + c_0 + c)$. But by construction, $\mathbb{E}_{\pi^\Delta}[1 + c_0 + c] = 1 + \mathbb{E}_{\mathrm{P}}[c(\cdot, \xi_0)] + \mathbb{E}_{\pi_\Delta}[c]$ which goes to $\mathbb{E}_\pi[1 + c_0 + c]$ as $\Delta \to 0$. The last part of the result then follows from [38], Lemma 4.3 with $c \leftarrow -\alpha(1 + c_0 + c)$ and $h \leftarrow g$. □

Lemma A.16. *Under the assumptions of Theorem A.8, the values of the two problems coincide:*

$$\sup\left\{\langle \pi, \widetilde{f}\rangle - R(\pi) : \pi \in \overline{\mathcal{P}_{\mathrm{P}}(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*}^*, \langle \pi, c\rangle + S(\pi) \leq \rho\right\} \tag{A.4}$$

and

$$\sup\left\{\langle \pi, \widetilde{f}\rangle - R(\pi) : \pi \in \mathcal{P}_{\mathrm{P}}(\Xi \times \Xi), \langle \pi, c\rangle + S(\pi) \leq \rho\right\}.$$

Proof. We divide the proof in two parts, based on the two regularity assumption in Theorem A.8.
Case (*a*): *R is $\sigma(\mathcal{X}^*, \mathcal{C}(\Xi \times \Xi) \cap L^{H_c})$-u.s.c., f is l.s.c. and $\widetilde{f}_- \in L^{H_c}$.* Thanks to the existence of a strictly feasible point π_S ($\langle \pi_S, c\rangle + S(\pi_S) < \rho$) and the convexity (resp. concavity) of the constraint (resp. objective), Lemma A.11 ensures that the values of the two problems do not change when the inequality constraints are replaced by strict inequality constraints.

The value of (A.4) is thus equal to

$$\begin{aligned}
&\sup\left\{\langle \pi, \widetilde{f}\rangle - R(\pi) : \pi \in \overline{\mathcal{P}_{\mathrm{P}}(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*}^*, \langle \pi, c\rangle + S(\pi) < \rho\right\} \\
=&\sup\left\{\langle \pi, \widetilde{f}\rangle - R(\pi) - \iota_{\langle \cdot, c\rangle + S < \rho}(\pi) : \pi \in \overline{\mathcal{P}_{\mathrm{P}}(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*}^*\right\},
\end{aligned}$$

with $\langle \cdot, \widetilde{f}\rangle - R - \iota_{\langle \cdot, c\rangle + S < \rho}$ now $\sigma(\mathcal{X}^*, \mathcal{X})$-l.s.c. since $\mathcal{C}(\Xi \times \Xi) \cap L^{H_c} \subset \mathcal{X}$ so that $\sigma(\mathcal{X}^*, \mathcal{C}(\Xi \times \Xi) \cap L^{H_c})$ is weaker than $\sigma(\mathcal{X}^*, \mathcal{X})$. Thus, the weak-$*$ closure can be removed, and (A.4) is equal to

$$\begin{aligned}
&\sup\left\{\langle \pi, \widetilde{f}\rangle - R(\pi) - \iota_{\langle \cdot, c\rangle + S < \rho}(\pi) : \pi \in \mathcal{P}_{\mathrm{P}}(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*\right\} \\
=&\sup\left\{\langle \pi, \widetilde{f}\rangle - R(\pi) : \pi \in \mathcal{P}_{\mathrm{P}}(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*, \langle \pi, c\rangle + S(\pi) < \rho\right\} \\
=&\sup\left\{\langle \pi, \widetilde{f}\rangle - R(\pi) : \pi \in \mathcal{P}_{\mathrm{P}}(\Xi \times \Xi) \cap L^0(\pi_0), \langle \pi, c\rangle + S(\pi) < \rho\right\},
\end{aligned}$$

where the last equality follows from Lemma A.7 since $S(\pi)$ and $\langle \pi, c\rangle$ have to be both finite for π to be feasible and $R \propto S$. All that is left to show is the equality

$$\begin{aligned}
&\sup\left\{\langle \pi, \widetilde{f}\rangle - R(\pi) : \pi \in \mathcal{P}_{\mathrm{P}}(\Xi \times \Xi) \cap L^0(\pi_0), \langle \pi, c\rangle + S(\pi) < \rho\right\} \\
=&\sup\left\{\langle \pi, \widetilde{f}\rangle - R(\pi) : \pi \in \mathcal{P}_{\mathrm{P}}(\Xi \times \Xi), \langle \pi, c\rangle + S(\pi) < \rho\right\}.
\end{aligned}$$

The inequality ($\leq$) is straightforward. Concerning the other side, for any $\pi \in \mathcal{P}_{\mathrm{P}}(\Xi \times \Xi)$ such that $\langle \pi, c\rangle + S(\pi) < \rho$, we can consider $(\pi_t)_t$ the sequence of probability distributions in $\mathcal{P}_{\mathrm{P}}(\Xi \times \Xi) \cap L^0(\pi_0)$ given by Lemma A.15. Thanks to c belonging to $\mathcal{C}(\Xi \times \Xi) \cap L^{H_c}$ and the u.s.c. assumption on $S \propto R$, $\limsup_{t \to \infty} \langle \pi_t, c\rangle + S(\pi_t) \leq \langle \pi, c\rangle + S(\pi) < \rho$ showing that π_t is strictly feasible after some time. Moreover, since $\widetilde{f}$ is l.s.c. and $\widetilde{f}_-$ belongs to L^{H_c}, we get that

$$\begin{aligned}
\langle \pi, \widetilde{f}\rangle - R(\pi) &\leq \liminf_{t \to \infty} \langle \pi_t, \widetilde{f}\rangle - R(\pi_t) \\
&\leq \sup\left\{\langle \pi, \widetilde{f}\rangle - R(\pi) : \pi \in \mathcal{P}_{\mathrm{P}}(\Xi \times \Xi) \cap L^0(\pi_0), \langle \pi, c\rangle + S(\pi) < \rho\right\},
\end{aligned}$$

which achieves the proof.

Case (2): $\operatorname{dom} R \subset L^0(\pi_0)$ *and, for any sequence* $g_t \in \operatorname{dom} H_R$ *non-negative such that* $g_t \to 0$ *as* $t \to \infty$, $H_R(g_t) \to 0$. In this case, H_R satisfies (H4) of [28], Theorem 22 so that $(L^{H_R}_\heartsuit)^* = L^{H_R^*}$. As a consequence, $\mathcal{P}_{\mathrm{P}}(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*$ is $\sigma(\mathcal{X}^*, \mathcal{X})$ closed so that

$$\overline{\mathcal{P}_{\mathrm{P}}(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*}^* = \mathcal{P}_{\mathrm{P}}(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*.$$

Therefore, using Lemma A.7, one gets that

$$\sup \left\{ \langle \pi, \widetilde{f} \rangle - R(\pi) : \pi \in \mathcal{P}_{\mathrm{P}}(\Xi \times \Xi) \cap L^0(\pi_0), \langle \pi, c \rangle + S(\pi) \leq \rho \right\},$$

and the result follows from the assumption on the domain of R. □

A.5 Examples

We finally develop the two examples of Section 2.2 by choosing carefully the coupling π_0.

Corollary A.17 (Cost-regularized WDRO)**.** *Let Assumption A.1 holds. Suppose that the function* $f : \Xi \to \mathbb{R}_+$ *in* $L^1(\mathrm{P})$*, l.s.c., and such that the function* $(\xi, \zeta) \mapsto f(\zeta)/\big(1 + c_0(\xi, \zeta) + c(\xi, \zeta)\big)$ *is bounded, above and below, over* Ξ^2*. Then, for any* $\varepsilon, \delta > 0$*, the problem*

$$\sup \{ \mathbb{E}_{\pi_2}[f] - \varepsilon \mathbb{E}_\pi[c] : \pi \in \mathcal{P}_{\mathrm{P}}(\Xi \times \Xi), \mathbb{E}_\pi[c] + \delta \mathbb{E}_\pi[c] \leq \rho \}$$

coincides with its dual

$$\inf \left\{ \lambda \rho + \mathbb{E}_{\xi \sim \mathrm{P}} \left[\operatorname*{ess\,sup}_{Leb} \{ f - (\varepsilon + (1+\delta)\lambda) c(\xi, \cdot) \} \right] : \lambda \geq 0 \right\}.$$

Proof. In order to prove this result, we choose $\pi_0 \in \mathcal{P}_{\mathrm{P}}(\Xi \times \Xi)$ such that, for any $\xi \in \Xi$,

$$\pi_0(\mathrm{d}\zeta | \xi) \propto \frac{\mathbb{1}_\Xi(\zeta)}{1 + c_0(\xi, \zeta) + c(\xi, \zeta)} e^{-\|\zeta\|^2/2} \mathrm{d}\zeta.$$

We also define R for all $\pi \in (L^b(\Xi^2))^*$,

$$R(\pi) = \begin{cases} \varepsilon \langle \pi, c \rangle & \text{if } \pi \in (L^H_c)^* \text{ non-negative} \\ \varepsilon \int c \, \mathrm{d}\, \pi & \text{if } \pi \in \mathcal{M}(\Xi \times \Xi) \text{ non-negative} \\ +\infty & \text{otherwise.} \end{cases}$$

With this definition, R satisfies the assumption of Lemma A.5 with $\pi_R = 0$ and we have

$$H_R(g) = \sup_{\pi \in L^0(\pi_0) : \pi \geq 0} \int (|g| - \varepsilon c) \, \pi \, \mathrm{d}\pi_0 = \iota_{|g| \leq \varepsilon c},$$

and thus $L^{H_R}_\heartsuit = \{0\}$ so that $\mathcal{X}$ is equal to L^{H_c}.

It remains to check the regularity conditions on R. $R|_{\mathcal{X}^*}$ is $\sigma(\mathcal{X}^*, \mathcal{X})$-l.s.c. since c belongs to $\mathcal{X}$ and the set of non-negative linear forms in $\mathcal{X}^*$ is $\sigma(\mathcal{X}^*, \mathcal{X})$-closed. On $\mathcal{P}_{\mathrm{P}}(\Xi \times \Xi)$, R coincides with $\pi \mapsto \varepsilon \int c \, \mathrm{d}\, \pi$ and c is inside $\mathcal{C}(\Xi \times \Xi) \cap L^{H_c}$ so it is actually $\sigma(\mathcal{X}^*, \mathcal{C}(\Xi \times \Xi) \cap L^{H_c})$ continuous and thus regularity condition (a) of

Theorem A.8 holds. Finally, the strict feasibility condition is satisfied thanks to the distribution $\pi(\mathrm{d}\xi,\mathrm{d}\zeta) = \mathrm{P}(\mathrm{d}\xi)\delta_\xi(\mathrm{d}\zeta)$ so Theorem A.8 applies. Since $L^\infty(\pi_0) \subset \mathcal{X}^*$, the preconjugates of $R + \lambda S$ is given by,

$$(R|_{\mathcal{X}^*} + \lambda S|_{\mathcal{X}^*})_*(g) = \iota_{g \leq (\varepsilon+\lambda\delta)c}.$$

Hence, since the function $(\varepsilon + \lambda\delta)c$ belongs to $\mathcal{X} = L^{H_c}$,

$$\begin{aligned}&\inf\left\{\lambda\rho + \mathbb{E}_{\xi\sim\mathrm{P}}\left[\operatorname*{ess\,sup}_{\pi_0(\cdot|\xi)}\{f - \lambda c(\xi,\cdot) - g(\xi,\cdot)\}\right] + (R|_{\mathcal{X}^*} + \lambda S|_{\mathcal{X}^*})_*(g) : g \in \mathcal{X}\right\}\\ &= \mathbb{E}_{\xi\sim\mathrm{P}}\left[\operatorname*{ess\,sup}_{Leb}\{f - (\varepsilon + (1+\delta)\lambda)c(\xi,\cdot)\}\right],\end{aligned}$$

which gives the result. □

Corollary A.18 (ϕ-divergence-regularized WDRO)**.** *Let Assumption A.1 hold and take $\varepsilon, \delta > 0$. Consider $\pi_0 \in \mathcal{P}(\Xi\times\Xi)$ with $\mathbb{E}_{\pi_0}[c] < +\infty$, a convex l.s.c. function $\phi : \mathbb{R} \to \mathbb{R}\cup\{+\infty\}$ with such that $\phi(1) = 0$, $[1,+\infty) \subset \operatorname{dom}\phi \subset [0,+\infty)$, $\phi'(+\infty) = +\infty$ and define the associated divergence*

$$\forall\,\pi \in \mathcal{P}(\Xi\times\Xi), \qquad D_\phi(\pi\,|\pi_0) \coloneqq \begin{cases}\int_{\Xi^2}\phi\left(\frac{\mathrm{d}\,\pi}{\mathrm{d}\pi_0}\right)\mathrm{d}\pi_0 & \text{if } \pi \text{ is absolutely continuous w.r.t. } \pi_0\\ +\infty & \text{otherwise.}\end{cases}$$

Assume also that f satisfies the growth condition:

$$\exists g \in L^{H_c}, \forall\alpha > 0, \int_{\Xi^2}\phi^*\Big(\alpha|\widetilde{f} - g|\Big)\mathrm{d}\pi_0 < +\infty.$$

Then, with $R \leftarrow \varepsilon D_\phi(\,\cdot\,|\pi_0)$ and $S \leftarrow \delta D_\phi(\,\cdot\,|\pi_0)$, if (R-WDRO) is strictly feasible, its value is equal to

$$\inf_{\lambda\geq 0}\inf_{\psi\in\mathcal{C}(\Xi\times\Xi)}\lambda\rho + \mathbb{E}_{\xi\sim\mathrm{P}}\left[\sup_{\zeta\in\Xi} f(\zeta) - \lambda c(\xi,\zeta) - \psi(\xi,\zeta)\right] + (\varepsilon+\lambda\delta)\int_{\Xi^2}\phi^*\left(\frac{\psi(\xi,\zeta)}{\varepsilon+\lambda\delta}\right)\mathrm{d}\pi_0(\xi,\zeta).$$

Proof. D_ϕ admits the following variational formula, see *e.g.*, [1], Proposition 4.2.8 applied with the pair $(L^\infty(\pi_0), \mathcal{M}(\Xi\times\Xi))$,

$$\forall\,\pi \in \mathcal{M}(\Xi\times\Xi), \quad D_\phi(\pi\,|\pi_0) = \sup\left\{\langle\pi,\psi\rangle - \int_{\Xi^2}\phi^*\circ\psi\;\mathrm{d}\pi_0 : \psi \in L^\infty(\pi_0)\right\}.$$

As seen in the second case of Lemma A.16, $\operatorname{dom} R \subset L^0(\pi_0)$ so that $\mathcal{X}^* \subset L^0(\pi_0)$. Hence, this variational formula applies to the whole $\mathcal{X}^*$ and since $L^\infty(\pi_0) \subset L^{H_c} \subset \mathcal{X}$, $D_\phi(\cdot|\pi_0)$ is $\sigma(\mathcal{X}^*,\mathcal{X})$-l.s.c.. Moreover, [1], Proposition 4.2.6 with the $(\mathcal{X}, \mathcal{M}(\Xi\times\Xi))$ and $(\mathcal{X},\mathcal{X}^*)$ gives

$$H_R(g) = (R|_{\mathcal{X}^*})_*(|g|) = \int_{\Xi^2}\phi^*\circ g\;\mathrm{d}\pi_0.$$

The continuity at 0 of H_R follows from the dominated convergence theorem and the fact that ϕ^* is non-decreasing. The growth condition means that $\widetilde{f}$ lies in $\mathcal{X}$, and we can apply Theorem A.8. □

Acknowledgements. We thank the associate editor and the two referees for their careful reading and numerous suggestions. In particular, a referee suggested us to study duality beyond compactness/continuity, which has led to the developments

sketched in Section 2.3 and detailed in Appendix. This work has been supported by MIAI Grenoble Alpes (ANR-19-P3IA-0003).

References

[1] R. Agrawal and T. Horel, Optimal bounds between f-divergences and integral probability metrics. *J. Mach. Learn. Res.* **22** (2021) 5662–5720.

[2] Y. An and R. Gao, Generalization bounds for (Wasserstein) robust Optimization. *Adv. Neural Inform. Process. Syst.* **34** (2021).

[3] J. Blanchet and Y. Kang, Semi-supervised learning based on distributionally robust optimization. *Data Anal. Applic. 3: Comput. Classif. Finan. Stat. Stochast. Methods* **5** (2020) 1–33.

[4] J. Blanchet, Y. Kang and K. Murthy, Robust Wasserstein profile inference and applications to machine learning. *J. Appl. Probab.* **56** (2019) 830–857.

[5] J. Blanchet and K. Murthy, Quantifying distributional model risk via optimal Transport. *Math. Oper. Res.* **44** (2019) 565–600.

[6] J. Blanchet, K. Murthy and F. Zhang, Optimal transport-based distributionally robust optimization: structural properties and iterative schemes. *Math. Oper. Res.* (2021).

[7] V.I. Bogachev, Measure Theory, Vol. 1, 1st edn. Springer (2007).

[8] R.I. Bot, S.-M. Grad and G. Wanka, Duality in Vector Optimization, Vector Optimization Springer Berlin Heidelberg (2009).

[9] S. Boucheron, O. Bousquet, G. Lugosi and P. Massart, Moment inequalities for functions of independent random variables. *Ann. Probab.* **33** (2005) 514–560.

[10] H. Brezis, Functional Analysis, Sobolev Spaces and Partial Differential Equations Springer New York (2010).

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

[12] R. Chen and I.C. Paschalidis, Distributionally robust learning. *Found. Trends Optim.* **4** (2020) 1–243.

[13] C. Clason and T. Valkonen, Introduction to Nonsmooth Analysis and Optimization, arXiv:2001.00216 (2020).

[14] M. Cuturi, Sinkhorn distances: lightspeed computation of optimal transport, *Adv. Neural Inform. Process. Syst.* Vol. 26 Curran Associates, Inc. (2013).

[15] P.M. Esfahani and D. Kuhn, Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Math. Programm.* **171** (2018) 115–166.

[16] R. Gao and A. Kleywegt, Distributionally robust stochastic optimization with Wasserstein distance. *Math. Oper. Res.* (2022).

[17] A. Genevay, L. Chizat, F. Bach, M. Cuturi and G. Peyré, Sample complexity of Sinkhorn divergences, in 22nd International Conference on Artificial Intelligence and Statistics, PMLR (2019) 1574–1583.

[18] A. Genevay, M. Cuturi, G. Peyré and F. Bach, Stochastic optimization for large-scale optimal transport, in *Adv. Neural Inform. Process Syst.* (2016).

[19] O. Kallenberg, Random Measures, Theory and Applications, Probability Theory and Stochastic Modelling. Springer International Publishing (2017).

[20] N.J. Kalton, N.T. Peck and J.W. Roberts, An F-space Sampler, London Mathematical Society Lecture Note Series. Cambridge University Press (1984).

[21] A. Klenke, *Probability Theory*: A Comprehensive Course, Universitext. Springer (2014).

[22] D. Kuhn, P.M. Esfahani, V.A. Nguyen and S. Shafieezadeh-Abadeh, Wasserstein distributionally robust optimization: theory and applications in machine learning, in Operations Research & Management Science in the Age of Analytics. INFORMS (2019).

[23] S.Y. Lee, Gibbs sampler and coordinate ascent variational inference: a set-theoretical review. *Commun. Stat. Theory Methods* (2021) 1–21.

[24] A. Lunardi, Interpolation Theory, Vol. 9. Springer (2009).

[25] Q. Merigot and B. Thibert, Optimal transport: discretization and algorithms, in Handbook of Numerical Analysis, Vol. 22. Elsevier (2021) 133–212.

[26] J. Musielak, Orlicz Spaces and Modular Spaces, Vol. 1034. Springer (2006).

[27] F.-P. Paty and M. Cuturi, Regularized optimal transport is ground cost adversarial, in International Conference on Machine Learning, PMLR (2020) 7532–7542.

[28] T. Pennanen and A.-P. Perkkiö, Topological duals of locally convex function Spaces. *Positivity* **26** (2022) 1–38.

[29] J. Peypouquet, Convex Optimization in Normed Spaces: Theory, Methods and Examples, Springer Briefs in Optimization. Springer International Publishing (2015).

[30] G. Peyré, M. Cuturi, Computational optimal transport: with applications to data science. *Found. Trends Mach. Learn.* **11** (2019) 355–607.

[31] R.T. Rockafellar and R.J.-B. Wets, Variational Analysis, Grundlehren Der Mathematischen Wissenschaften. Springer-Verlag (1998).

[32] W. Rudin, Real and Complex Analysis. McGraw-Hill (1987).

[33] W. Rudin, Functional Analysis, International Series in Pure and Applied Mathematics. McGraw-Hill (1991).

[34] F. Santambrogio, Optimal transport for applied mathematicians, *Progress in Nonlinear Differential Equations and Their Applications*, Vol. 87. Springer International Publishing (2015).

[35] S. Shafieezadeh-Abadeh, D. Kuhn and P.M. Esfahani, regularization via mass transportation. *J. Mach. Learn. Res.* **20** (2019) 1–68.
[36] A. Sinha, H. Namkoong and J. Duchi, Certifying some distributional robustness with principled adversarial training, in International Conference on Learning Representations (2018).
[37] C. Villani, Topics in Optimal Transportation. American Mathematical Society (2003).
[38] C. Villani, Optimal Transport: Old and New. Springer (2008).
[39] J. Wang, R. Gao and Y. Xie, Sinkhorn Distributionally Robust Optimization, arXiv:2109.11926 (2021).
[40] Y. Yu, T. Lin, E.V. Mazumdar and M. Jordan, Fast distributionally robust learning with variance-reduced min-max optimization, in International Conference on Artificial Intelligence and Statistics, PMLR (2022) 1219–1250.
[41] C. Zalinescu, Convex Analysis in General Vector Spaces. World Scientific (2002).

Please help to maintain this journal in open access!

This journal is currently published in open access under the Subscribe to Open model (S2O). We are thankful to our subscribers and supporters for making it possible to publish this journal in open access in the current year, free of charge for authors and readers.

Check with your library that it subscribes to the journal, or consider making a personal donation to the S2O programme by contacting subscribers@edpsciences.org.

More information, including a list of supporters and financial transparency reports, is available at https://edpsciences.org/en/subscribe-to-open-s2o.