VARIATIONAL ANALYSIS IN THE WASSERSTEIN SPACE*

NICOLAS LANZETTI[†], ANTONIO TERPIN[‡], AND FLORIAN DÖRFLER[†]

Abstract. We study optimization problems whereby the optimization variable is a probability measure. Since the probability space is not a vector space, many classical and powerful methods for optimization (e.g., gradients) are of little help. Thus, one typically resorts to the abstract machinery of infinite-dimensional analysis or other ad-hoc methodologies – not tailored to the probability space – which however involve projections or rely on convexity-type assumptions. We believe instead that these problems call for a comprehensive methodological framework for calculus in probability spaces. In this work, we combine ideas from optimal transport, variational analysis, and Wasserstein gradient flows to equip the Wasserstein space (i.e., the space of probability measures endowed with the Wasserstein distance) with a variational structure, both by combining and extending existing results and introducing novel tools. Our theoretical analysis culminates in very general necessary optimality conditions for optimality. Notably, our conditions (i) resemble the rationales of Euclidean spaces, such as the Karush-Kuhn-Tucker conditions, and (ii) are intuitive, informative, and easy to study. We believe this framework lays the foundation for new algorithmic and theoretical advancements in the study of optimization problems in probability spaces.

Key words. variational analysis, optimal transport

MSC codes. 49J53, 49Q22

1. Main result.

This work considers the optimization problem

$$\inf_{\mu\in\mathcal{C}} \mathcal{J}(\mu), \tag{1.1}$$

where $\mathcal{C} \subseteq \mathcal{P}(\mathbb{R}^d)$ is a set of admissible probability measures and $\mathcal{J} : \mathcal{P}(\mathbb{R}^d) \to \bar{\mathbb{R}}$ is a functional to minimize. This abstract problem setting stems from the observation that numerous fields, including machine learning, robust optimization, and biology, tackle their own version of (1.1), but with ad hoc methods that often cease to be effective as soon as the problem structure changes. We believe, instead, that these problems still demand a comprehensive theory for the optimization problem (1.1), despite the recent efforts in the literature [46].

In this paper, we present a general and flexible toolbox for optimization in probability spaces. Specifically, we derive novel necessary first-order optimality conditions for (1.1), for arbitrary functionals and constraints. These formally resemble the rationales of Euclidean spaces (e.g., Karush–Kuhn–Tucker conditions) and are intuitive, informative, and easy to study. As a byproduct of our analysis, we translate tools from variational analysis (e.g., generalized subgradients, normal cones, etc.) to the Wasserstein space (i.e., the probability space endowed with the Wasserstein distance). After practicing these novel tools in numerous pedagogical examples, we tackle open problems arising in machine learning and Distributionally Robust Optimization (DRO).

Our main result are general first-order optimality conditions of (1.1):

*Submitted to the editors DATE. Nicolas Lanzetti and Antonio Terpin contributed equally to this work .

Funding: This research was supported by the Swiss National Science Foundation under the NCCR Automation, grant agreement 51NF40_180545.

[†]Nicolas Lanzetti and Florian Dörfler are with the Automatic Control Laboratory, Department of Information Technology and Electrical Engineering, ETH Zürich, Physikstrasse 3, 8092 Zürich, Zürich, Switzerland ({lnicolas,dorfler}@ethz.ch).

[‡]Antonio Terpin is with the Institute for Dynamic Systems and Control, Department of Mechanical and Process Engineering, ETH Zürich, Sonnegstrasse 3, 8092 Zürich, Zürich, Switzerland (aterpin@ethz.ch).

Informal statement 1.1 (First-order optimality conditions). *If $\mu^* \in \mathcal{P}(\mathbb{R}^d)$ is an optimal solution of* (1.1) *with finite second moment and provided that a constraint qualification holds, then the "Wasserstein subgradients" are "aligned" with the constraints at "optimality", i.e.,*

$$\mathbf{0}_{\mu^*} \in \partial \mathcal{J}(\mu^*) + \mathrm{N}_{\mathcal{C}}(\mu^*),$$

where $\partial \mathcal{J}(\mu^*)$ is the "Wasserstein subgradient" of $\mathcal{J}$ at μ^*, $\mathrm{N}_{\mathcal{C}}(\mu^*)$ is the "Wasserstein normal cone" of $\mathcal{C}$ at μ^* and $\mathbf{0}_{\mu^*}$ is a "null Wasserstein tangent vector" at μ^*.

As corollaries of our theorem, we obtain the "Wasserstein counterparts" of Fermat's rule in the unconstrained setting (i.e., the gradient vanishes at optimality) and the Lagrange conditions for (in)equality-constrained settings (i.e., the "gradients" of the objective and the constraint are "aligned" at "optimality", see Figure 1).

Before diving into variational analysis in the Wasserstein space, we illustrate our optimality conditions by studying a simple and accessible version of (1.1). For $\theta \neq 0$ and $\varepsilon > 0$, consider the problem

$$\inf_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathbb{E}^{x\sim\mu}\left[\langle \theta, x\rangle\right] \qquad \text{subject to} \qquad \mathbb{E}^{x\sim\mu}\left[\|x\|^2\right] \leq \varepsilon^2. \tag{1.2}$$

To get some intuition, let us restrict to Dirac's delta of the form δ_x for $x \in \mathbb{R}^d$. Accordingly, (1.2) reduces to $\inf_{\|x\|^2 \leq \varepsilon^2} \langle \theta, x\rangle$. This optimization problem can be studied through standard first-order optimality conditions in Euclidean spaces. Since the gradient of the objective $\nabla_x \langle \theta, x\rangle = \theta$ never vanishes, the optimal solution (if it exists) lies at the boundary. We thus seek the Lagrange multiplier $\lambda > 0$ such that

$$0 = \nabla_x \langle \theta, x\rangle + \lambda \nabla_x \|x\|^2 = \theta + 2\lambda x \quad \text{and} \quad \varepsilon^2 = \|x\|^2, \tag{1.3}$$

which yields $x = -\varepsilon \frac{\theta}{\|\theta\|}$ and $\lambda = \frac{\|\theta\|}{2\varepsilon}$. Now back to (1.2): After basic algebraic manipulations, our main result (stated informally on page 2) tells us that any solution $\mu^* \in \mathcal{P}_2(\mathbb{R}^d)$ of (1.2) satisfies the condition

$$0 = 2\lambda x + \theta \quad \mu^* \text{-a.e.} \quad \text{and} \quad \varepsilon^2 = \mathbb{E}^{x\sim\mu^*}\left[\|x\|^2\right] = \frac{\|\theta\|^2}{4\lambda^2}, \tag{1.4}$$

for some $\lambda \geq 0$ constant across the support of μ^*. We conclude that the mass of any candidate solution is necessarily located at $\varepsilon \frac{\theta}{\|\theta\|}$. In particular, our optimality conditions in (1.4) perfectly mirror their counterpart on $\mathbb{R}^d$ in (1.3).

1.1. Introduction to the broader context. Optimization problems over the probability space are ubiquitous across a variety of fields.

First, DRO emerges as a paradigm for decision-making under uncertainty [57]. In DRO, the goal is to identify solutions that are robust against a range of possible (adversary) probability measures, acknowledging the inherent ambiguity in real-world data, where the true underlying probability measure is uncertain and difficult to ascertain. Thus, DRO falls within the scope of (1.1) where $\mathcal{J}$ is a risk measure [5, 25, 38, 42] and $\mathcal{C}$ is a so-called ambiguity set of probability measures, often defined in terms of the Kullback-Leibler divergence [24, 51, 72] or an optimal transport discrepancy [8, 9, 26, 31, 43, 64, 74].

Second, in inverse problems one seeks the state of a system given some noisy observations. It is well-known [41] that Bayesian inference amounts to solving

$$\inf_{\mu \in \mathcal{P}(\mathbb{R}^d)} \mathbb{E}^{\theta\sim\mu}\left[\sum_{i=1}^N -\log(p(x_i|\theta))\right] + \mathrm{KL}(\mu \,||\, \hat{\mu}), \tag{1.5}$$

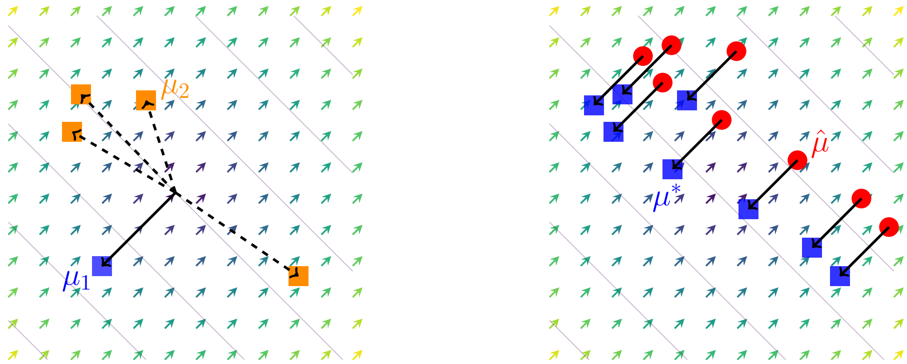


Fig. 1: "Gradients" are "aligned" with the constraints at "optimality". The figure on the left depicts two candidate solutions μ_1 and μ_2 for (1.1) with $\mathcal{C} = \{\mu \in \mathcal{P}_2(\mathbb{R}^2) \,|\, \mathbb{E}^{(x,y)\sim\mu}\left[x^2+y^2\right] \leq \varepsilon^2\}$ (i.e., bounded second moment) and $\mathcal{J}(\mu) = \mathbb{E}^{(x,y)\sim\mu}\left[x+y\right]$, of which we show the contours and the gradient vector field. The solid (dashed) black arrows represent the gradient of the constraint function $\mathbb{E}^{(x,y)\sim\mu}\left[x^2+y^2\right] - \varepsilon^2$ at μ_1 (μ_2). Here, μ_1 is indeed a candidate optimal solution: The gradient of the objective is aligned with the gradient of the constraint. For μ_2, instead, these two are not aligned. Thus, μ_2 cannot be optimal. The figure on the right shows that μ^* satisfies the optimality condition for (1.1) with $\mathcal{C} = \bar{\mathbb{B}}_{W_2}(\hat{\mu}; \varepsilon) = \{\mu \in \mathcal{P}_2(\mathbb{R}^2) \,|\, W_2(\mu, \hat{\mu}) \leq \varepsilon\}$ (i.e., Wasserstein ball centered at $\hat{\mu}$ of radius ε) and $\mathcal{J}(\mu) = \mathbb{E}^{(x,y)\sim\mu}\left[x+y\right])$, of which the contours and the gradient vector field are shown. The black arrows connecting $\hat{\mu}$ and μ^* represent the gradient of the constraint function $W_2(\mu, \hat{\mu})^2 - \varepsilon^2$. Since μ^* is optimal, the gradient of the objective and the constraint are aligned at all the "particles" of μ^*.

where x_i for $i \in \{1, \ldots, N\}$ are the observations, $p(x_i|\theta)$ is the probability of the observation x_i given the value of the state $\theta \sim \mu$, and KL is the Kullback-Leibler divergence between a candidate posterior μ and $\hat{\mu} \in \mathcal{P}(\mathbb{R}^d)$, the prior. Various inference problems (e.g., see [41, Table 1] and [23, 59]) result from modifications of (1.5).

Third, the dual formulation [56] of the Reinforcement Learning (RL) problem seeks the optimal stationary state-action distribution compatible with the dynamics $\mu^* \in \Delta(\mathcal{S} \times \mathcal{A}) \subseteq \mathcal{P}(\mathcal{S} \times \mathcal{A})$, where $\mathcal{S}$ is the state space and $\mathcal{A}$ the action space, that maximizes the expected reward $r : \mathcal{S} \times \mathcal{A} \to \bar{\mathbb{R}}$,

$$\mu^* \in \underset{\mu \in \Delta(\mathcal{S}\times\mathcal{A})}{\operatorname{argmax}}\ \mathbb{E}^{(s,a)\sim\mu}\left[r(s,a)\right], \tag{1.6}$$

a special case of (1.1). Variations of (1.6) yield different settings of the problem [33, 34, 36, 63, 67, 71, 77].

Finally, a growing number of fields are tackling optimal decision problems formulated in the probability space, including weather forecasting [28], single-cell perturbation responses [17], control of dynamical systems [35, 54, 65, 69], neural network training [22], mean-field control [10, 13], and finance [42, 48, 57], among others.

1.2. Related work. Our work studies (1.1) through the lens of optimal transport. The theory of optimal transport, dating back to the seminal work of Monge [49] and Kantorovich [39], defines a metric, the Wasserstein metric, on the space of probability measures. While the probability space is not a vector space, which makes most optimization tools in Banach space (e.g., [44, 47]) inapplicable, the theory of optimal transport enables a notion of differentiability, called Wasserstein differentiability, specifically tailored for the probability space [3, 37, 62]. Wasserstein differentiability was exploited in [46] to derive first-order optimality conditions for (1.1) for differentiable objective and constraint sets being sublevel sets of differentiable functionals. In the context of optimal control, [11, 12, 13] use Wasserstein differentiability to derive optimality conditions for optimal control problems, whereas [65] explores the properties of the dynamic programming algorithm in probability spaces for discrete-time. The theory has also been applied to derive algorithms to compute Wasserstein barycenters [1, 21, 53], to analyze over-parametrized neural networks [6, 22], approximate inference [29], and reinforcement learning [78, 67].

Unfortunately, many functionals over the probability space, including the Wasserstein distance itself, fail to be differentiable. This, together with the rigidity of so-far-considered feasible sets $\mathcal{C}$, effectively hinders the deployment of these tools in many practical instances. From a technical standpoint, these limitations are intrinsic in the choice of the perturbation of probability measures: Perturbations induced by the pushforward of (sufficiently regular) *transport maps*, as in [46], are not expressive enough. For instance, the pushforward of an empirical probability measure is always empirical (in particular, mass cannot be split). Thus, while attractive (e.g., perturbations can be captured by well-behaved functions, which form a vector space), a comprehensive theory of calculus requires a more general way of perturbing probability measures.

Intuitively, we can approximate a Dirac's delta by Gaussians with vanishing variance. However, there is no transport map describing a variation from the Dirac's delta to the approximating Gaussians. In this work, we therefore adopt a different approach and consider perturbations induced by *transport plans*. This more general way of perturbing probability measure prompts us to dive into the generalized notion of Wasserstein subdifferential proposed in [3, §10.3] and, importantly and contrary to the literature, in the generalized tangent space first introduced in [3, §12]. While these perturbations entail significant challenges (e.g., transport plans do not form a vector space), we show in this work that, if judiciously combined with traditional ideas from traditional variational analysis [60], they result in general necessary optimality conditions for (1.1).

Our approach offers several advantages over alternative methods for optimization in the probability space. First, our analysis is specifically tailored to the probability space and, in particular, does not require the introduction of non-negativity and normalization. For instance, if (1.1) is unconstrained, the corresponding optimality condition simply predicates that, at optimality, the Wasserstein gradient vanishes, just like in Euclidean settings. Second, our optimality conditions hold in full generality, and, in particular, do not rely on convexity-type assumptions or linearity assumptions, which in some cases might allow one to study (1.1) through infinite-dimensional linear programming or convex analysis. Third, analogously to the traditional Karush-Kuhn-Tucker conditions, we demonstrate that our optimality conditions can be used to both solve (1.1) in closed form, when the problem admits a closed-form solution, and devise numerical methods, when a closed-form solution is not available. We believe that this offers an advantage over alternative methods, e.g., [2], which instead are by design purely computational.

1.3. More details on our contributions. More specifically, our main contribution consists of four key aspects. First, we import various fundamental concepts from variational analysis, such as generalized subdifferential, normal cone, and tangent cone, to the Wasserstein space. To this extent, we need to resort to very general perturbations induced by transport plans. While invisible to the end user, these perturbations entail working with a tangent space which is not a linear space and carefully selecting an appropriate notion of convergence to resolve compactness issues. Second, we provide closed-form and easy-to-use expressions for the subdifferential of many functionals and for the normal cone of feasible sets of practical interests. In particular, we show that the Wasserstein distance is not regularly subdifferentiable and only admits a generalized subgradient. This result, which we believe to be of independent interest, also confirms that a general theory of optimality conditions is required already for simple functions (in particular, the distance itself), and not only to cover all corner cases. The key technique to characterize the general subgradient of the Wasserstein distance involves approximation arguments and its differentiability at regular measures, which is a promising technique to explore in future work addressing the differentiability of other functionals not covered in this work. Third, we derive general first-order optimality conditions for (1.1). Inspired by classical variational analysis in Euclidean spaces, we prove our result by reformulating (1.1) as an optimization problem over the epigraph of $\mathcal{J}$. This way, we can establish our optimality conditions under very weak assumptions on the functionals – not even continuity – and constraint qualification. Notably, as we demonstrate with several pedagogical examples, this complexity is hidden from the end user, and the deployment of our optimality conditions is effectively analogous to what one would do in Euclidean spaces. Fourth, we deploy our optimality to study a wide variety of optimization problems of the form (1.1) arising in machine learning and DRO. Across all these settings, we show that our optimality conditions both enable novel insights and, when the problem of interest does not admit a closed-form solution, can be used to design computational methods. We believe our tools enable the development of novel algorithms and results in machine learning, robust optimization, and biology, among others.

2. Subgradients and variational geometry in the Wasserstein space.

In this section, we introduce various elements for variational analysis in the Wasserstein space. After recalling preliminaries in measure theory and optimal transport in subsection 2.1, we present variations in the Wasserstein space in subsection 2.2. Armed with a way of perturbing probability measures, we then introduce Wasserstein subgradients in subsection 2.3 and study the variational geometry of the Wasserstein space in subsection 2.4, where, among others, we present normal and tangent cones.

2.1. Preliminaries.

All the maps considered in this work are tacitly assumed to be Borel, i.e., measurable w.r.t. the Borel topology. The set of Borel probability measures on $\mathbb{R}^d$ is $\mathcal{P}(\mathbb{R}^d)$, and we denote the set of finite second moment probability distributions by $\mathcal{P}_2(\mathbb{R}^d)$. We write $\mu \ll \bar{\mu}$ to indicate that μ is absolutely continuous w.r.t. $\bar{\mu}$, and by $\mathcal{P}_{2,\mathrm{abs}}(\mathbb{R}^d)$ the set of absolutely continuous probability measures with finite second moment. The support $\mathrm{supp}(\mu) \subseteq \mathbb{R}^d$ (cf. [3, Equation 5.0.1]) of a probability measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ is the closed set $\left\{x \in \mathbb{R}^d \mid \mu(U) > 0 \text{ for each neighborhood } U \text{ of } x \right\}$. The identity map on $\mathbb{R}^d$ is $\mathrm{Id}_{\mathbb{R}^d}(x) = x$ and when clear from the context, we simply write Id. The gradient of a function $h : \mathbb{R}^d \to \mathbb{R}$ at $x \in \mathbb{R}^d$ is $\nabla h(x)$, and the partial derivatives of a function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ at $(x, y) \in \mathbb{R}^d$ are denoted $\nabla_x c(x, y)$ and $\nabla_y c(x, y)$, respectively.

The *pushforward* of a probability measure $\mu \in \mathcal{P}(\mathbb{R}^n)$ through $T : \mathbb{R}^d \to \mathbb{R}^m$ (cf. [27, Definition 1.2.2]), denoted by $T_\# \mu \in \mathcal{P}(\mathbb{R}^m)$, is defined by $(T_\# \mu)(B) = \mu(T^{-1}(B))$ for all Borel sets $B \subseteq \mathbb{R}^m$, and it is a probability measure; see [27, Lemma 1.2.3]. Then, for any $T_{\#}\mu$-integrable $\phi : \mathbb{R}^m \to \mathbb{R}$, it holds $\int_{\mathbb{R}^m} \phi(y)\mathrm{d}(T_{\#}\mu)(y) = \int_{\mathbb{R}^n} \phi(T(x))\mathrm{d}\,\mu(x)$; see [27, Corollary 1.2.6]. Furthermore (cf. [27, Lemma 1.2.7]), for any $T : \mathbb{R}^n \to \mathbb{R}^m$ and $S : \mathbb{R}^m \to \mathbb{R}^t$ measurable, $(T \circ S)_{\#}\mu = S_{\#}(T_{\#}\mu)$.

Given $\mu_1 \in \mathcal{P}_2(\mathbb{R}^n)$ and $\mu_2 \in \mathcal{P}_2(\mathbb{R}^m)$, their product measure is $\mu_1 \times \mu_2$. We say that $T : \mathbb{R}^n \to \mathbb{R}^m$ is a *transport map* from μ_1 to μ_2 if $T_{\#}\mu = \nu$ or, equivalently, $\int_{\mathbb{R}^m} \phi(y)\mathrm{d}\,\mu_2(y) = \int_{\mathbb{R}^m} \phi(y)\mathrm{d}(T_{\#}\mu_1)(y)$ for all $\phi \in C_b^0(\mathbb{R}^m)$ [27, Lemma 1.2.5], where $C_b^0(\mathbb{R}^m)$ denotes the space of real-valued bounded continuous functions on $\mathbb{R}^m$. For some $i, m \in \mathbb{N}$, $1 \leq i \leq m$, we denote by $\pi_i : (\mathbb{R}^d)^m \to \mathbb{R}^d$ the projection map on the i^{th} component, i.e. $\pi_i(x_1, x_2, \ldots, x_m) = x_i$. A *transport plan* between μ_1 and μ_2 is a probability measure $\gamma \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ so that $\pi_{1\#}\gamma = \mu_1$ and $\pi_{2\#}\gamma = \mu_2$. A transport map may not exist between μ_1 and μ_2 (for instance, when $\mu_1 = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ and $\mu_2 = \frac{1}{M}\sum_{j=1}^M \delta_{y_j}$ with $M > N$), but a transport plan always does (e.g., the product measure $\gamma = \mu_1 \times \mu_2$). We collect them in the set (of *couplings*) $\Gamma(\mu_1, \mu_2)$. Given a lower semi-continuous function $c : \mathbb{R}^n \times \mathbb{R}^m \to \bar{\mathbb{R}} \coloneqq \mathbb{R} \cup \{\pm\infty\}$, the optimal transport problem reads:

$$W_c\left(\mu_1, \mu_2\right) \coloneqq \min_{\gamma \in \Gamma(\mu_1, \mu_2)} \int_{\mathbb{R}^n \times \mathbb{R}^m} c(x, y)\mathrm{d}\,\gamma(x, y). \tag{2.1}$$

The celebrated Wasserstein distance is a special case of (2.1) (cf. [27, Definition 3.1.3]):

$$W_2\left(\mu_1, \mu_2\right) \coloneqq \left(\min_{\gamma \in \Gamma(\mu_1, \mu_2)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \left\|x - y\right\|^2 \mathrm{d}\,\gamma(x, y)\right)^{\frac{1}{2}}, \tag{2.2}$$

where $\|\cdot\|$ is the standard euclidean norm in $\mathbb{R}^d$. We write $\Gamma_c(\mu_1, \mu_1)$ and $\Gamma_o(\mu_1, \mu_1)$ for the set of minimizers of (2.1) and (2.2), respectively. The Wasserstein distance is a distance on $\mathcal{P}_2(\mathbb{R}^d)$ [70, §6]. We define the Wasserstein ball of radius ε as $\mathbb{B}_{W_2}(\varepsilon; \bar{\mu}) \coloneqq \left\{\mu \in \mathcal{P}_2(\mathbb{R}^d) \,|\, W_2\left(\mu, \bar{\mu}\right) < \varepsilon\right\}$.

Throughout the work, we use two notions of convergence for probability measures. A sequence $(\mu_n)_{n \in \mathbb{N}} \subseteq \mathcal{P}(\mathbb{R}^d)$ (i) narrowly converges to $\mu \in \mathcal{P}(\mathbb{R}^d)$, denoted by $\mu_n \rightharpoonup \mu$, if for all $\phi \in C_b^0(\mathbb{R}^d)$ we have $\lim_{n\to\infty} \int_{\mathbb{R}^d} \phi(x)\mathrm{d}\,\mu_n(x) \to \int_{\mathbb{R}^d} \phi(x)\mathrm{d}\,\mu(x)$ (cf. [27, Definition 2.1.5]) and (ii) converges in the Wasserstein topology to $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$, $\mu_n \to \mu$, if $\lim_{n\to\infty} W_2\left(\mu, \bar{\mu}\right) = 0$. The narrow topology is weaker than the Wasserstein topology on $\mathcal{P}_2(\mathbb{R}^d)$. Indeed, by [3, Proposition 7.1.5], $\mu_n \to \bar{\mu}$ if and only if $\mu_n \rightharpoonup \bar{\mu}$ and $\int_{\mathbb{R}^d} \|x\|^2 \mathrm{d}\,\mu_n(x) \to \int_{\mathbb{R}^d} \|x\|^2 \mathrm{d}\,\bar{\mu}(x)$ or, equivalently, $\int_{\mathbb{R}^d} \phi(x)\mathrm{d}\,\mu_n(x) \to \int_{\mathbb{R}^d} \phi(x)\mathrm{d}\,\bar{\mu}(x)$ for all real-valued continuous ϕ with $|\phi(x)| \leq C(1 + \|x\|^2)$.

2.2. Variations in the Wasserstein space. In the Euclidean space $\mathbb{R}^d$, a variation at $x \in \mathbb{R}^d$ can be interpreted as an "arrow" $\upsilon \in \mathbb{R}^d$ rooted at x. In the same spirit, in the probability space, a variation at $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ is a "(weighted) collection of arrows" for each point in the support of $\bar{\mu}$ [3, Chapter 12]. Formally, a variation at $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ is a probability measure $\xi \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ whose first marginal is $\bar{\mu}$ (i.e., $\pi_{1\#}\xi = \bar{\mu}$) (see Figure 2, left). We can disintegrate [3, Theorem 5.3.1] ξ to obtain a collection $\{\xi_x\}_{x \in \mathbb{R}^d} \subseteq \mathcal{P}_2(\mathbb{R}^d)$ where each ξ_x denotes the probability measure over the "tangent" vectors (i.e., the "(weighted) collection of arrows") at each $x \in \operatorname{supp}\bar{\mu}$.

The distance between two variations $\xi_1, \xi_2 \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ at $\bar{\mu}$ can be defined in terms of the distance between their (weighted) arrows. To do so, we have to account for the fact that different arrows starting from the same point can be coupled in different ways. We express this coupling as a transport plan $\gamma \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d)$ so that $\pi_{12\#}\gamma = \xi_1$ and $\pi_{13\#}\gamma = \xi_2$, where $(x, \upsilon_1, \upsilon_2) \in \operatorname{supp}(\gamma)$ if and only if the arrows $v_1 \in \mathbb{R}^d$ and $v_2 \in \mathbb{R}^d$ are anchored at x, and v_1 and v_2 are coupled. Then, the distance between ξ_1 and ξ_2 amounts to the minimum distance that can be obtained among all such couplings:

$$W_{\bar{\mu}}(\xi_1, \xi_2) \coloneqq \left(\min_{\gamma \in \Gamma^1(\xi_1,\xi_2)} \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \|v_1 - v_2\|^2 \,\mathrm{d}\,\alpha(\bar{x}, v_1, v_2) \right)^{\frac{1}{2}}. \tag{2.3}$$

where $\Gamma^1(\xi_1, \xi_2)$ is the set of all couplings, as defined above, between ξ_1 and ξ_2:

$$\Gamma^1(\xi_1, \xi_2) \coloneqq \left\{ \gamma \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d) \mid \pi_{12\#}\gamma = \xi_1, \pi_{13\#}\gamma = \xi_2 \right\}.$$

In view of [3, Proposition 12.4.6], $W_{\bar{\mu}}(\xi_1, \xi_2) = 0 \Rightarrow W_2(\xi_1, \xi_2) = 0$. Similarly, we can also define the inner product and norm:

$$\langle \xi_1, \xi_2 \rangle_{\bar{\mu}} \coloneqq \max_{\alpha \in \Gamma^1(\xi_1,\xi_2)} \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \langle v_1, v_2 \rangle \,\mathrm{d}\,\alpha(\bar{x}, v_1, v_2), \tag{2.4}$$

$$\|\xi\|_{\bar{\mu}} \coloneqq \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|v\|^2 \,\mathrm{d}\,\xi_1(\bar{x}, v) \right)^{\frac{1}{2}}. \tag{2.5}$$

With these definitions, the Euclidean-like identity $W_{\bar{\mu}}(\xi_1, \xi_2)^2 = \|\xi_1\|_{\bar{\mu}}^2 - 2\langle \xi_1, \xi_2 \rangle_{\bar{\mu}} + \|\xi_1\|_{\bar{\mu}}^2$ holds.

In Euclidean spaces, given a variation v at a point x we can construct a new point by altering x according to v: $y = x + v$. Analogously, given a variation ξ at a probability measure $\bar{\mu}$, we obtain a new probability measure summing to all $x \in \operatorname{supp}(\bar{\mu})$ the variations v allocating mass according to the weight of $(x, v) \in \operatorname{supp}(\xi)$: $\mu = (\pi_1 + \pi_2)_\# \xi$. However, differently from the Euclidean counterpart where the variation displacing x to y is $v = y - x$, in the probability spaces we have multiple ways of connecting $\bar{\mu}$ and μ, described by the transport plans $\Gamma(\bar{\mu}, \mu)$. Each of these $\gamma \in \Gamma(\bar{\mu}, \mu)$ induces a variation $\xi = (\pi_1, \pi_2 - \pi_1)_\# \gamma$. Another difference is that while v describes a geodesic in the Euclidean space, i.e., $\|v\| = \|y - x\|$ and $\|\varepsilon\, v\| = \varepsilon \|y - x\|$ for $\varepsilon \in \mathbb{R}$ this is not the case for the generic variation ξ: $\|\xi\|_{\bar{\mu}} \geq W_2(\bar{\mu}, \mu)$ and $\left\|(\pi_1, \varepsilon(\pi_2 - \pi_1))_\# \gamma\right\|_{\bar{\mu}} \geq \varepsilon W_2(\bar{\mu}, \mu)$. In light of this, to define a meaningful *tangent space*, we consider only the variations that if "scaled enough" would describe geodesics in the Wasserstein space: The *tangent space* $\mathrm{T}_{\mathcal{P}_2(\mathbb{R}^d)}(\bar{\mu})$ of $\mathcal{P}_2(\mathbb{R}^d)$ at $\bar{\mu} \in \mathcal{P}(\mathbb{R}^d)$ is the closure with respect to $W_{\bar{\mu}}$ of

$$\begin{aligned} \Big\{ \xi \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d) \mid \exists \bar{\varepsilon} > 0, \forall \varepsilon \in (-\bar{\varepsilon}, \bar{\varepsilon}) \\ \pi_{1\#}\xi = \bar{\mu}, \\ (\pi_1, \pi_1 + \varepsilon\pi_2)_\# \xi \in \Gamma_o(\bar{\mu}, (\pi_1 + \varepsilon\pi_2)_\# \xi) \Big\}. \end{aligned} \tag{2.6}$$

Critically, the tangent space (2.6) is not a vector space. However, we can equip it with an "almost" linear structure (see Figure 2 for an intuition):

Definition 2.1 (An "almost linear" structure of the tangent space). *Given $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$, $\xi_1, \xi_2 \in \mathrm{T}_{\mathcal{P}_2(\mathbb{R}^d)}(\bar{\mu})$ and $\tau \in \mathbb{R}$, we define:*

(i) A "local" zero: $\mathbf{0}_{\bar{\mu}} \coloneqq \bar{\mu} \times \delta_0$.

(ii) A scalar multiplication: $\tau\,\xi_1 \coloneqq (\pi_1, \tau\pi_2)_\# \xi_1$.

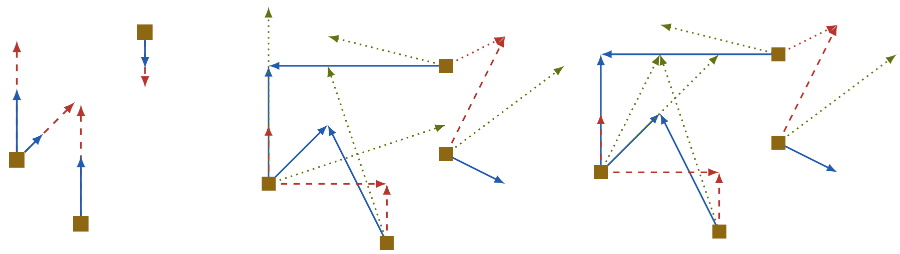


Fig. 2: Wasserstein geometry. On the left, we consider example variations for $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ (denoted by a brown square), ξ_1 in dotted red and ξ_2 in solid blue. They are scaled version of each other: $\xi_1 = 2\xi_2$. The other two images show two possible sums (in dotted green) of the variations ξ_1 (in dotted red) and ξ_2 (in solid blue) at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$: Each sum results from a different coupling $\alpha \in \Gamma^1(\xi_1, \xi_2)$ of ξ_1 and ξ_2 (i.e., different coupling of the arrows starting from the same square).

(iii) An addition: *For* $\alpha \in \Gamma^1(\xi_1, \xi_2)$, $\xi_1 +_\alpha \xi_2 \coloneqq (\pi_1, \pi_2 + \pi_3)_\# \alpha$.

In the appendix (Propositions SM1.4 and SM1.5), we show that Definition 2.1(ii)-(iii) are well-posed (i.e., $\tau\,\xi$ and $\xi_1 +_\alpha \xi_2$ belong to the tangent space) as well as additional properties of this "almost linear" structure of the tangent space.

Remark 2.2. The sum of n elements $\xi_1, \ldots, \xi_n$ results from iteratively applying Definition 2.1(iii) and we write $\xi_1 +_\alpha \ldots +_\alpha \xi_n = (\pi_1, \pi_2 + \ldots + \pi_{n+1})_\# \alpha$ for some $\alpha \in \Gamma^1(\xi_1, \ldots, \xi_n)$.

Comparison with the literature. Most of the literature uses the simplified tangent space

$$\begin{aligned} &\overline{\{\nabla\varphi \,|\, \varphi \in C_c^\infty(\mathbb{R}^d)\}}^{L^2(\mathbb{R}^d;\bar{\mu})} \\ &\qquad = \overline{\left\{\varepsilon(T - \mathrm{Id}) \,|\, (\mathrm{Id}, T)_\#\bar{\mu} \in \Gamma_o(\bar{\mu}, T_\#\bar{\mu}), \varepsilon > 0\right\}}^{L^2(\mathbb{R}^d;\bar{\mu})} \end{aligned} \tag{2.7}$$

first introduced in [3, Chapter 8]; see e.g. [46, 13, 11, 12] and [3, Theorem 8.5.1] for the proof of the equality in (2.7). The tangent space (2.7) is a subset of (2.6) since for $\nabla\varphi$ with $\varphi \in C_c^\infty(\mathbb{R}^d)$, the tangent vector $\xi = (\mathrm{Id}, \nabla\varphi)_\#\bar{\mu}$ belongs of $\mathrm{T}_{\mathcal{P}_2(\mathbb{R}^d)}(\bar{\mu})$ of $\mathcal{P}_2(\mathbb{R}^d)$. In fact, (i) its first marginal is $\bar{\mu}$ and (ii) $(\pi_1, \pi_1 + \varepsilon\pi_2)_\#\xi = (\mathrm{Id}, \mathrm{Id} + \varepsilon\nabla\varphi)_\#\bar{\mu}$ is, for ε sufficiently small, a transport plan. The latter follows from $\mathrm{Id} + \varepsilon\nabla\varphi$ being the gradient of a convex function and, thus, an optimal transport map [46, Proposition 2.3]. While attractive for its simplicity (it is a vector space), the tangent space (2.7) limits the perturbations to transport maps. For absolutely continuous measures, this comes with no loss of generality [15]. However, for probability measures with atoms (such as the ones in data-driven applications), the restriction to transport maps is effectively limiting the possible perturbations (e.g., a Dirac's delta can only be transported to another Dirac's delta with a transport map). Finally, the tangent space (2.6), with ε restricted to be positive, was first defined in [3, Chapter 12]. Instead, we drop the non-negativity of ε, which simplifies various technical results.

2.3. Wasserstein subgradients. We now define the *regular* and *general* subgradient, along the lines of [60, Definition 8.3] and [3, §10], for the Wasserstein space. The definition is inspired from the Euclidean setting, where $\bar{v}$ is a subgradient of

$f : \mathbb{R}^d \to \bar{\mathbb{R}}$ at $x \in \mathbb{R}^d$ if $f(x) - f(\bar{x}) \geq \langle \bar{\xi}, x - \bar{x} \rangle + o(\|x - \bar{x}\|)$ for all $x \in \mathbb{R}^d$. General subgradients are then defined via limits of regular subgradients.In the probability space, we can proceed analogously. In particular, we use the (local) inner product (2.4) and replace the displacement "$x - \bar{x}$" with $(\pi_1, \pi_2 - \pi_1)_{\#}\gamma$ where $\gamma \in \Gamma_o(\bar{\mu}, \mu)$ is an optimal transport plan between $\bar{\mu}$ and μ (which reduces to $(\mathrm{Id}, T - \mathrm{Id})_{\#}\bar{\mu}$ when γ is induced by a map T):

Definition 2.3 (Wasserstein subgradients). *Consider a functional $\mathcal{J} : \mathcal{P}_2(\mathbb{R}^d) \to \bar{\mathbb{R}}$ and a probability measure $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ with $\mathcal{J}(\bar{\mu}) \in \mathbb{R}$. For $\bar{\xi} \in \mathrm{T}_{\mathcal{P}_2(\mathbb{R}^d)}(\bar{\mu})$, we say that*

(i) $\bar{\xi}$ is a regular subgradient of $\mathcal{J}$ at $\bar{\mu}$, *written $\bar{\xi} \in \hat{\partial}\,\mathcal{J}(\bar{\mu})$, if for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and all $\xi \in (\pi_1, \pi_2 - \pi_1)_{\#}\Gamma_o(\bar{\mu}, \mu)$ $\mathcal{J}(\mu) - \mathcal{J}(\bar{\mu}) \geq \langle \bar{\xi}, \xi \rangle_{\bar{\mu}} + o(W_2(\bar{\mu}, \mu))$. In this case, $\mathcal{J}$ is regularly sub-differentiable at $\bar{\mu}$.*

(ii) $\bar{\xi}$ is a (general) subgradient of $\mathcal{J}$ at $\bar{\mu}$, *written $\bar{\xi} \in \partial\,\mathcal{J}(\bar{\mu})$ if there are sequences $\mu_n \to \bar{\mu}$ with $\mathcal{J}(\mu_n) \to \mathcal{J}(\mu)$ and $\xi_n \in \hat{\partial}\,\mathcal{J}(\mu_n)$ with $\xi_n \to \bar{\xi}$. In this case, $\mathcal{J}$ is sub-differentiable at $\bar{\mu}$.*

With this definition, we can then define regular supergradients as the elements of $-\hat{\partial}(-\mathcal{J})(\bar{\mu})$, and supergradients as elements of $-\partial(-\mathcal{J})(\bar{\mu}))$. We consider the Wasserstein convergence for the general subgradient since the first marginal may differ (we consider sequences $\mu_n \to \bar{\mu}$). For this reason, the general subgradient may not be in the tangent space. Similarly, we can then define differentiability:

Definition 2.4 (Differentiable functional). *A functional $\mathcal{J} : \mathcal{P}_2(\mathbb{R}^d) \to \bar{\mathbb{R}}$ is* differentiable *at $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ if it admits both a regular subgradient and supergradient at $\bar{\mu}$; i.e., $-\hat{\partial}(-\mathcal{J})(\bar{\mu}) \cap \hat{\partial}\,\mathcal{J}(\bar{\mu}) \neq \emptyset$.*

Wasserstein subgradients enjoy similar properties to their Euclidean counterpart:

Proposition 2.5 (Characterization of the (sub-)gradients). *Consider a functional $\mathcal{J} : \mathcal{P}_2(\mathbb{R}^d) \to \bar{\mathbb{R}}$ and a probability measure $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$, where $\mathcal{J}(\bar{\mu}) \in \mathbb{R}$. Then,*

(i) Either $\mathcal{J}$ is differentiable or at least one between $\hat{\partial}\,\mathcal{J}(\bar{\mu})$ and $\hat{\partial}(-\mathcal{J})(\bar{\mu})$ is empty.

(ii) If $G : \mathcal{P}_2(\mathbb{R}^d) \to \bar{\mathbb{R}}$ is a lower bound on $\mathcal{J}$, i.e., $G(\mu) \leq \mathcal{J}(\mu)$ for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $G(\bar{\mu}) = \mathcal{J}(\bar{\mu})$, then $\hat{\partial} G(\bar{\mu}) \subseteq \hat{\partial}\,\mathcal{J}(\bar{\mu})$.

If $\mathcal{J}$ is differentiable at $\bar{\mu}$, the following statements hold:

(iii) For all $\xi_1, \xi_2 \in -\hat{\partial}(-\mathcal{J})(\bar{\mu}) \cap \hat{\partial}\,\mathcal{J}(\bar{\mu})$, $\xi_1 = \xi_2$. Thus, we call the gradient of $\mathcal{J}$ at $\bar{\mu}$, written $\nabla \mathcal{J}(\bar{\mu})$, this unique element. In particular, it satisfies $\hat{\partial}\,\mathcal{J}(\bar{\mu}) = \{\nabla \mathcal{J}(\bar{\mu})\} \subseteq \partial\,\mathcal{J}(\bar{\mu})$.

(iv) A tangent element $\xi \in \mathrm{T}_{\mathcal{P}_2(\mathbb{R}^d)}(\bar{\mu})$ is the gradient of $\mathcal{J}$ at $\bar{\mu}$ if and only if for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $\xi' \in (\pi_1, \pi_2 - \pi_1)_{\#}\xi'$, $\alpha \in \Gamma^1(\xi, \xi')$ it holds:

$$\mathcal{J}(\mu) - \mathcal{J}(\bar{\mu}) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle x_2, x_3 - x_1 \rangle \,\mathrm{d}\,\alpha(x_1, x_2, x_3) + o(W_2(\bar{\mu}, \mu)). \tag{2.8}$$

In particular, Proposition 2.5(iii) allows us to talk about *the* gradient. The (general) subgradient might not be a singleton, even for a differentiable functional; see [60, §8.B] for an example in $\mathbb{R}^d$. Proposition 2.5(iv), instead, is the analogous of $f(y) - f(x) = \nabla f(x)^\top (y - x) + o(\|x - y\|)$: Namely, Wasserstein gradients provide a "linear approximation" of $\mathcal{J}$, with errors that are small in terms of the Wasserstein distance.

Comparison with the literature. Our definition of regular subgradient ((i) in Definition 2.4) coincides with the definition of extended Fréchet subdifferential, introduced in [3, Definition 10.3.1] and employed, for instance, in [13]. It is a weaker notion of subdifferentiability compared to the ones in [30, 46], whereby subgradients are of the form $(\mathrm{Id}, \phi)_{\#}\mu$ for some Borel map $\phi : \mathbb{R}^d \to \mathbb{R}^d$. From an optimization perspective, these definitions are however not satisfying since, as we will discuss shortly, the Wasserstein distance itself fails to be subdifferentiable at measures that are not absolutely continuous. Inspired by classical variational analysis in Euclidean spaces [60], we therefore introduce the definition of general and horizon subgradients ((ii) in Definition 2.3). To the best of our knowledge, this definition is novel in the Wasserstein space.

2.3.1. Why do we need such machinery?. We argue that our machinery is required for at least two reasons. First, already in Euclidean spaces, variational analysis is needed to deal with non-differentiable and non-convex settings; the (general) subgradient usually appears in the necessary conditions, while the horizon subgradient (which we shall discuss in subsection 2.4.2) appears in the constraint qualification (e.g., see [60, Theorem 8.15]). We do not expect things to simplify in the more general probability space. Second, the Wasserstein distance itself fails to be regularly subdifferentiable. The existing theory of gradient flows [3, 46] bypasses this complexity resorting to absolutely continuous measures. However, this introduces a substantial practical and theoretical limitation since it excludes any data-driven setting. We characterize below the general subgradient which is, instead, non-empty.

The Wasserstein distance is not regularly subdifferentiable. To start, we study the regular differentiability properties of the Wasserstein distance:

PROPOSITION 2.6 (Wasserstein subgradients and optimal plans). *Consider the functional $\mathcal{J} : \mathcal{P}_2(\mathbb{R}^d) \to \bar{\mathbb{R}}$, $\mathcal{J}(\mu) := \frac{1}{2}W_2(\mu, \hat{\mu})^2$. Then, for any $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$:*

(i) $\mathcal{J}$ is regularly super-differentiable at $\bar{\mu}$.

(ii) If $\Gamma_o(\bar{\mu}, \hat{\mu})$ is a singleton and induced by an optimal transport map (e.g., $\bar{\mu}$ is absolutely continuous), then $\mathcal{J}$ is regularly sub-differentiable, and thus differentiable, at $\bar{\mu}$.

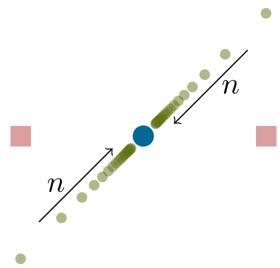
(iii) If $\xi \in \hat{\partial}\,\mathcal{J}(\bar{\mu})$, then $(\pi_1, \pi_1 - \pi_2)_{\#}\xi \in \Gamma_o(\bar{\mu}, \hat{\mu})$.

(iv) If $\Gamma_o(\bar{\mu}, \hat{\mu})$ is not a singleton, then $\mathcal{J}$ is not regularly sub-differentiable at μ.

Proposition 2.6 indicates, perhaps surprisingly, that the squared Wasserstein distance from a reference measure $\hat{\mu}$ is not subdifferentiable at all measures μ for which there exist multiple optimal transport plans between μ and $\hat{\mu}$. Instead, it is regularly differentiable, and thus differentiable, whenever there is a unique optimal transport plan between μ and $\hat{\mu}$ and this plan is induced by an optimal transport map, a result first established in [3, Theorem 10.2.6]. By Brenier's theorem, if $\bar{\mu}$ is absolutely continuous, then there is a unique optimal transport plan induced by an optimal transport map between $\bar{\mu}$ and $\hat{\mu}$, and so the squared Wasserstein distance is differentiable at absolutely continuous measures. This result is sharp: Even when a single optimal transport plan exists, the Wasserstein distance may lack regular subdifferentiability, as we illustrate in the next example.

EXAMPLE 2.7 (Non regular-subdifferentiability of the squared Wasserstein distance). *Let $\bar{\mu} := \delta_{(0,0)}$ (blue in the picture) and $\hat{\mu} := \frac{1}{2}\delta_{(-1,0)} + \frac{1}{2}\delta_{(1,0)}$ (red in the picture). Clearly, $\mathcal{J}(\bar{\mu}) = \frac{1}{2}W_2(\bar{\mu}, \hat{\mu})^2 = \frac{1}{2}$ and there is a unique optimal transport plan between $\bar{\mu}$ and $\hat{\mu}$; in particular, $\Gamma_o(\bar{\mu}, \hat{\mu}) = \{\bar{\mu} \times \hat{\mu}\}$. We now show that $\mathcal{J}(\mu) = \frac{1}{2}W_2(\mu, \hat{\mu})^2$ is not regularly subdifferentiable at $\bar{\mu}$. By Proposition 2.6(iii), it suffices to prove that $\xi = (\pi_1, \pi_1 - \pi_2)_{\#}(\bar{\mu} \times \hat{\mu})$ is not a regular subgradient of $\mathcal{J}$.*

Consider $\mu_n = \frac{1}{2}\delta_{(-\frac{1}{n},-\frac{1}{n})} + \frac{1}{2}\delta_{(\frac{1}{n},\frac{1}{n})}$. *Clearly, there is a unique optimal transport plan* $\Gamma_o(\bar{\mu},\mu_n) = \{\bar{\mu}\times\mu_n\}$, *and* $\mathcal{J}(\mu_n) = \frac{1}{2}W_2(\mu_n,\hat{\mu})^2 = \frac{1}{2}((1-\frac{1}{n})^2+\frac{1}{n^2}) = \frac{1}{2}-\frac{1}{n}+\frac{1}{n^2}$, *and* $W_2(\bar{\mu},\mu_n) = (\frac{1}{n^2}+\frac{1}{n^2})^{1/2} = \frac{\sqrt{2}}{n}$. *Following Definition* 2.3, *we consider the variations* $\xi_n \in (\pi_1,\pi_1-\pi_2)_{\#}\Gamma_o(\bar{\mu},\mu_n)$. *Using the expressions of* $\bar{\mu}$ *and* μ_n, *we conclude that the only variation between* $\bar{\mu}$ *and* μ_n *is* $\xi_n = \bar{\mu}\times\mu_n$. *With this, we can now compute the inner product between* ξ *and* ξ_n*:*



$$
\begin{aligned}
\langle \xi,\xi_n\rangle_{\bar{\mu}} &= \max_{\alpha\in\Gamma^1(\xi,\xi_n)} \int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \langle v, v_n\rangle \,\mathrm{d}\,\alpha(x,v,v_n) \\
&= \max_{\tau\in[0,1]} \frac{\tau}{2}\left(\left\langle (1,0),\left(\frac{1}{n},\frac{1}{n}\right)\right\rangle + \left\langle (-1,0),\left(-\frac{1}{n},-\frac{1}{n}\right)\right\rangle\right) \\
&\qquad + \frac{(1-\tau)}{2}\left(\left\langle (1,0),\left(-\frac{1}{n},-\frac{1}{n}\right)\right\rangle + \left\langle (-1,0),\left(\frac{1}{n},\frac{1}{n}\right)\right\rangle\right) = \frac{1}{n}.
\end{aligned}
$$

Thus,

$$
\liminf_{n\to\infty} \frac{\mathcal{J}(\mu_n)-\mathcal{J}(\bar{\mu})-\langle\xi,\xi_n\rangle_{\bar{\mu}}}{W_2(\bar{\mu},\mu_n)} = \liminf_{n\to\infty}\frac{-\frac{2}{n}+\frac{1}{n^2}}{\frac{\sqrt{2}}{n}} = -\sqrt{2} < 0
$$

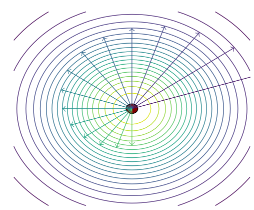
and so $\mathcal{J}(\mu_n)-\mathcal{J}(\bar{\mu}) < \langle\xi,\xi_n\rangle_{\bar{\mu}} + o(W_2(\bar{\mu},\mu_n))$ *That is,* $\mathcal{J}$ *is not regularly subdifferentiable at* $\bar{\mu}$ *and, in particular, uniqueness of the optimal plan does not imply regular subdifferentiability.*

The absence of regular subgradients prompts us to study the general subgradient of the Wasserstein distance. We will do so in the next section.

The (general) subgradient of the Wasserstein distance. The regular subdifferentiability of the Wasserstein distance at absolutely continuous probability measures (cf. Proposition 2.6(ii)) fuels the hope that the general subgradient can be characterized by approximating each probability measure via absolutely continuous ones. We illustrate the intuition by approximating a Dirac's delta at 0 with Gaussian probability measures with vanishing variance:

Example 2.8 (Subgradients via Gaussians). *Let* $\mathcal{J}:\mathcal{P}_2(\mathbb{R}^d)\to\bar{\mathbb{R}}$, $\mathcal{J}(\mu) = \frac{1}{2}W_2(\mu,\hat{\mu})^2$ *and fix* $\bar{\mu}=\delta_0$ *and* $\hat{\mu}=\mathcal{N}(0,I_d)$, *with* I_d *being the identity matrix in* $\mathbb{R}^{d\times d}$. *It is easy to verify that there is only a unique plan* $\bar{\mu}\times\hat{\mu}$ *from* $\bar{\mu}$ *to* $\hat{\mu}$, *but it is not induced by a transport map (indeed, the mass at 0 needs to be split, see picture below for an intuition). Here, we show that* $\xi = (\pi_1,\pi_1-\pi_2)_{\#}(\bar{\mu}\times\hat{\mu})$ *belongs to the general subgradient* $\partial\mathcal{J}(\bar{\mu})$. *To do so, we approximate* δ_0 *with Gaussians with vanishing variance. Specifically, consider the sequence* $\mu_n = \mathcal{N}\left(0,\frac{1}{n^2}I_d\right)$ *so that* $W_2(\bar{\mu},\mu_n) = \frac{1}{n}\searrow 0$ *and* $\mu_n\to\bar{\mu}$ *and* $\mathcal{J}(\mu_n)\to\mathcal{J}(\bar{\mu})$. *Since both* $\bar{\mu}$ *and* μ_n *are Gaussians, the optimal transport plans is a singleton* $\Gamma_o(\mu_n,\hat{\mu}) = \{(\mathrm{Id},T_n)_{\#}\mu_n\}$, *where the optimal transport plan is induced by the optimal transport map* $T_n(x)=nx$. *By Proposition* 2.6, $\mathcal{J}$ *is differentiable at* μ *and, in particular,* $\xi_n := (\mathrm{Id},\mathrm{Id}-T_n)_{\#}\mu_n \in \hat{\partial}\mathcal{J}(\mu_n)$. *We now show that* $\xi_n\to\xi$ *and so* ξ *is a (general) subgradient of* $\mathcal{J}$. *Convergence (in the Wasserstein topology) of the first marginal follows directly from* $(\pi_1)_{\#}\xi_n=\mu_n$, $(\pi_1)_{\#}\xi=\bar{\mu}$, *and* $\mu_n\to\bar{\mu}$. *For narrow convergence of the second marginal, consider* $\phi\in C_b^0(\mathbb{R}^d)$. *Then,*

$$
\begin{aligned}
\lim_{n\to\infty}\int_{\mathbb{R}^d}\phi(y)\mathrm{d}\,\xi_n(x,y) &= \lim_{n\to\infty}\int_{\mathbb{R}^d}\phi(x-T_n(x))\mathrm{d}\,\mu_n(x)\\
&= \lim_{n\to\infty}\int_{\mathbb{R}^d}\phi(x-T_n(x))\mathrm{d}(T_n^{-1})_{\#}\hat{\mu}(x)\\
&= \int_{\mathbb{R}^d}\lim_{n\to\infty}\phi\left(\frac{1}{n}x-x\right)\mathrm{d}\,\hat{\mu}(x)\\
&= \int_{\mathbb{R}^d}\phi(0-x)\mathrm{d}\,\hat{\mu}(x)\\
&= \int_{\mathbb{R}^d}\phi(y)\mathrm{d}\,\xi(x,y),
\end{aligned}
$$



where we used dominated convergence to exchange limit and integral and continuity of ϕ. Thus, $(\pi_2)_{\#}\xi_n \rightharpoonup (\pi_2)_{\#}\xi$, and so $\xi \in \partial\mathcal{J}(\bar{\mu})$ and the Wasserstein distance is subdifferentiable at $\bar{\mu}$.

Since $\mathcal{P}_{2,\mathrm{abs}}(\mathbb{R}^d)$ is dense in $\mathcal{P}_2(\mathbb{R}^d)$ [53, Theorem 2.2.7], we can always construct a sequence $(\mu_n)_{n\in\mathbb{N}} \subseteq \mathcal{P}_{2,\mathrm{abs}}(\mathbb{R}^d)$ with $\mu_n \to \bar{\mu}$, compute $\xi_n \in \hat{\partial}\mathcal{J}(\mu_n)$, which exists by absolute continuity of μ_n, and study the limit of ξ_n. One way to construct such μ_n is via convolution with a Gaussian kernel; see [3, Lemma 7.1.10]. Overall, this procedure yields a sufficient characterization (for the sake of necessary optimality conditions) of the (general) subgradient of the Wasserstein distance:

PROPOSITION 2.9 (General subgradient of the squared Wasserstein distance). *For $\bar{\mu}, \hat{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$, $\mathcal{J} : \mathbb{R}^d \to \bar{\mathbb{R}}$, $\mathcal{J}(\mu) = \frac{1}{2}W_2(\mu,\hat{\mu})^2$, we have $\partial\mathcal{J}(\bar{\mu}) \neq \emptyset$ and $\partial\mathcal{J}(\bar{\mu}) \subseteq (\pi_1, \pi_1 - \pi_2)_{\#}\Gamma_o(\bar{\mu},\hat{\mu})$.*

In particular, Proposition 2.9 establishes subdifferentiability of the Wasserstein distance at *all* probability measures.

2.3.2. Examples of subgradients. Next, we characterize the subdifferential of several functionals of interest. We start with general optimal transport discrepancies, defined in (2.1):

PROPOSITION 2.10 (Optimal transport discrepancy). *Let $\hat{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ and let $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ be twice continuously differentiable with bounded Hessian. Suppose that c satisfies the twist condition (i.e., $y \mapsto \nabla_x c(x,y)$ is injective for all $x \in \mathbb{R}^d$). Consider the functional $\mathcal{J} : \mathcal{P}_2(\mathbb{R}^d) \to \bar{\mathbb{R}}$, $\mathcal{J}(\mu) \coloneqq W_c(\mu,\hat{\mu})$ as defined in* (2.1)*. Then,*

(i) $\mathcal{J}$ is proper, non-negative, and lower semi-continuous.

(ii) At any $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$, $\mathcal{J}$ is subdifferentiable with $\emptyset \neq \partial\mathcal{J}(\bar{\mu}) \subseteq (\pi_1, \nabla_x c)_{\#}\Gamma_c(\bar{\mu},\hat{\mu})$, where $\Gamma_c(\bar{\mu},\hat{\mu})$ is the set of transport plans between $\bar{\mu}$ and $\hat{\mu}$ that yields the minimum for $W_c(\mu,\hat{\mu})$.
Moreover, if $\Gamma_c(\bar{\mu},\hat{\mu})$ is a singleton and induced by an optimal transport map (e.g., $\bar{\mu}$ is absolutely continuous), then $\mathcal{J}$ is differentiable with $\nabla\mathcal{J}(\bar{\mu}) = (\pi_1, \nabla_x c)_{\#}\xi$, where ξ is the unique element in $\Gamma_c(\bar{\mu},\hat{\mu})$.

As observed for the Wasserstein distance, optimal transport discrepancies are subdifferentiable (but not regularly subdifferentiable), and their subgradient is characterized in terms of the gradient of the transportation cost and the set of optimal transport plans. Remarkably, when $\bar{\mu} = \delta_{\bar{x}}$ and $\hat{\mu} = \delta_{\hat{x}}$, then $\mathcal{J}(\bar{\mu}) = W_c(\bar{\mu},\hat{\mu}) = c(x,\hat{x})$ and $\Gamma_c(\bar{\mu},\hat{\mu})) = \{\delta_{(\bar{x},\hat{x})}\}$. Thus, $\mathcal{J}$ is Wasserstein differentiable with gradient $\nabla\mathcal{J}(\bar{\mu}) = (\pi_1, \nabla_x c)_{\#}\delta_{(\bar{x},\hat{x})} = \delta_{\bar{x}} \times \delta_{\nabla_x c(\bar{x},\hat{x})}$, which is formally analogous to the standard Euclidean gradient of $x \mapsto c(x,\hat{x})$. We now exemplify our results in the case of strictly convex transportation costs:

Example 2.11 (Strictly convex transportation cost). *Consider* $c(x,y) = h(x-y)$ *for* $h : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ *strictly convex and twice continuously differentiable with bounded Hessian. By strict convexity of* h, $y \mapsto \nabla_x c(x,y) = \nabla h(x-y)$ *is injective. Thus,* c *satisfies the assumptions of Proposition* 2.10 *and the subgradient of* $\mathcal{J}(\mu) := W_c(\mu, \hat{\mu})$ *at* $\bar{\mu}$ *is* $\emptyset \neq \partial \mathcal{J}(\bar{\mu}) \subseteq (\pi_1, \nabla h \circ (\pi_1 - \pi_2))_{\#} \Gamma_c(\bar{\mu}, \hat{\mu})$. *Moreover,* $\mathcal{J}$ *is differentiable whenever* $\Gamma_c(\bar{\mu}, \hat{\mu})$ *is a singleton and induced by an optimal transport map and, in particular, at all absolutely continuous measures.*

With $c(x,y) = \frac{1}{2}\|x-y\|^2$ (i.e., $h(z) = \frac{1}{2}\|z\|^2$ with $\nabla h = \mathrm{Id}$), we can study the differentiability properties of the squared Wasserstein distance and, among others, recover the subdifferentiability result of Proposition 2.9:

Corollary 2.12 (Squared Wasserstein distance). *Fix* $\hat{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ *and consider the functional* $\mathcal{J} : \mathcal{P}_2(\mathbb{R}^d) \to \bar{\mathbb{R}}$, $\mathcal{J}(\mu) := \frac{1}{2} W_2(\mu, \hat{\mu})^2$. *Then, the subgradient of* $\mathcal{J}$ *at* $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ *is* $\emptyset \neq \partial \mathcal{J}(\bar{\mu}) \subseteq (\pi_1, \pi_1 - \pi_2)_{\#} \Gamma_o(\bar{\mu}, \hat{\mu})$. *Moreover, if* $\Gamma_o(\bar{\mu}, \hat{\mu})$ *is a singleton and induced by an optimal transport map (e.g.,* $\bar{\mu}$ *is absolutely continuous), then* $\mathcal{J}$ *is differentiable* $\bar{\mu}$ *and* $\nabla \mathcal{J}(\bar{\mu}) = (\pi_1, \pi_1 - \pi_2)_{\#} \xi$, *where* ξ *is the unique optimal transport plan in* $\Gamma_o(\bar{\mu}, \hat{\mu})$.

In particular, when the set of optimal transport plans consists of a unique optimal transport map $T : \mathbb{R}^d \to \mathbb{R}^d$, then $\nabla \mathcal{J}(\bar{\mu}) = (\mathrm{Id}, \mathrm{Id} - T)_{\#} \bar{\mu}$, which is formally analogous the gradient of $x \mapsto \frac{1}{2}\|x - \hat{x}\|^2$ being $x - \hat{x}$ in Euclidean spaces (with "Id as x" and "T as $\hat{x}$").

Next, we recall and extend, by characterizing their general subgradients, the existing subdifferentiability results for expected values:

Proposition 2.13 (Expected value). *Let* $V : \mathbb{R}^d \to (-\infty, +\infty]$ *be a proper,* λ*-convex, lower semicontinuous functional whose negative part has 2-growth, i.e.* $V(x) \geq -A - B\|x\|^2$ *for all* $x \in \mathbb{R}^d$ *for some* $A, B \in \mathbb{R}$ *and consider the functional* $\mathcal{J} : \mathcal{P}_2(\mathbb{R}^d) \to \bar{\mathbb{R}}$, $\mathcal{J}(\mu) := \mathbb{E}^{x \sim \mu}[V(x)]$. *Then,*

(i) $\mathcal{J}$ *is proper and lower semi-continuous.*

(ii) At every $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ *at which* $\mathcal{J}(\bar{\mu}) \in \mathbb{R}$, $\xi \in \mathrm{T}_{\mathcal{P}_2(\mathbb{R}^d)}(\bar{\mu})$ *is in* $\hat{\partial} \mathcal{J}(\bar{\mu})$, $\partial \mathcal{J}(\bar{\mu})$ *if and only if for all* $(x,y) \in \mathrm{supp}(\xi)$, $y \in \hat{\partial} V(x), \partial V(x)$ *, respectively. Since* $\hat{\partial} V(x) = \partial V(x)$ *(cf. Proposition* SM1.1*),* $\hat{\partial} \mathcal{J}(\bar{\mu}) = \partial \mathcal{J}(\bar{\mu})$. *In particular, if* V *is differentiable and* $V(x) \leq A + B\|x\|^2$, $\mathcal{J}$ *is differentiable at any* $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ *and* $\nabla \mathcal{J}(\bar{\mu}) = (\mathrm{Id}, \nabla V)_{\#} \bar{\mu}$.

Any twice continuously differentiable function V with bounded Hessian is λ-convex and satisfies $|V(x)| \leq A + B\|x\|^2$ for some $A, B \in \mathbb{R}$. Thus, the corresponding functional $\mathcal{J}$ is differentiable and, in particular, $(x,y) \in \mathrm{supp}(\nabla \mathcal{J})$ if and only if $y = \nabla V(x)$. If, instead, V does not have negative quadratic growth, then $\mathcal{J}$ is not proper and might evaluate to $-\infty$ (e.g., $V(x) = -e^x$ and $\mu = \mathcal{N}(0,1)$). Similarly, if V grows more than quadratically, then $\mathcal{J}$ will attain $+\infty$ (e.g., $V(x) = e^x$ and $\mu = \mathcal{N}(0,1)$) and so cannot be differentiable.

Example 2.14 (Expected values of linear functions). *Consider the same functional* $\mathcal{J}$ *as in section 1, namely* $\mathcal{J}(\mu) = \mathbb{E}^{x \sim \mu}[\langle \theta, x \rangle]$ *and* $V(x) = \langle \theta, x \rangle$ *in Proposition* 2.13. *Since* V *is a linear function, it satisfies all the assumptions of Proposition* 2.13, *and in particular* $\mathcal{J}$ *is Wasserstein differentiable with* $\nabla \mathcal{J}(\mu) = (\mathrm{Id}, \theta)_{\#} \mu$.

Finally, we extend the existing subdifferentiability results for interaction-type functionals which model the energy associated with the interaction between particles in a distribution; e.g., see [61] for an introduction, [18] for an example in physics, and

[19, §3.3] for one in economics. For simplicity, we restrict here our attention to the smooth case.

PROPOSITION 2.15 (Interaction energy). *Let $U : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable with bounded Hessian and consider the functional $\mathcal{J} : \mathcal{P}_2(\mathbb{R}^d) \to \bar{\mathbb{R}}$, $\mathcal{J}(\mu) := \frac{1}{2}\mathbb{E}^{(x,y)\sim\mu\times\mu}\left[U(x-y)\right]$. Then,*
(i) $\mathcal{J}$ is proper and continuous.
(ii) $\mathcal{J}$ is Wasserstein differentiable and its Wasserstein gradient at $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ is $\nabla\mathcal{J}(\bar{\mu}) = (\mathrm{Id}, (\nabla U * \bar{\mu}))_{\#}\bar{\mu}$, where $*$ denotes the convolution operation.

We exemplify the result in the case of quadratic functions:

EXAMPLE 2.16 (Quadratic interaction energies). *Consider $U(z) = \frac{1}{2}\langle Qz, z\rangle$ for $Q \in \mathbb{R}^{d\times d}$. Then, U is twice continuosly differentiable with $\nabla U(z) = Qz$ and constant (and, thus, bounded) Hessian Q. In virtue of Proposition 2.15, the functional $\mathcal{J}(\mu) = \frac{1}{2}\mathbb{E}^{(x,y)\sim\mu\times\mu}\left[U(x-y)\right]$ is Wasserstein differentiable with $\nabla\mathcal{J}(\mu) = (\mathrm{Id}, Q\mathbb{E}^{x\sim\mu}\left[x\right])_{\#}\mu$.*

2.3.3. Subdifferential calculus. We now show some useful calculus rules in the Wasserstein space. More can be studied and we expect this to be a source of interesting future work.

PROPOSITION 2.17 (Subdifferential calculus). *Consider the proper functionals $\mathcal{J}_1, \mathcal{J}_2 : \mathcal{P}_2(\mathbb{R}^d) \to \bar{\mathbb{R}}$, a monotone map $g : \bar{\mathbb{R}} \to \bar{\mathbb{R}}$ with $g(\infty) = \infty$, and $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ so that $\mathcal{J}_i(\bar{\mu}) \in \mathbb{R}$, g is continuously differentiable at $\mathcal{J}_1(\bar{\mu}) \in \mathbb{R}$ with $g(\mathcal{J}_1(\bar{\mu})) \in \mathbb{R}$, and $\mathcal{J}_1$ continuous in the Wasserstein topology at $\bar{\mu}$. Then, the following rules hold:*
(i) Sum rule:

$$\begin{aligned}\hat{\partial}(\mathcal{J}_1 + \mathcal{J}_2)(\bar{\mu}) &\supseteq \hat{\partial}(\mathcal{J}_1)(\bar{\mu}) + \hat{\partial}(\mathcal{J}_2)(\bar{\mu}) \\ &:= \left\{\xi_1 +_\alpha \xi_2 \mid \xi_i \in \hat{\partial}\,\mathcal{J}_i(\bar{\mu}), \alpha \in \Gamma^1(\xi_1, \xi_2)\right\}.\end{aligned}$$

In particular, if $\mathcal{J}_2$ is differentiable at $\bar{\mu}$, then

$$\hat{\partial}(\mathcal{J}_1 + \mathcal{J}_2)(\bar{\mu}) = \hat{\partial}(\mathcal{J}_1)(\bar{\mu}) + \{\nabla\mathcal{J}_2(\bar{\mu})\}$$

and, thus,

$$\partial(\mathcal{J}_1 + \mathcal{J}_2)(\bar{\mu}) = \partial\,\mathcal{J}_1(\bar{\mu}) + \{\nabla\mathcal{J}_2(\bar{\mu})\}.$$

(ii) Chain rule: If $g'(\mathcal{J}(\bar{\mu})) > 0$, then

$$\begin{aligned}\hat{\partial}(g \circ \mathcal{J}_1)(\bar{\mu}) &= g'(\mathcal{J}_1(\bar{\mu}))\hat{\partial}\,\mathcal{J}_1(\bar{\mu}) \\ &:= \{g'(\mathcal{J}_1(\bar{\mu}))\,\xi \mid \xi \in \hat{\partial}\,\mathcal{J}_1(\bar{\mu})\} \quad \text{and} \\ \partial(g \circ \mathcal{J}_1)(\bar{\mu}) &= g'(\mathcal{J}_1(\bar{\mu}))\partial\,\mathcal{J}_1(\bar{\mu}).\end{aligned}$$

With $g_i(x) = \rho_i x$ for $\rho_1, \rho_2 \in \mathbb{R}_{\geq 0}$, Proposition 2.17 readily characterizes the subgradients of the linear combination of functionals:

COROLLARY 2.18 (Sum and multiplication rule). *Consider proper functionals $\mathcal{J}_1, \mathcal{J}_2 : \mathcal{P}_2(\mathbb{R}^d) \to \bar{\mathbb{R}}$, $\rho_1, \rho_2 \in \mathbb{R}_{\geq 0}$, and $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$. Then,*

$$\hat{\partial}(\rho_1\,\mathcal{J}_1 + \rho_2\,\mathcal{J}_2)(\bar{\mu}) \supseteq \rho_1\hat{\partial}\,\mathcal{J}_1(\bar{\mu}) + \rho_2\hat{\partial}\,\mathcal{J}_2(\bar{\mu}).$$

If additionally $\mathcal{J}_2$ is differentiable,

$$\begin{aligned}\hat{\partial}(\rho_1\,\mathcal{J}_1 + \rho_2\,\mathcal{J}_2)(\bar{\mu}) &= \rho_1\hat{\partial}\,\mathcal{J}_1(\bar{\mu}) + \rho_2\nabla\mathcal{J}_2(\bar{\mu}) \quad \text{and} \\ \partial(\rho_1\,\mathcal{J}_1 + \rho_2\,\mathcal{J}_2)(\bar{\mu}) &= \rho_1\partial\,\mathcal{J}_1(\bar{\mu}) + \rho_2\nabla\mathcal{J}_2(\bar{\mu}).\end{aligned}$$

Next, we deploy Proposition 2.17 to study the variance:

COROLLARY 2.19 (Variance). *Consider the functional* $\mathcal{J} : \mathcal{P}_2(\mathbb{R}^d) \to \bar{\mathbb{R}}$, $\mathcal{J}(\mu) \coloneqq \mathrm{Var}^{x\sim\mu}[V(x)]$. *If both* V *and* $x \mapsto V(x)^2$ *satisfy the assumptions of Proposition* 2.13*, at every* $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ *at which* $\mathcal{J}(\bar{\mu}) \in \mathbb{R}$*,*

$$\partial \mathcal{J}(\bar{\mu}) = \hat{\partial} \mathcal{J}(\bar{\mu}) = (\pi_1, 2(\pi_1 - \mathbb{E}^{x\sim\bar{\mu}}[V(x)])\pi_2)_{\#}\left\{\xi \in \mathrm{T}_{\mathcal{P}_2(\mathbb{R}^d)}(\bar{\mu}) \mid \mathrm{supp}\,(\xi) \subseteq \hat{\partial} V\right\}$$

When V *is continuously differentiable and* $V(x)^2, V(x)^4 \leq A + B\|x\|^2$ *for some* $A, B > 0$*,* $\mathcal{J}$ *is differentiable and its gradient is* $\nabla\mathcal{J}(\bar{\mu}) = (\mathrm{Id}, 2(V - \mathbb{E}^{x\sim\mu}[V(x)])\nabla V)_{\#}\bar{\mu}$*.*

2.4. Variational geometry. So far, we focused on the subdifferentability of the objective function. In this section, we study the geometry of the constraint set $\mathcal{C}$. We combine the two in section 3 to derive our necessary optimality conditions. Similarly to Euclidean settings, all the necessary information is encoded in tangent and normal cones (see Figure 3 for an intuition in $\mathbb{R}^d$). We start with the tangent cone:

DEFINITION 2.20 (Tangent cone). *The* tangent cone *of* $\mathcal{C} \subseteq \mathcal{P}_2(\mathbb{R}^d)$ *at* $\bar{\mu} \in \mathcal{C}$*, denoted by* $\mathrm{T}_{\mathcal{C}}(\bar{\mu})$*, is characterized by all the* $\bar{\xi} \in \mathrm{T}_{\mathcal{P}_2(\mathbb{R}^d)}(\bar{\mu})$ *such that for some* $(\mu_n)_{n\in\mathbb{N}} \subseteq \mathcal{C}$*,* $\mu_n \to \bar{\mu}$*,* $(\xi_n)_{n\in\mathbb{N}} \subseteq \mathrm{T}_{\mathcal{P}_2(\mathbb{R}^d)}(\bar{\mu})$*,* $\xi_n \in (\pi_1, \pi_2 - \pi_1)_{\#}\Gamma_o(\bar{\mu}, \mu_n)$ *and* $\tau_n \searrow 0$*, one has* $W_{\bar{\mu}}\left(\frac{1}{\tau_n}\xi_n, \bar{\xi}\right) \to 0$*.*

We highlight the formal similarity with the tangent cone in Euclidean spaces, defined by all the $\bar{v}$ for which there is $(x_n)_{n\in\mathbb{N}} \subset \mathcal{C}$, $x_n \to \bar{x}$, and $\tau_n \searrow 0$ so that $\frac{1}{\tau_n}(x_n - \bar{x}) \to \bar{v}$. In particular, if $\mu_n = \delta_{x_n}$ and $\bar{\mu} = \delta_{\bar{x}}$, then $\Gamma_o(\mu_n, \bar{\mu}) = \{\delta_{(x_n,\bar{x})}\}$ and $\xi_n = (\bar{x}, x_n - \bar{x})$. Namely, the definitions in the probability space and in the Euclidean space agree. Intuitively , the tangent cone is a restriction of the tangent space, and it describes the variations admissible within the set $\mathcal{C}$. Dual to tangent cones are the normal cones:

DEFINITION 2.21 (Normal cone). *The* regular normal cone *of* $\mathcal{C} \subseteq \mathcal{P}_2(\mathbb{R}^d)$ *at* $\bar{\mu} \in \mathcal{C}$*, denoted by* $\hat{\mathrm{N}}_{\mathcal{C}}(\bar{\mu})$*, is characterized by all the* $\bar{\xi} \in \mathrm{T}_{\mathcal{P}_2(\mathbb{R}^d)}(\bar{\mu})$ *such that for all* $\mu \in \mathcal{C}$*, one has* $\langle \bar{\xi}, \xi \rangle_{\bar{\mu}} \leq o\left(\|\xi\|_{\bar{\mu}}\right)$ *for* $\xi = (\pi_1, \pi_2 - \pi_1)_{\#}\Gamma_o(\bar{\mu}, \nu)$ *and* $\nu \in \mathcal{C}$*. The* normal cone *of* $\mathcal{C}$ *at* $\bar{\mu}$*, denoted by* $\mathrm{N}_{\mathcal{C}}(\bar{\mu})$*, is then defined as the limit point of the regular normals; i.e.,* $\bar{\xi} \in \mathrm{T}_{\mathcal{P}_2(\mathbb{R}^d)}(\bar{\mu})$ *belongs to* $\mathrm{N}_{\mathcal{C}}(\bar{\mu})$ *if there exists* $(\mu_n)_{n\in\mathbb{N}} \subset \mathcal{C}$*,* $\mu_n \to \bar{\mu}$*,* $\xi_n \in \hat{\mathrm{N}}_{\mathcal{C}}(\mu_n)$ *so that* $\xi_n \to \bar{\xi}$*.*

Similarly to the subgradient definition Definition 2.3, for general normals we consider the Wasserstein convergence of regular normals. This implies that the general normals may not be in the tangent space, and need not be. Analogously to the tangent cone, if we specify the definition of normal cone to the Dirac's deltas, we see that it generalizes the notion of normal cone in Euclidean spaces; see also Proposition 2.23. The next proposition characterizes the dual relationship between normal and tangent cones:

PROPOSITION 2.22 (Normal and tangent cones relationships). *At any* $\bar{\mu} \in \mathcal{C} \subseteq \mathcal{P}_2(\mathbb{R}^d)$*, the sets* $\hat{\mathrm{N}}_{\mathcal{C}}(\bar{\mu})$ *and* $\mathrm{N}_{\mathcal{C}}(\bar{\mu})$ *are closed cones. Additionally, if* $\bar{\xi} \in \hat{\mathrm{N}}_{\mathcal{C}}(\bar{\mu})$ *then* $\langle \bar{\xi}, \xi \rangle_{\bar{\mu}} \leq 0$ *for all* $\xi \in \mathrm{T}_{\mathcal{C}}(\bar{\mu})$*.*

2.4.1. Examples of normal cones. We start with a few simple examples of normal cones:

PROPOSITION 2.23 (Trivial normal cones). *Let* $\mathcal{C} \subseteq \mathcal{P}_2(\mathbb{R}^d)$ *and* $\bar{\mu} \in \mathcal{C}$*. Then,*

Fig. 3: Variational Geometry in $\mathbb{R}^d$. On the left, the arrows depict some examples of regular normals. At the same location $x \in \mathcal{C}$ they may have different scales and directions. There also may not be any regular normal. On the right, the solid, blue, arrows depict some examples of normals. They are the limit points of sequences of regular normals. The sequence is colored with increasing opacity as the regular normals approach the general normal. In dotted red there are sequences $\tau_n(x_n - x) \to w \in \mathrm{T}_{\mathbb{R}^2}(x)$, with the tangent vector w being depicted in solid red. The inner product between tangent vectors and regular normals is non-positive, whereas the same might not hold with general normals.

(i) if $\mathcal{C} = \mathcal{P}_2(\mathbb{R}^d)$, then $\hat{\mathrm{N}}_{\mathcal{C}}(\bar{\mu}) = \mathrm{N}_{\mathcal{C}}(\bar{\mu}) = \mathrm{N}_{\mathcal{P}_2(\mathbb{R}^d)}(\bar{\mu}) = \{\mathbf{0}_{\bar{\mu}}\}$;

(ii) if $\bar{\mu} \in \mathrm{int}_{W_2}\mathcal{C}$, then $\hat{\mathrm{N}}_{\mathcal{C}}(\bar{\mu}) = \mathrm{N}_{\mathcal{C}}(\bar{\mu}) = \{\mathbf{0}_{\bar{\mu}}\}$;

(iii) for $\Xi \subseteq \mathbb{R}^d$, if $\mathcal{C} \subseteq \{\delta_x \in \mathcal{P}_2(\mathbb{R}^d) \,|\, x \in \Xi\}$, then at any $\bar{\mu} = \delta_{\bar{x}} \in \mathcal{C}$

$$\hat{\mathrm{N}}_{\mathcal{C}}(\bar{\mu}) = \{\delta_{\bar{x}} \times \nu \,|\, \mathbb{E}^{x\sim\nu}[x] \in \hat{\mathrm{N}}_{\Xi}(\bar{x})\} \quad \textit{and}$$
$$\mathrm{N}_{\mathcal{C}}(\bar{\mu}) = \{\delta_{\bar{x}} \times \nu \,|\, \mathbb{E}^{x\sim\nu}[x] \in \mathrm{N}_{\Xi}(\bar{x})\}.$$

In particular, analogously to the Euclidean setting, the normal cone at $\bar{\mu}$ trivializes if the feasible set covers the whole space (Proposition 2.23(i)) or if $\bar{\mu}$ lies in its interior (Proposition 2.23(ii)). Moreover, when we restrict ourselves to the space of Dirac's deltas, the normal cone is consistent with its Euclidean counterpart (Proposition 2.23(iii)); namely, a "tangent vector" ν lies in the normal cone if and only if its "average" tangent vector lies in the Euclidean normal cone.

The remainder of the section characterizes the normal cones of two important classes of constraints: support constraints and level sets of functionals. We start with support constraints:

Proposition 2.24 (Normal cone for support constraint). *For a closed convex $\Xi \subseteq \mathbb{R}^d$ define*

$$\mathcal{C}_{\Xi} \coloneqq \left\{\mu \in \mathcal{P}_2(\mathbb{R}^d) \,|\, \mathrm{supp}\,(\mu) \subseteq \Xi\right\}. \tag{2.9}$$

Then,

(i) The set $\mathcal{C}_{\Xi}$ has empty interior: $\partial\mathcal{C}_{\Xi} = \mathcal{C}_{\Xi}$.

(ii) At any $\bar{\mu} \in \partial\mathcal{C}_{\Xi}$,

$$\hat{\mathrm{N}}_{\mathcal{C}_{\Xi}}(\mu) = \left\{\xi \in \mathrm{T}_{\mathcal{P}_2(\mathbb{R}^d)}(\mu) \,|\, (x,y) \in \mathrm{supp}\,(\xi) \Leftrightarrow y \in \hat{\mathrm{N}}_{\Xi}(x)\right\}$$
$$\mathrm{N}_{\mathcal{C}_{\Xi}}(\mu) = \left\{\xi \in \mathrm{T}_{\mathcal{P}_2(\mathbb{R}^d)}(\mu) \,|\, (x,y) \in \mathrm{supp}\,(\xi) \Leftrightarrow y \in \mathrm{N}_{\Xi}(x)\right\}.$$

(iii) At any $\bar{\mu} \in \partial \mathcal{C}_\Xi$, $\xi \in \mathrm{N}_{\mathcal{C}_\Xi}(\bar{\mu})$ if and only if a Borel family $\{\xi_x\}_{x\in\Xi}, \xi_x \in \mathcal{P}_2(\mathrm{N}_\Xi(x))$ exists such that $\xi = \int_{\mathbb{R}^d} \xi_x \,\mathrm{d}\,\bar{\mu}(x)$.

Intuitively, a tangent vector ξ belongs to the normal cone $\mathrm{N}_{\mathcal{C}_\Xi}(\bar{\mu})$ if and only if each of its "particles" $(x, y) \in \operatorname{supp}(\xi)$ "belongs" to the normal cone of Ξ (i.e., $y \in \mathrm{N}_\Xi(x)$). We instantiate Proposition 2.24 to the case of box constraints.

EXAMPLE 2.25 (Normal cone for box support constraint). *Given the intervals $\Xi_i = [l_i, u_i] \subset \mathbb{R}$ (with $l_i < u_i$), define $\Xi \coloneqq \Xi_1 \times \ldots \times \Xi_d$ and consider $\mathcal{C}_\Xi$ as in (2.9). For every $x = (x_1, \ldots, x_d) \in \Xi$, the normal cone of Ξ reads (cf. [60, Example 6.10])*

$$\hat{\mathrm{N}}_\Xi(x) = \mathrm{N}_\Xi(x) = \mathrm{N}_{\Xi_1}(x_1) \times \ldots \times \mathrm{N}_{\Xi_d}(x_d) \qquad \mathrm{N}_{\Xi_i}(x_i) = \begin{cases} [0, \infty) & \text{if } x_i = l_i, \\ (-\infty, 0] & \text{if } x_i = u_i, \\ \{0\} & \text{else.} \end{cases}$$

Then, at any $\bar{\mu} \in \mathcal{C}_\Xi$,

$$\hat{\mathrm{N}}_{\mathcal{C}_\Xi}(\bar{\mu}) = \mathrm{N}_{\mathcal{C}_\Xi}(\bar{\mu}) = \left\{ \xi \in \mathrm{T}_{\mathcal{P}_2(\mathbb{R}^d)}(\mu) \,|\, (x, y) \in \operatorname{supp}(\xi) \Leftrightarrow y_i \in \hat{\mathrm{N}}_{\Xi_i}(x_i) \right\}.$$

Second, we study the normal cones of level sets of continuous functionals:

PROPOSITION 2.26 (Level sets). *Suppose $\mathcal{C} \coloneqq \left\{\mu \in \mathcal{P}_2(\mathbb{R}^d) \,|\, \mathcal{J}(\mu) \leq 0\right\}$ for a proper, continuous functional $\mathcal{J} : \mathcal{P}_2(\mathbb{R}^d) \to \bar{\mathbb{R}}$. Then, at any $\bar{\mu} \in \mathcal{C}$ so that $\mathcal{J}(\bar{\mu}) = 0$, $\hat{\mathrm{N}}_{\mathcal{C}}(\bar{\mu}) \supseteq \mathbb{R}_{\geq 0} \hat{\partial} \mathcal{J}(\bar{\mu})$ and, if $\mathbf{0}_{\bar{\mu}} \notin \partial \mathcal{J}(\bar{\mu})$ and $\partial \mathcal{J}(\bar{\mu}) \neq \emptyset$, $\mathrm{N}_{\mathcal{C}}(\bar{\mu}) \subseteq \mathbb{R}_{\geq 0} \partial \mathcal{J}(\bar{\mu})$. At any $\bar{\mu} \in \mathcal{C}$ so that $\mathcal{J}(\bar{\mu}) < 0$ it holds that $\hat{\mathrm{N}}_{\mathcal{C}}(\bar{\mu}) = \mathrm{N}_{\mathcal{C}}(\bar{\mu}) = \{\mathbf{0}_{\bar{\mu}}\}$.*

Proposition 2.26 intimately relates to Lagrange multipliers in the Wasserstein space: The normal cone of a sublevel set of a functional $\mathcal{J}$ consists of any scaling (by a multiplier) of the subgradients of $\mathcal{J}$. In particular, it includes the case where the constraint is not active (i.e., the multiplier is zero and the normal cone trivializes) and the case where the constraint is active (i.e., the multiplier is non-zero). The condition $\mathbf{0}_{\bar{\mu}} \notin \partial \mathcal{J}(\bar{\mu})$ resembles the same "constraint qualification" required in Euclidean settings (cf. [60, Proposition 10.3]), and in practice it is often not restrictive. As an example, we consider the normal cone of (closed) Wasserstein balls:

EXAMPLE 2.27 (Normal cone of closed Wasserstein balls). *For some $\varepsilon > 0$, consider $\mathcal{C} = \bar{\mathbb{B}}_{W_2}(\varepsilon; \hat{\mu}) = \{\mu \in \mathcal{P}_2(\mathbb{R}^d) \,|\, W_2(\mu, \hat{\mu})^2 \leq \varepsilon^2\}$. Since at any interior point the normal cone of $\mathcal{C}$ trivializes, we only need to consider $\bar{\mu}$ on the boundary. By Proposition 2.9, the subgradient of the Wasserstein distance is not empty and is contained in $(\pi_1, 2\pi_1 - 2\pi_2)_{\#}\Gamma_o(\bar{\mu}, \hat{\mu})$[1]. Since $\bar{\mu}$ lies at the boundary, we have $\bar{\mu} \neq \hat{\mu}$ and so $\Gamma_o(\bar{\mu}, \hat{\mu})$ does not contain the "identity plan" $(\mathrm{Id}, \mathrm{Id})_{\#}\bar{\mu}$. Thus, $\mathbf{0}_{\bar{\mu}}$ does not belong to the subgradient of the Wasserstein distance, and the constraint qualification in Proposition 2.26 holds true. With this, we conclude that the normal cone satisfies*

$$\mathrm{N}_{\mathcal{C}}(\bar{\mu}) \subseteq \mathbb{R}_{\geq 0}(\pi_1, 2\pi_1 - 2\pi_2)_{\#}\Gamma_o(\bar{\mu}, \hat{\mu}) = \{\xi \mid \xi \in (\pi_1, 2\lambda(\pi_1 - \pi_2))_{\#}\Gamma_o(\bar{\mu}, \hat{\mu}), \lambda \geq 0\}.$$

In particular, if $\bar{\mu}$ is absolutely continuous, then $\Gamma_o(\bar{\mu}, \hat{\mu}) = \{(\mathrm{Id}, T)_{\#}\bar{\mu}\}$, where T is the unique optimal transport map T, and $\mathrm{N}_{\mathcal{C}}(\bar{\mu}) \subseteq \{(\mathrm{Id}, 2\lambda(\mathrm{Id} - T))_{\#}\bar{\mu} \,|\, \lambda \geq 0\}$.

The next example characterizes the feasible set of the optimization problem presented in section 1, namely the second moment constraint:

[1] Here, the factor 2 arises since there is no $\frac{1}{2}$ in front of the Wasserstein distance; cf. Corollary 2.18.

EXAMPLE 2.28 (Second moment constraint). *For some* $\varepsilon > 0$*, consider* $\mathcal{C} = \{\mu \in \mathcal{P}_2(\mathbb{R}^d) \,|\, \mathbb{E}^{x\sim\mu}\left[\|x\|^2\right] \le \varepsilon^2\}$*. Then,* $\mathcal{C} = \bar{\mathbb{B}}_{W_2}(\varepsilon;\delta_0)$ *and, thus, for any* $\bar\mu \in \mathcal{P}_2(\mathbb{R}^d)$ *in the interior of* $\mathcal{C}$ *the normal cone trivializes, whereas for all the ones with* $\mathbb{E}^{x\sim\bar\mu}\left[\|x\|^2\right] = \varepsilon^2$ *we have* $\mathrm{N}_{\mathcal{C}}(\bar\mu) \subseteq \mathbb{R}_{\ge 0}(\mathrm{Id}, 2\mathrm{Id})_{\#}\bar\mu$*.*

2.4.2. Variational geometry and epigraphs. In section 3, we provide the first-order optimality conditions for general, constrained, optimization problems and present differentiable functionals as a special case. From a technical standpoint, however, we *first* establish optimality conditions for differentiable functions, and *then* study the non-differentiable case via an epigraphical argument. More formally, recall that the epigraph of a proper functional $\mathcal{J} : \mathcal{P}_2(\mathbb{R}^d) \to \bar{\mathbb{R}}$ is the set

$$\widetilde{\mathrm{epi}}\,\mathcal{J} \coloneqq \left\{(\mu,\beta) \mid \mu \in \mathcal{P}_2(\mathbb{R}^d), \beta \ge \mathcal{J}(\mu)\right\} \subset \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}.$$

Then, we immediately observe that

$$\inf_{\mu\in\mathcal{C}} \mathcal{J}(\mu) = \inf_{(\mu,\beta)\in\widetilde{\mathrm{epi}}(\mathcal{J})\cap(\mathcal{C}\times\mathbb{R})} \beta = \inf_{(\mu,\beta)\in\widetilde{\mathrm{epi}}(\mathcal{J})\cap(\mathcal{C}\times\mathbb{R})} \tilde{l}(\mu,\beta),$$

with $\tilde{l} : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R} \to \mathbb{R}$ being the linear function $\tilde{l}(\mu,\beta) = \beta$. Being linear, $\tilde{l}$ is supposedly easy to optimize. Unfortunately, this argument does not directly carry over to the Wasserstein space: While on $\mathbb{R}^d$ the epigraph is a subset of $\mathbb{R}^{d+1}$, on $\mathcal{P}_2(\mathbb{R}^d)$ is a subset of $\mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}$ and not $\mathcal{P}_2(\mathbb{R}^{d+1})$. Thus, one has to study the differential structure and variational geometry of $\mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}$. To avoid this burden, we take a different angle. Since each $x \in \mathbb{R}^d$ can be "embedded" in the probability space as the Dirac's delta $\delta_x \in \mathcal{P}_2(\mathbb{R}^d)$, we define the epigraph of a functional $\mathcal{J}$ as

$$\mathrm{epi}\,\mathcal{J} \coloneqq \left\{\mu \times \delta_\beta \mid \mu \in \mathcal{P}_2(\mathbb{R}^d), \beta \ge \mathcal{J}(\mu)\right\} \subseteq \mathcal{P}_2(\mathbb{R}^{d+1}). \tag{2.10}$$

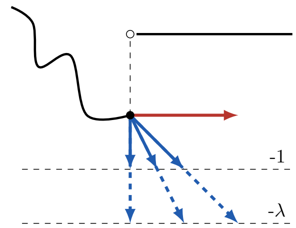
In particular, contrary to $\widetilde{\mathrm{epi}}\,\mathcal{J}$, $\mathrm{epi}\,\mathcal{J}$ is now a subset of $\mathcal{P}_2(\mathbb{R}^{d+1})$. Then, the optimization of $\mathcal{J}$ over $\mathcal{C}$ is equivalent to the optimization of $l : \mathcal{P}_2(\mathbb{R}^{d+1}) \to \mathbb{R}$, $l(\boldsymbol{\mu}) = \mathbb{E}^{\boldsymbol{x}\sim\boldsymbol{\mu}}\left[\langle(0,\dots,0,1),\boldsymbol{x}\rangle\right]$, over $\mathrm{epi}\,\mathcal{J} \cap \mathcal{C}_e$, with $\mathcal{C}_e \coloneqq \{\mu \times \delta_\beta \mid \mu \in \mathcal{C}, \beta \in \mathbb{R}\}$. Note that we use bold symbols when working in $\mathbb{R}^{d+1}$ and $\mathcal{P}_2(\mathbb{R}^{d+1})$ (e.g., $\boldsymbol{x} = (x,\beta)$ with $x \in \mathbb{R}^d$ and $\beta \in \mathbb{R}$). Given $\boldsymbol{\mu}$, we can "extract" μ and δ_β via $(\pi_1,\dots,\pi_d)_{\#}\boldsymbol{\mu} = \mu$ (i.e., we forget the last component) and $(\pi_{d+1})_{\#}\boldsymbol{\mu} = \delta_\beta$ (i.e., we only keep the last component). Similarly, given a variation $\boldsymbol{\xi} \in \mathrm{T}_{\mathcal{P}_2(\mathbb{R}^{d+1})}(\mu)$, we can extract the variation corresponding to μ via $T_{\#}\boldsymbol{\xi}$, where

$$T \coloneqq (\pi_1,\dots,\pi_d,\pi_{d+2},\dots,\pi_{2d+1})$$

(i.e., we forget the last component of the first and the second marginal), and the variation "corresponding" to δ_β via $(\pi_{d+1},\pi_{d+2})_{\#}\boldsymbol{\xi}$ (i.e., we only keep the last component of the first and the second marginal).

With this formalism, the gradient of l follows directly from Proposition 2.13 and it reads $\boldsymbol{\mu} \times \boldsymbol{\delta}_{(\mathbf{0},\dots,\mathbf{0},\mathbf{1})}$. Thus, the complexity of our main result moves to prove first-order optimality conditions in the differentiable case subject to epigraphical constraints, for which, in particular, we need to study the variational geometry of $\mathrm{epi}\,\mathcal{J}$, which we do next. To start, observe that epigraphical normals "point downwards in expectation":

PROPOSITION 2.29 (Characterization of the epigraphical normals). *Let* $\mathcal{J} : \mathcal{P}_2(\mathbb{R}^d) \to \bar{\mathbb{R}}$ *and consider any* $\bar\mu \in \mathcal{P}_2(\mathbb{R}^d)$ *at which* $\mathcal{J}(\bar\mu) \in \mathbb{R}$*. Then, for all* $\boldsymbol{\xi} \in \mathrm{N}_{\mathrm{epi}\,\mathcal{J}}(\bar\mu \times \delta_{\mathcal{J}(\bar\mu)})$ *we have* $\mathbb{E}^{(\boldsymbol{x},\boldsymbol{v})\sim\boldsymbol{\xi}}\left[\langle(0,\dots,0,1),\boldsymbol{v}\rangle\right] \le 0$*.*



Proposition 2.29 is reminiscent of the well-known fact that, in Euclidean settings, epigraphical normals always point downwards (see figure on the right). Analogously, in the probability space, epigraphical normals point downward in expected value. Next, we show that the non-horizontal normals to epi $\mathcal{J}$ are intimately related to the subgradients of $\mathcal{J}$:

PROPOSITION 2.30 (Subgradients and epigraphical normals). *Let $\mathcal{J} : \mathcal{P}_2(\mathbb{R}^d) \to \bar{\mathbb{R}}$ and any $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ at which $\mathcal{J}(\bar{\mu}) \in \mathbb{R}$ and let $\boldsymbol{\xi} \in \mathrm{N}_{\mathrm{epi}\,\mathcal{J}}(\bar{\mu} \times \delta_{\mathcal{J}(\bar{\mu})})$ with $\mathbb{E}^{(\boldsymbol{x},\boldsymbol{v})\sim\boldsymbol{\xi}}\left[\langle (0,\ldots,0,1), \boldsymbol{v}\rangle\right] = -1$. Then, $T_{\#}\boldsymbol{\xi} \in \partial\,\mathcal{J}(\bar{\mu})$.*

Note that Proposition 2.30 characterizes *all* non-horizontal normals. Indeed, if $\boldsymbol{\xi} \in \mathrm{N}_{\mathrm{epi}\,\mathcal{J}}(\bar{\mu} \times \delta_{\mathcal{J}(\bar{\mu})})$ satisfies $\mathbb{E}^{(\boldsymbol{x},\boldsymbol{v})\sim\boldsymbol{\xi}}\left[\langle (0,\ldots,0,1), \boldsymbol{v}\rangle\right] = -\tau < 0$, then, since the normal cone is a cone (cf. Proposition 2.22), $\frac{1}{\tau}\boldsymbol{\xi}$ also belongs to $\mathrm{N}_{\mathrm{epi}\,\mathcal{J}}(\bar{\mu} \times \delta_{\mathcal{J}(\bar{\mu})})$ and it satifies $\mathbb{E}^{(\boldsymbol{x},\boldsymbol{v})\sim\frac{1}{\tau}\boldsymbol{\xi}}\left[\langle (0,\ldots,0,1), \boldsymbol{v}\rangle\right] = \mathbb{E}^{(\boldsymbol{x},\boldsymbol{v})\sim\boldsymbol{\xi}}\left[\left\langle (0,\ldots,0,1), \frac{1}{\tau}\boldsymbol{v}\right\rangle\right] = -1$. Thus, $(\pi_1,\ldots,\pi_d)_{\#}\frac{1}{\tau}\boldsymbol{\xi} \in \partial\,\mathcal{J}(\bar{\mu})$. Thus, we only need to characterize the horizontal normals. We do so by introducing a third notion of subgradients:

DEFINITION 2.31 (Wasserstein horizon subgradients). *For $\mathcal{J} : \mathcal{P}_2(\mathbb{R}^d) \to \bar{\mathbb{R}}$ and any $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ at which $\mathcal{J}(\bar{\mu}) \in \mathbb{R}$, we say that ξ is a* horizon subgradient of $\mathcal{J}$ at $\bar{\mu}$, *$\xi \in \partial^\infty\,\mathcal{J}(\bar{\mu})$, if there exists $\boldsymbol{\xi} \in \mathrm{N}_{\mathrm{epi}\,\mathcal{J}}(\bar{\mu} \times \delta_{\mathcal{J}(\bar{\mu})})$ such that $T_{\#}\boldsymbol{\xi} = \xi$ and $\mathbb{E}^{(\boldsymbol{x},\boldsymbol{v})}\left[\langle (0,\ldots,0,1), \boldsymbol{v}\rangle\right] = 0$.*

The horizon subgradients relate to the constraint qualifications and, as we shall see in section 3, are required for a comprehensive theory of optimization. Since they are defined starting from the general normals to the epigraph, they are not guaranteed to be in the tangent space. Nonetheless, for sufficiently well-behaved functionals they are trivial:

PROPOSITION 2.32 (Horizon subgradients for strictly continuous functions). *Let $\mathcal{J} : \mathcal{P}_2(\mathbb{R}^d) \to \bar{\mathbb{R}}$ and any $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$. If $\mathcal{J}$ is* strictly continuous *at $\bar{\mu}$, i.e.,*

$$\lim_{\mu\to\bar{\mu}} \frac{\|\mathcal{J}(\mu) - \mathcal{J}(\bar{\mu})\|}{W_2(\mu,\bar{\mu})} < \infty,$$

then the horizon subgradient of $\mathcal{J}$ at $\bar{\mu}$ is trivial; i.e., $\partial^\infty\,\mathcal{J}\,\bar{\mu} = \{\mathbf{0}_{\bar{\mu}}\}$.

For instance, all locally Lipschitz functionals are strictly continuous and therefore have trivial horizon subgradients. This is also the case for many of the functionals discussed so far:

COROLLARY 2.33 (Some horizon subgradients). *At any $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$, the functionals defined in Propositions 2.10, 2.13, and 2.15 and Corollaries 2.12 and 2.19 have trivial horizon subgradient $\partial^\infty\,\mathcal{J} = \{\mathbf{0}_{\bar{\mu}}\}$.*

With horizon subgradients, we can finally provide a full characterization of epigraphical normals:

PROPOSITION 2.34 (Subgradients and epigraphical normals). *Let $\mathcal{J} : \mathcal{P}_2(\mathbb{R}^d) \to \bar{\mathbb{R}}$ and $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ at which $\mathcal{J}(\bar{\mu}) \in \mathbb{R}$. Then, $\boldsymbol{\xi} \in \mathrm{N}_{\mathrm{epi}\,\mathcal{J}}(\mu \times \delta_{\mathcal{J}(\mu)})$ if and only if either*

(i) $\mathbb{E}^{(\boldsymbol{x},\boldsymbol{v})\sim\boldsymbol{\xi}}\left[\langle (0,\ldots,0,1), \boldsymbol{v}\rangle\right] = -\lambda, \lambda > 0$ and $\frac{1}{\lambda}T_{\#}\boldsymbol{\xi} \in \partial\,\mathcal{J}(\mu)$; or

(ii) $\mathbb{E}^{(\boldsymbol{x},\boldsymbol{v})\sim\boldsymbol{\xi}}\left[\langle (0,\ldots,0,1), \boldsymbol{v}\rangle\right] = 0$ and $T_{\#}\boldsymbol{\xi} \in \partial^\infty\,\mathcal{J}(\mu)$.

Discussion. The definition of horizon subgradients introduced in Definition 2.31 follows the geometrical ideas used for Euclidean spaces in [50, Definition 1.18]. It departs from the alternative definition of the horizon subgradients as the scaled limit of regular subgradients [60, Definition 8.3], according to which $\xi \in \partial^\infty \mathcal{J}(\bar{\mu})$ if there are sequences $\mu_n \to \bar{\mu}, \mathcal{J}(\mu_n) \to \mathcal{J}(\bar{\mu})$ and $\xi_n \in \hat{\partial}\,\mathcal{J}(\mu_n)$ so that there exists $(\tau_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}_{\geq 0}, \tau_n \searrow 0$ and $\tau_n\,\xi_n \to \bar{\xi}$. Nonetheless, it is well known that, at least in Euclidean settings, the two definitions are identical, cf. [60, Theorem 8.9]. Finally, we could have defined the subgradients also as the non-horizontal epigraphical normals, as in [50, Definition 1.18]. Nonetheless, to ease the presentation, we opted for the more standard Definition 2.3, as in [60, Definition 8.3].

3. Optimality conditions.

In this section, we rigorously introduce the first-order necessary optimality conditions for (1.1). To start, we need a notion of local optimality:

DEFINITION 3.1 (Local optimality in the Wasserstein space). *A functional $\mathcal{J} : \mathcal{P}_2(\mathbb{R}^d) \to \bar{\mathbb{R}}$ attains a local minimum over a constraint set $\mathcal{C} \subseteq \mathcal{P}_2(\mathbb{R}^d)$ at $\mu^* \in \mathcal{C}$ if there exists $\varepsilon > 0$ such that for all $\mu \in \mathbb{B}_{W_2}(\varepsilon; \mu^*) \cap \mathcal{C}$ we have $\mathcal{J}(\mu) \geq \mathcal{J}(\mu^*)$.*

Armed with the tools of section 2, we now present the main result of this paper: first-order necessary optimality conditions in the Wasserstein space.

THEOREM 3.2 (First-order optimality conditions).
If a proper functional $\mathcal{J}$ attains a local minimum at $\mu^* \in \mathcal{C} \subseteq \mathcal{P}_2(\mathbb{R}^d)$ and the constraint qualification

$$\begin{cases} -\partial^\infty \mathcal{J}(\mu^*) \cap_{\mu^*} \mathrm{N}_{\mathcal{C}}(\mu^*) \subseteq \{\mathbf{0}_{\mu^*}\} \\ \mathcal{J}(\mathcal{C} \cap \mathbb{B}_{W_2}(\mu^*; \varepsilon)) \subseteq \mathbb{R} \end{cases} \tag{3.1}$$

holds, then

$$\mathbf{0}_{\mu^*} \in \partial\,\mathcal{J}(\mu^*) + \mathrm{N}_{\mathcal{C}}(\mu^*).$$

We highlight the similarity between our result and its counterpart in the Euclidean setting: At optimality, at least one element of the subgradient of the cost functional (direction of improvement) must align and have the same magnitude (but opposite direction) with one element of the normal cone (the "blocked" directions). In the unconstrained case, the normal cone trivializes (cf. Proposition 2.23) and we recover Fermat's rule:

THEOREM 3.3 (Fermat's rule in the Wasserstein space). *If a proper functional $\mathcal{J} : \mathcal{P}_2(\mathbb{R}^d) \to \bar{\mathbb{R}}$ attains a local minimum at μ^*, then $\mathbf{0}_{\mu^*} \in \hat{\partial}\,\mathcal{J}(\mu^*)$, which implies $\mathbf{0}_{\mu^*} \in \partial\,\mathcal{J}(\mu^*)$.*

Both results deserve a separate statement for the differentiable case[2]:

COROLLARY 3.4 (First-order optimality conditions, the differentiable case). *If a differentiable functional $\mathcal{J}$ has a local minimum $\mu^* \in \mathcal{C} \subseteq \mathcal{P}_2(\mathbb{R}^d)$, then*

$$-\nabla\,\mathcal{J}(\mu^*) \in \hat{\mathrm{N}}_{\mathcal{C}}(\mu^*) \subseteq \mathrm{N}_{\mathcal{C}}(\mu^*).$$

COROLLARY 3.5 (Fermat's rule in the Wasserstein Space, the differentiable case). *If a proper functional $\mathcal{J} : \mathcal{P}_2(\mathbb{R}^d) \to \bar{\mathbb{R}}$ is differentiable at $\mu^* \in \mathcal{P}_2(\mathbb{R}^d)$, then local optimality implies $\mathbf{0}_{\mu^*} = \nabla\,\mathcal{J}(\mu^*)$.*

[2]In view of the discussion in subsection 2.4.2, the use of "Corollary" here and in the next result is an abuse of terminology from the perspective of the "developer", but in this exposition we pay more attention to the "user" of our tools.

3.1. Discussion. Before showcasing our first-order necessary optimality conditions in various examples, we discuss several related aspects.

Relation to classical variational analysis. The conditions in Theorem 3.2, as well as those in the subsequent results, very much resemble their Euclidean (or more generally Hilbertian) counterpart: the negative subgradient of the objective function must lie in the normal cone of the feasible set, provided that a constraint qualification holds; e.g., see [60].

Constraint qualification. Condition (3.1), which we require for our necessary conditions, and the regularity conditions of normal cones, which we require for an expression for the normal cone (e.g., see Proposition 2.26), are so-called constraint qualifications. Depending on the specifics of the feasible set $\mathcal{C}$, these conditions can be made more explicit. When $\mathcal{J}$ is strictly continuous on $\mathcal{C}$, the horizon subgradient trivializes and the domain of $\mathcal{J}$ contains $\mathcal{C}$. Thus, both conditions in (3.1) are satisfied and we are left with the regularity conditions of normal cones. In the particular (and popular) case where $\mathcal{C}$ is a finite set of sufficiently regular (smooth) equality and inequality constraints (i.e., $G(\mu) = 0$ and $F(\mu) \leq 0$ differentiable for $G : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}^h$ and $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}^l$), we are therefore left with well-known constraint qualifications in Euclidean spaces (e.g., linearity constraint qualification, linear independence constraint qualification, quasi-normality constraint qualification). Indeed, Proposition 2.26 requires that the gradient of the constraint does not vanish, as predicated in [46, Theorem 3.4].

Comparison with the literature. General necessary optimality conditions in the Wasserstein space are studied in [46], in the smooth (or non-smooth but convex) setting where the constraint is the level set of a real-valued functional, and in [11, 12, 13] in the different setting of optimal control problems in the Wasserstein space. With this work, we provide a general theory of variational analysis in the Wasserstein space and significantly generalize existing necessary conditions for static optimization.

Computational aspects. Our optimality conditions transform the problem of minimizing a function over the infinite-dimensional probability space into an inclusion problem at the level of transport plans which, in turn, amounts to a set of conditions over $\mathbb{R}^d \times \mathbb{R}^d$. Depending on the application, they can be solved in (quasi) closed-form or addressed via (nonlinear) function approximators [4, 68], as we show in our examples. For this reason, we argue that our necessary conditions are significantly more tractable than the initial infinite-dimensional optimization problem. Nonetheless, more rigorous statements of the computational aspects of necessary conditions for optimality depend on the specifics of the problem at hand, as is already the case in Euclidean settings.

Sufficient conditions. As in Euclidean settings, our conditions are not sufficient for optimality. We expect sufficient conditions for optimality to be intimately related to geodesic convexity [3, §7] and second-order calculus in the Wasserstein space [32]; see [46] for preliminary results. We leave this topic to future research.

3.2. Pedagogical examples. In the remainder of this section, we illustrate our necessary conditions across several examples. We start by solving rigorously the example in section 1:

EXAMPLE 3.6 (Expected value subject to second moment constraints). *For $\theta \neq 0$ and $\varepsilon > 0$, consider the problem*

$$\inf_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \mathcal{J}(\mu) := \mathbb{E}^{x\sim\mu}\left[\langle\theta, x\rangle\right] \qquad \textit{subject to} \qquad \mathbb{E}^{x\sim\mu}\left[\|x\|^2\right] \leq \varepsilon^2.$$

By Example 2.14*, $\mathcal{J}$ is differentiable with gradient $\nabla\mathcal{J}(\mu) = (\mathrm{Id}, \theta)_{\#}\mu$. The normal* *cone of the second moment constraint was studied in Example* 2.28, $\mathrm{N}_{\mathcal{C}}(\bar{\mu}) = \{\mathbf{0}_{\bar{\mu}}\}$ *if* $\mathbb{E}^{x\sim\bar{\mu}}\left[\|x\|^2\right] < \varepsilon^2$ *and* $\mathrm{N}_{\mathcal{C}}(\bar{\mu}) = \{(\mathrm{Id}, 2\mathrm{Id})_{\#}\bar{\mu}\}$ *if* $\mathbb{E}^{x\sim\bar{\mu}}\left[\|x\|^2\right] = \varepsilon^2$*. Any candidate optimal solution* $\mu^* \in \mathcal{P}_2(\mathbb{R}^d)$ *must satisfy the optimality conditions in Corollary* 3.4*. We study two cases. If* $\mathbb{E}^{x\sim\mu^*}\left[\|x\|^2\right] < \varepsilon^2$*, it holds*

$$\begin{aligned}\mathbf{0}_{\mu^*} = \nabla\mathcal{J}(\mu^*) \Leftrightarrow 0 &= \|\nabla\mathcal{J}(\mu^*)\|^2_{\mu^*}\\ &= \int_{\mathbb{R}^d} \|\theta\|^2 \,\mathrm{d}\,\mu^*(x) > 0,\end{aligned}$$

which is absurd. Therefore, the optimal solution (if exists) lies at the boundary, and a $\lambda > 0$ *and a "sum coupling"* $\alpha \in \Gamma^1(\nabla\mathcal{J}(\mu^*), \lambda(\mathrm{Id}, 2\mathrm{Id})_{\#}\bar{\mu})$ *must exist so that*

$$\begin{aligned}\mathbf{0}_{\mu^*} = \nabla\mathcal{J}(\mu^*) +_\alpha \lambda(\mathrm{Id}, 2\mathrm{Id})_{\#}\bar{\mu} \Leftrightarrow 0 &= \left\|\nabla\mathcal{J}(\mu^*) +_\alpha (\mathrm{Id}, 2\mathrm{Id})_{\#}\bar{\mu}\right\|^2_{\mu^*}\\ &= \int_{\mathbb{R}^d} \|\theta + 2\lambda x\|^2 \,\mathrm{d}\,\mu^*(x) > 0,\end{aligned}$$

In particular, $\mu^* = \delta_{-\frac{\theta}{2\lambda}}$ *and we find* $\lambda = \frac{\|\theta\|}{2\varepsilon}$ *recalling that* $\mathbb{E}^{x\sim\mu^*}\left[\|x\|^2\right] = \varepsilon^2$*.*

We now show that necessary conditions for optimality can be used to produce certificates that an optimal solution does not exist:

EXAMPLE 3.7 (A certificate of unfeasibility). *Consider the problem of finding the worst-case mean-variance risk over the full probability space. In this case, we seek for the minimum of the functional*

$$\mathcal{J}(\mu) = -\left(\mathbb{E}^{x\sim\mu}\left[\langle\theta, x\rangle\right] + \frac{\rho}{2}\mathrm{Var}^{x\sim\mu}\left[\langle\theta, x\rangle\right]\right).$$

By Propositions 2.13 *and* 2.17 *and Corollary* 2.19*,* $\mathcal{J}$ *is differentiable with gradient*

$$\nabla\mathcal{J}(\mu) = -(\mathrm{Id}, (1 + \rho(\langle\theta, \mathrm{Id}\rangle - \mathbb{E}^{x\sim\mu}\left[\langle\theta, x\rangle\right]))\theta)_{\#}\mu.$$

To find the unconstrained minimizer of $\mathcal{J}$*, we can deploy Corollary* 3.5*:*

$$\begin{aligned}\mathbf{0}_{\mu} = \nabla\mathcal{J}(\mu) \Leftrightarrow 0 &= \|\nabla\mathcal{J}(\mu)\|^2_{\mu}\\ &= \int_{\mathbb{R}^d} \|(1 + \rho(\langle\theta, x\rangle - \mathbb{E}^{y\sim\mu}\left[\langle\theta, y\rangle\right]))\theta\|^2 \,\mathrm{d}\,\mu(x)\end{aligned}$$

Thus, at any optimal $\mu^* \in \mathcal{P}_2(\mathbb{R}^d)$*, it holds*

$$0 = 1 + \rho(\langle\theta, x\rangle - \mathbb{E}^{y\sim\mu^*}\left[\langle\theta, y\rangle\right]) \quad \mu^*\text{-a.e.},$$

but taking the expected value w.r.t. μ^* *of both sides yields* $0 = 1$*, a contradiction. Thus, the infimum of* $\mathcal{J}$ *is not attained.*

We now consider an example in the setting of DRO, a ubiquitous framework for decision-making under uncertainty. We seek to evaluate the worst-case mean-variance of a linear portfolio with allocation θ but now constrained to a closed Wasserstein ball of a given radius ε and center $\bar{\mu}$. Formally, this example is similar to Example 3.7, but now the worst-case is taken over a Wasserstein ball instead of the entire probability space. As we shall see below, our necessary conditions lead to a closed-form solution for the worst-case probability measure, generalizing [46, Section 4] to non-absolutely continuous measures (which, among others, allows us to study the data-driven setting where $\bar{\mu}$ is empirical) and [52, Appendix E], which does not provide a closed-form solution and assume positive variance of $\bar{\mu}$.

EXAMPLE 3.8 (A constrained problem: mean-variance DRO). *Consider the mean-variance functional $\mathcal{J}: \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}^d$ defined in Example 3.7 and the constraint $\mu \in \bar{\mathbb{B}}_{W_2}(\hat{\mu};\varepsilon)$ for some $\hat{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ and $\varepsilon > 0$. We computed the gradient of $\mathcal{J}$ in Example 3.7 and we recall from Example 2.27 that the normal cone of $\bar{\mathbb{B}}_{W_2}(\hat{\mu};\varepsilon)$ at any $\bar{\mu} \in \bar{\mathbb{B}}_{W_2}(\hat{\mu};\varepsilon)$ satisfies $\mathrm{N}_{\bar{\mathbb{B}}_{W_2}(\hat{\mu};\varepsilon)}(\bar{\mu}) \subseteq \{\xi \mid \xi \in (\pi_1, 2\lambda(\pi_1-\pi_2))_{\#}\Gamma_o(\bar{\mu},\hat{\mu}), \lambda \geq 0\}$. Then, by Corollary 3.4, at any optimal $\mu^* \in \mathcal{P}_2(\mathbb{R}^d)$ an optimal transport plan $\gamma \in \Gamma_o(\mu^*,\hat{\mu})$, a Lagrange multiplier $\lambda \geq 0$, and a "sum coupling" $\alpha \in \Gamma^1(\nabla\mathcal{J}(\mu), (\pi_1, \pi_1-\pi_2)_{\#}\gamma)$ exist so that $\mathbf{0}_{\mu^*} = \nabla\mathcal{J}(\mu^*) +_\alpha 2\lambda(\pi_1,\pi_1-\pi_2)_{\#}\gamma$. Equivalently,*

$$
\begin{aligned}
0 &= \min_{\alpha \in \Gamma^1(\nabla\mathcal{J}(\mu), 2\lambda(\pi_1,\pi_1-\pi_2)_{\#}\gamma)} \left\| \nabla\mathcal{J}(\mu^*) +_\alpha 2\lambda(\pi_1,\pi_1-\pi_2)_{\#}\gamma \right\|^2_{\mu^*} \\
&= \min_\alpha \int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \|\upsilon_1+\upsilon_2\|^2 \,\mathrm{d}\,\alpha(x,\upsilon_1,\upsilon_2) \\
&= \min_\alpha \int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \|(1+\rho(\langle\theta,x\rangle - \mathbb{E}^{z\sim\mu}[\langle\theta,z\rangle]))\theta + \upsilon_2\|^2 \,\mathrm{d}\,\alpha(x,\upsilon_1,\upsilon_2) \\
&= \min_\alpha \int_{\mathbb{R}^d\times\mathbb{R}^d} \|(1+\rho(\langle\theta,x\rangle - \mathbb{E}^{z\sim\mu}[\langle\theta,z\rangle]))\theta + \upsilon_2\|^2 \,\mathrm{d}(\pi_{13\#}\alpha)(x,\upsilon_2) \\
&= \min_\alpha \int_{\mathbb{R}^d\times\mathbb{R}^d} \|(1+\rho(\langle\theta,x\rangle - \mathbb{E}^{z\sim\mu}[\langle\theta,z\rangle]))\theta + 2\lambda\,\upsilon\|^2 \,\mathrm{d}((\pi_1,\pi_1-\pi_2)_{\#}\gamma)(x,\upsilon) \\
&= \int_{\mathbb{R}^d\times\mathbb{R}^d} \|(1+\rho(\langle\theta,x\rangle - \mathbb{E}^{z\sim\mu}[\langle\theta,z\rangle]))\theta + 2\lambda(y-x)\|^2 \,\mathrm{d}\,\gamma(x,y).
\end{aligned}
$$

Thus, if $\mu^* \in \mathbb{B}_{W_2}(\hat{\mu};\varepsilon)$ is optimal, we must have

$$
(1+\rho(\langle\theta,x\rangle - \mathbb{E}^{z\sim\mu^*}[\langle\theta,z\rangle]))\theta = 2\lambda(x-y) \quad \gamma\text{-a.e.} \tag{3.2}
$$

We now proceed in three steps:

Step 1. *We take the expected value w.r.t. γ of both sides of (3.2) to get $\theta = 2\lambda(\mathbb{E}^{x\sim\mu^*}[x] - \mathbb{E}^{y\sim\hat{\mu}}[y])$. If $\lambda = 0$, we reach the contradiction $\theta = 0$, consistently with Example 3.7. Thus, $\lambda > 0$ and*

$$
\mathbb{E}^{x\sim\mu^*}[\langle\theta,x\rangle] = \frac{\|\theta\|^2}{2\lambda} + \mathbb{E}^{y\sim\hat{\mu}}[\langle\theta,y\rangle]. \tag{3.3}
$$

Step 2. *We take the expected value w.r.t. γ of the squared norm in (3.2). The right-hand-side gives*

$$
\mathbb{E}^{(x,y)\sim\gamma}\left[\|2\lambda(x-y)\|^2\right] = 4\lambda^2\mathbb{E}^{(x,y)\sim\gamma}\left[\|x-y\|^2\right] = 4\lambda^2\varepsilon^2,
$$

where the last equality comes from $\lambda > 0$ and so $\mu^* \in \partial\mathbb{B}_{W_2}(\hat{\mu};\varepsilon)$. The left-hand-side, instead, reads

$$
\begin{aligned}
&\mathbb{E}^{(x,y)\sim\gamma}\left[\left\|(1+\rho(\langle\theta,x\rangle - \mathbb{E}^{z\sim\mu^*}[\langle\theta,z\rangle]))\theta\right\|^2\right] \\
&= \|\theta\|^2\left(1+\rho^2\mathbb{E}^{(x,y)\sim\gamma}\left[\left(\langle\theta,x\rangle - \mathbb{E}^{z\sim\mu^*}[\langle\theta,z\rangle]\right)^2\right]\right)
\end{aligned}
$$

$$= \|\theta\|^2 \left(1 + \rho^2 \mathrm{Var}^{x\sim\mu^*}\left[\langle \theta, x\rangle\right]\right).$$

Overall,

$$\mathrm{Var}^{x\sim\mu^*}\left[\langle \theta, x\rangle\right] = \frac{4\lambda^2\varepsilon^2 - \|\theta\|^2}{\rho^2 \|\theta\|^2}. \tag{3.4}$$

Step 3: *We write* (3.2) *as* $(\rho\|\theta\|^2 - 2\lambda)\langle \theta, x\rangle + const. = -2\lambda\langle \theta, y\rangle$ *and take the variance w.r.t.* γ *on both sides to get*

$$(\rho\|\theta\|^2 - 2\lambda)^2 \mathrm{Var}^{x\sim\mu^*}\left[\langle \theta, x\rangle\right] = 4\lambda^2 \mathrm{Var}^{y\sim\hat{\mu}}\left[\langle \theta, y\rangle\right]. \tag{3.5}$$

Thus, if $\mathrm{Var}^{y\sim\hat{\mu}}\left[\langle \theta, y\rangle\right] > 0$*, then* $\lambda \neq \frac{\rho\|\theta\|^2}{2}$*, and* (3.4) *and* (3.5) *yield* λ *from a polynomial equation (cf. [46, Proposition 4.7]). In this case, the optimal cost is*

$$\frac{\|\theta\|^2}{2\lambda} + \mathbb{E}^{y\sim\hat{\mu}}\left[\langle \theta, y\rangle\right] + \frac{\rho}{4\lambda^2(\rho\|\theta\|^2 - 2\lambda)^2}\mathrm{Var}^{y\sim\hat{\mu}}\left[\langle \theta, y\rangle\right]. \tag{3.6}$$

If, instead, $\mathrm{Var}^{y\sim\hat{\mu}}\left[\langle \theta, y\rangle\right] = 0$ *(and so* $\hat{\mu}$ *is not absolutely continuous), then* (3.5) *yields two cases. If* $\lambda = \frac{\rho\|\theta\|^2}{2}$*, then* (3.3) *and* (3.4) *yields the optimal cost*

$$\mathbb{E}^{y\sim\hat{\mu}}\left[\langle \theta, y\rangle\right] + \rho\|\theta\|^2\varepsilon^2. \tag{3.7}$$

If $\mathrm{Var}^{x\sim\mu^*}\left[\langle \theta, x\rangle\right] = 0$*, then* (3.4) *gives* $\lambda = \frac{\|\theta\|}{2\varepsilon}$ *and* (3.3) *yields the optimal cost*

$$\mathbb{E}^{y\sim\hat{\mu}}\left[\langle \theta, y\rangle\right] + \|\theta\|\varepsilon. \tag{3.8}$$

With λ*, the optimal solution follows from the inverses of the linear map (cf.* (3.2)*):*

$$y = T(x) = \left(I - \frac{\rho\theta\theta^\top}{2\lambda}\right)x - \frac{1}{2\lambda}\left(1 - \mathbb{E}^{y\sim\hat{\mu}}\left[\langle \theta, y\rangle\right] - \frac{\|\theta\|^2}{2\lambda}\right)\theta. \tag{3.9}$$

Finally, since $(\mathrm{Id}, T)_{\#}\mu^* = (\pi_1, \pi_1 - \pi_2)_{\#}\xi$*,* T *is an optimal transport map. By [70, Theorem 5.10], it is also a gradient of a convex function, which implies* $\lambda \geq \frac{\rho\|\theta\|^2}{2}$*.*

4. Applications. In this section, we showcase the practical appeal of our first-order optimality conditions by discussing several applications in machine learning and non-linear DRO.

4.1. Machine learning. In [66], we showcase how our first-order optimality conditions improve the computational efficiency of learning the dynamics of particles undergoing diffusion, bypassing cumbersome bi-level optimization procedures [16, 66]. In this section, we study applications to drug discovery and distribution fitting.

4.1.1. Drug discovery. In euclidean settings, the proximal operator appears as a (often closed-form) sub-routine of various first-order optimization algorithms [7]. It is intimately related to DRO [8] and recently, it has found applications in learning the dynamics of particles undergoing diffusion [16, 66], single cell perturbation analysis [17], and drug discovery [2, §6]. We use the first-order optimality conditions provided in this work to characterize the proximal operator, and we showcase how these results can be applied to molecular discovery. More precisely, following [2, §6], the task of drug discovery and drug re-purposing can be cast as an iterative process which aims to increasing the *drug-likeness* $\mathbb{E}^{x\sim\mu}[V(x)]$ of a given distribution of molecules μ while staying close to the original distribution μ_t,

$$\mu_{t+1} = \operatorname*{argmin}_{\mu\in\mathcal{P}(\mathbb{R}^d)} \mathbb{E}^{x\sim\mu}[V(x)] + \frac{1}{2}W_2(\mu,\mu_t)^2.$$

Differently from [2, §6], we do not require any convex proxy for the drug-likeness potential V:

EXAMPLE 4.1 (Proximal operator in the Wasserstein space). *Let $\bar{\mu}\in\mathcal{P}_2(\mathbb{R}^d)$ and $V:\mathbb{R}^d\to\mathbb{R}$ be differentiable function with quadratic growth (i.e., $|V(x)|\leq A+B\|x\|^2$). Consider the proximal operator in the Wasserstein space, defined by*

$$\operatorname*{argmin}_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \mathcal{J}(\mu) := \mathbb{E}^{x\sim\mu}[V(x)] + \frac{1}{2}W_2(\mu,\bar{\mu})^2. \tag{4.1}$$

We aim to characterize optimal solutions of (4.1)*. Since V is differentiable, the chain rule (with $\mathcal{J}_1 = \mathbb{E}^{x\sim\mu}[V(x)]$, $\mathcal{J}_2 = \frac{1}{2}W_2(\mu,\bar{\mu})^2$, and $g(x_1,x_2)=x_1+x_2$) gives*

$$\begin{aligned}\partial\mathcal{J}(\mu) &\subseteq \{v\in \mathrm{T}_{\mathcal{P}_2(\mathbb{R}^d)}(\mu) \,|\, \mathrm{supp}(v)\subseteq\partial V\} + (\pi_1,\pi_1-\pi_2)_{\#}\Gamma_o(\mu,\bar{\mu})\\ &= \{(\mathrm{Id},\nabla V)_{\#}\mu\} + (\pi_1,\pi_1-\pi_2)_{\#}\Gamma_o(\mu,\bar{\mu}).\end{aligned}$$

By Theorem 3.3*, the subgradient of $\mathcal{J}$ vanishes at an optimal solution μ^*. Thus, at optimality there exist $\gamma\in\Gamma_o(\mu^*,\bar{\mu})$ and $\alpha\in\Gamma^1((\mathrm{Id},\nabla V)_{\#}\mu^*,(\pi_1,\pi_1-\pi_2)_{\#}\gamma)$ so that $\mathbf{0}_{\mu^*} = (\mathrm{Id},\nabla V)_{\#}\mu^* +_\alpha (\pi_1,\pi_1-\pi_2)_{\#}\gamma$. Equivalently,*

$$\begin{aligned}0 &= \min_{\alpha\in\Gamma^1((\mathrm{Id},\nabla V)_{\#}\mu^*,(\pi_1,\pi_1-\pi_2)_{\#}\gamma)} \left\|(\mathrm{Id},\nabla V)_{\#}\mu^* +_\alpha (\pi_1,\pi_1-\pi_2)_{\#}\gamma\right\|_{\mu^*}^2\\ &= \min_\alpha \int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \|x_2+x_3\|^2 \,\mathrm{d}\,\alpha(x_1,x_2,x_3)\\ &= \min_\alpha \int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \|\nabla V(x_1)+x_3\|^2 \,\mathrm{d}\,\alpha(x_1,x_2,x_3)\\ &= \min_\alpha \int_{\mathbb{R}^d\times\mathbb{R}^d} \|\nabla V(x_1)+x_3\|^2 \,\mathrm{d}(\pi_{13\#}\alpha)(x_1,x_3)\\ &= \int_{\mathbb{R}^d\times\mathbb{R}^d} \|\nabla V(x_1)+x_3\|^2 \,\mathrm{d}((\pi_1,\pi_1-\pi_2)_{\#}\gamma)(x_1,x_3)\\ &= \int_{\mathbb{R}^d\times\mathbb{R}^d} \|\nabla V(x)+(x-y)\|^2 \,\mathrm{d}\,\gamma(x,y).\end{aligned}$$

Thus, for almost all $(x,y)\in\mathrm{supp}(\gamma)$, it necessarily holds

$$0 = \nabla V(x) + (x-y) = \nabla\left(V(x) + \frac{\|x-y\|^2}{2}\right).$$

That is, $(x,y)\in\mathrm{supp}(\gamma)$ implies that the gradient of $V(x)+\frac{\|x-y\|^2}{2}$ vanishes at x. The corresponding optimal cost is

$$\mathcal{J}(\mu^*) = \mathbb{E}^{(x,y)\sim\gamma}\left[V(x)+\frac{1}{2}\|x-y\|^2\right].$$

Thus, the cost-minimizing choice is $x \in \operatorname{argmin}_{z\in\mathbb{R}^d} V(z) + \frac{\|z-y\|^2}{2}$ *for* γ*-almost all* y,

$$\mathcal{J}(\mu^*) = \mathbb{E}^{y\sim\bar{\mu}}\left[\min_{x\in\mathbb{R}^d} V(x) + \frac{1}{2}\|x-y\|^2\right].$$

In particular, the proximal operator on $\mathbb{R}^d$ *characterizes the proximal operator* (4.1) *in* $\mathcal{P}_2(\mathbb{R}^d)$. *In the special case where the function* $x \mapsto V(x) + \frac{\|x\|^2}{2}$ *is strictly convex, we conclude* $\mu^* = (\mathrm{Id} + \nabla V)^{-1}{}_{\#}\bar{\mu}$.

We can further extend the result to include support constraints:

Example 4.2 (Proximal operator in the Wasserstein space, revisited). *Consider the setting of Example* 4.1 *and suppose that additionally the support of* μ *is restricted to a convex and closed set* $\mathbb{R}^d i \subseteq \mathbb{R}^d$, *so that the proximal operator reads*

$$\underset{\substack{\mu\in\mathcal{P}_2(\mathbb{R}^d)\\ \mathrm{supp}(\mu)\subseteq\mathbb{R}^d i}}{\operatorname{argmin}}\ \mathcal{J}(\mu) := \mathbb{E}^{x\sim\mu}[V(x)] + \frac{1}{2}W_2(\mu,\bar{\mu})^2. \tag{4.2}$$

In this case, the optimal solution can be characterized via Theorem 3.2. *To start, since* $\partial^\infty\mathcal{J}_1$ *and* $\partial^\infty\mathcal{J}_2$ *are trivial, the horizon subgradient* $\partial^\infty\mathcal{J}$ *also trivializes and the constraint qualification holds necessarily. Thus, we only need to study the normal cone of the feasible set. By Proposition* 2.24,

$$\mathrm{N}_{\mathcal{C}}(\mu) = \left\{\xi \in \mathrm{T}_{\mathcal{P}_2(\mathbb{R}^d)}(\mu) \,|\, (x,y)\in\mathrm{supp}(\xi) \Leftrightarrow y \in \mathrm{N}_{\mathbb{R}^d i}(x)\right\}.$$

Thus, by Theorem 3.2, *at the any optimal* μ^* *it holds*

$$\begin{aligned}\mathbf{0}_{\mu^*} = \{&(\mathrm{Id}, \nabla V)_{\#}\mu^*\}\\ &+_\alpha (\pi_1, \pi_1 - \pi_2)_{\#}\Gamma_o(\mu,\bar{\mu})\\ &+_\alpha \left\{\xi \in \mathrm{T}_{\mathcal{P}_2(\mathbb{R}^d)}(\mu^*) \,|\, (x,y)\in\mathrm{supp}(\xi) \Leftrightarrow y\in\mathrm{N}_{\mathbb{R}^d i}(x)\right\}\end{aligned}$$

for some "sum coupling" α. *Thus, at optimality there exist* $\gamma \in \Gamma_o(\mu^*,\bar{\mu})$, $\xi \in \mathrm{N}_{\mathcal{C}}(\mu)$, *and* $\alpha \in \Gamma^1((\mathrm{Id}, \nabla V)_{\#}\mu^*, (\pi_1,\pi_1-\pi_2)_{\#}\gamma, \xi)$ *so that* $\mathbf{0}_{\mu^*} = (\mathrm{Id}, \nabla V)_{\#}\mu^* +_\alpha (\pi_1,\pi_1-\pi_2)_{\#}\gamma +_\alpha v$ *(see Remark* 2.2 *for the sum of three elements). Equivalently,*

$$\begin{aligned}0 &= \left\|(\mathrm{Id}, \nabla V)_{\#}\mu^* +_\alpha (\pi_1,\pi_1-\pi_2)_{\#}\gamma +_\alpha \xi\right\|^2_{\mu^*}\\ &= \min_{\alpha\in\Gamma^1((\mathrm{Id},\nabla V)_{\#}\mu^*, (\pi_1,\pi_1-\pi_2)_{\#}\gamma, \xi)} \int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \|x_2+x_3+x_4\|^2 \,\mathrm{d}\,\alpha(x_1,x_2,x_3,x_4)\\ &= \int_{\mathbb{R}^d i\times\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \|\nabla V(x_1)+x_3+x_4\|^2 \,\mathrm{d}\,\alpha(x_1,x_2,x_3,x_4).\end{aligned}$$

Thus, for all $(x_1,x_2,x_3,x_4)\in\mathrm{supp}(\alpha)$, *it necessarily holds* $0 = \nabla V(x_1) + x_3 + x_4$ *or, equivalently, that for all* $(x,y)\in\mathrm{supp}(\gamma)$ *there is* $n \in \mathrm{N}_{\mathbb{R}^d i}(x)$ *so that*

$$0 = \nabla V(x) + (x-y) + n = \nabla\left(V(x) + \frac{\|x-y\|^2}{2}\right) + n$$

That is, x *satisfies necessary conditions for optimality of* $\min_{x\in\mathbb{R}^d i} V(x) + \frac{\|x-y\|^2}{2}$. *We then can argue as in Example* 4.1 *to conclude that it is optimal to select* $x \in$

$\mathrm{argmin}_{z\in\mathbb{R}^d i} V(z) + \frac{\|z-y\|^2}{2}$ *for γ-almost all y, which yields the optimal cost*

$$\mathcal{J}(\mu^*) = \mathbb{E}^{y\sim\bar{\mu}}\left[\min_{x\in\mathbb{R}^d i} V(x) + \frac{1}{2}\|x-y\|^2\right].$$

That is, the proximal operator on $\mathbb{R}^d$ characterizes the proximal operator (4.2) *in $\mathcal{P}_2(\mathbb{R}^d)$.*

To deploy these tools for drug discovery, we follow [2] and proceed in three steps:

(i) We train a Variational Auto-Encoder [40] on SMILES, a string representation of molecules [20, 55, 73] to obtain a 128-dimensional latent space.

(ii) We fit a neural network to the *drug-likeness* of the molecules, which can be computed with the RDKit library [45].

(iii) Here, we depart from [2] and, rather than attempting an optimization in the probability space, we use our theoretic results in Examples 4.1 and 4.2 to decouple the computation of the proximal operator for the single molecule. This computation, which is highly parallelizable, can be efficiently implemented by gradient descent and automatic differentiation [14].

4.1.2. Learning gaussians. Gaussian mixture models represent one of the most widespread techniques for distribution fitting [58, 66, 75]. As is well-known (e.g., see [75] and references therein) this problem can be formulated as an optimization problem in the probability spaces: Since a Gaussian mixture model can be written as $\mu * \phi$ with $\phi(x)$ being the Gaussian kernel $\frac{1}{(2\pi)^{-d/2}}\exp\left(-\frac{\|x\|^2}{2}\right)$[3], the learning problem can be cast as the optimization problem

$$\min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} -\sum_{i=1}^{n}\log\left(\int_{\mathbb{R}^d}\phi(x-X_i)\mathrm{d}\,\mu(x)\right), \tag{4.3}$$

where $\{X_1,\ldots,X_n\}$ is a given dataset [75]. Since the objective is differentiable, Corollary 3.5 yields the first-order necessary optimality condition

$$\begin{aligned} 0 &= \int_{\mathbb{R}^d}\left\|-\sum_{i=1}^{n}\frac{\nabla\phi(x-X_i)}{\int_{\mathbb{R}^d}\phi(y-X_i)\mathrm{d}\,\mu^*(y)}\right\|^2\mathrm{d}\,\mu^*(x) \\ &= \int_{\mathbb{R}^d}\left\|\sum_{i=1}^{n}\frac{(x-X_i)\exp\left(-\frac{\|x-X_i\|^2}{2}\right)}{\int_{\mathbb{R}^d}\exp\left(-\frac{\|y-X_i\|^2}{2}\right)\mathrm{d}\,\mu^*(y)}\right\|^2\mathrm{d}\,\mu^*(x), \end{aligned} \tag{4.4}$$

where we used Proposition 2.17 to evaluate the subgradient. First, it is apparent that any probability measure supported on $\{X_1,\ldots,X_n\}$ satisfies the first-order optimality conditions, but this solution is of course not interesting. Second, one can parametrize μ with finitely many samples $x_1,\ldots,x_m$, with $m<n$, with weights $\rho_1,\ldots,\rho_m$, and search for the $x_1,\ldots,x_m$ and $\rho_1,\ldots,\rho_m$ which best "fits" the first-order optimality condition. We showcase the results of the method in Figure 4.

4.2. Non-linear mean-variance optimization. In this section, we extend the existing duality result for linear mean-variance optimization [8, 9, 43, 76] to the non-linear case. Specifically, for $\hat{\mu}\in\mathcal{P}_2(\mathbb{R}^d), \varepsilon>0$ and $V:\mathbb{R}^d\to\mathbb{R}$ satisfying the

[3] Here, we consider mixture of Gaussians with unit variance, but the discussion in this section can easily be extended for general mixtures.

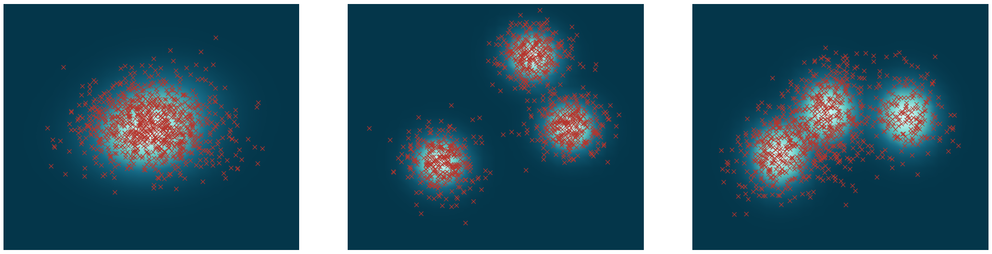


Fig. 4: Data (blue crosses) generated from a mixture of different number of Gaussians (left: one, middle: three, right: five) is fit using the method outlined in subsection 4.1.2 with three components.

assumptions of Corollary 2.19, consider the problem of evaluating the *worst-case* risk over an Wasserstein ball:

$$\inf_{\mu \in \mathbb{B}_{W_2}(\hat{\mu};\varepsilon)} -\left(\mathbb{E}^{x\sim\mu}\left[V(x)\right] + \rho \mathrm{Var}^{x\sim\mu}\left[V(x)\right]\right). \tag{4.5}$$

In this section, we deploy our results to prove that any candidate optimal solution $\mu^* \in \mathcal{P}_2(\mathbb{R}^d)$ yields the worst-case cost

$$\mathcal{J}(\mu^*) = \min_{\substack{\lambda \geq 0 \\ \beta \in \mathbb{R}}} \lambda\varepsilon^2 + \int_{\mathbb{R}^d} \max_{y \in \mathbb{R}^d} \left\{V(y) + (V(y)-\beta)^2 - \lambda \left\|y - \hat{x}\right\|^2\right\} \mathrm{d}\,\hat{\mu}(\hat{x}). \tag{4.6}$$

Moreover, with λ^*, β^* being the minimizers above, there exists $\gamma \in \Gamma_o(\mu^*, \hat{\mu})$ such that

$$x^* \in \operatorname*{argmax}_{y \in \mathbb{R}^d} \left\{V(y) + (V(y)-\beta^*)^2 - \lambda^* \left\|y - \hat{x}\right\|^2\right\} \quad \forall (x^*, \hat{x}) \in \mathrm{supp}(\gamma). \tag{4.7}$$

We start with the observation that, for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\mathrm{Var}^{x\sim\mu}\left[V(x)\right] = \inf_{\beta \in \mathbb{R}} \mathbb{E}^{x\sim\mu}\left[(V(x)-\beta)^2\right].$$

For $\beta \in \mathbb{R}$, consider the parametric problem

$$\inf_{\mu \in \mathbb{B}_{W_2}(\hat{\mu};\varepsilon)} -\mathbb{E}^{x\sim\mu}\left[V(x) + \rho(V(x)-\beta)^2\right]. \tag{4.8}$$

Theorem 3.2 combined with Proposition 2.13 and Example 2.27 requires that, for any optimal solution $\mu^*_\beta \in \mathcal{P}_2(\mathbb{R}^d)$, $\xi_V \in \mathrm{T}_{\mathcal{P}_2(\mathbb{R}^d)}(\mu^*)$, $\gamma \in \Gamma_o(\mu^*_\beta, \hat{\mu})$, $\xi = (\pi_1, \pi_2 - \pi_1)_\# \gamma$, and $\lambda^*_\beta > 0$ (for $\lambda^*_\beta = 0$, see Example 3.7, so that $W_2(\mu^*, \hat{\mu}) = \varepsilon$) exist such that

$$\mathbf{0}_{\mu^*{}_\beta} \in \left(\pi_1, \pi_2(1 + 2\rho(V \circ \pi_1 - \beta)) + 2\lambda^*_\beta(\pi_1 - \pi_3)\right)_\# \Gamma^1(\xi_V, \xi), \qquad \mathrm{supp}(\xi_V) \subseteq \partial V.$$

After basic algebraic manipulations, these conditions provide the inclusion relation:

$$\hat{x} \in x^*_\beta + \frac{1}{2\lambda^*_\beta} \partial V(x^*_\beta)(1 + 2\rho(V(x^*_\beta) - \beta)) \quad \forall (x^*_\beta, \hat{y}) \in \mathrm{supp}(\gamma).$$

Analogously to Examples 4.1 and 4.2, we conclude that it is optimal to select

$$x^*_\beta \in \operatorname*{argmax}_{y \in \mathbb{R}^d} \left\{V(y) + \rho(V(y)-\beta)^2 - \lambda^*_\beta \left\|y - \hat{x}\right\|^2\right\}$$

for γ-almost all $\hat{x}$, which yields the optimal value of (4.8):

$$\begin{aligned}
&\int_{\mathbb{R}^d} \max_{y\in\mathbb{R}^d} \left\{ V(y) + \rho(V(y)-\beta)^2 - \lambda^*_\beta \left\|y-\hat{x}\right\|^2 \right\} \mathrm{d}\,\hat{\mu}(\hat{x}) \\
&\qquad = \int_{\mathbb{R}^d\times\mathbb{R}^d} V(x^*_\beta) + \rho(V(x^*_\beta)-\beta)^2 - \lambda^*_\beta \left\|x^*_\beta-\hat{x}\right\|^2 \mathrm{d}\,\gamma(x^*_\beta,\hat{x}) \\
&\qquad = \mathbb{E}^{x^*_\beta\sim\mu^*_\beta}\left[V(x^*_\beta) + \rho(V(x^*_\beta)-\beta)^2\right] - \lambda^*_\beta\varepsilon^2
\end{aligned}$$

Since this holds for any $\beta\in\mathbb{R}$, we can take the infimum over $\beta\in\mathbb{R}$ to obtain

$$\mathcal{J}(\mu^*) = \inf_{\beta\geq 0} -\left(\lambda^*_\beta\varepsilon^2 + \int_{\mathbb{R}^d}\max_{y\in\mathbb{R}^d}\left\{V(y)+\rho(V(y)-\beta)^2 - \lambda^*_\beta\left\|y-\hat{x}\right\|^2\right\}\mathrm{d}\,\hat{\mu}(\hat{x})\right)$$

and, thus, the inequality

$$\mathcal{J}(\mu^*) \geq \inf_{\substack{\lambda\geq 0\\ \beta\in\mathbb{R}}} -\left(\lambda\varepsilon^2 + \int_{\mathbb{R}^d}\max_{y\in\mathbb{R}^d}\left\{V(y)+\rho(V(y)-\beta)^2 - \lambda\left\|y-\hat{x}\right\|^2\right\}\mathrm{d}\,\hat{\mu}(\hat{x})\right). \tag{4.9}$$

For the other inequality, we deploy the standard Lagrangian argument. Define $\mathcal{L}(\mu,\lambda) \coloneqq \mathcal{J}(\mu) + \lambda(\varepsilon^2 - W_2\left(\mu,\hat{\mu}\right)^2)$. Then, for all $\lambda\geq 0$,

$$\begin{aligned}
\mathcal{L}(\mu,\lambda) &\geq \inf_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \mathcal{J}(\mu) + \lambda(\varepsilon^2 - W_2\left(\mu,\hat{\mu}\right)^2) \\
&\geq \inf_{\mu\in\mathbb{B}_{W_2}(\hat{\mu};\varepsilon)} \mathcal{J}(\mu) + \lambda(\varepsilon^2 - W_2\left(\mu,\hat{\mu}\right)^2) \\
&\overset{\heartsuit}{=} \mathcal{J}(\mu^*) + \lambda(\varepsilon^2 - W_2\left(\mu^*,\hat{\mu}\right)^2) \\
&\geq \mathcal{J}(\mu^*).
\end{aligned}$$

In particular, $\heartsuit$ holds because $\mathcal{L}(\mu^*,\lambda)$ is non-decreasing in $\lambda\geq 0$, so that we obtain

$$\begin{aligned}
\mathcal{J}(\mu^*) &\leq \inf_{\lambda\geq 0}\mathcal{L}(\mu^*,\lambda) \\
&= \inf_{\lambda\geq 0}\mathcal{J}(\mu^*) + \lambda\underbrace{(\varepsilon^2 - W_2\left(\mu^*,\hat{\mu}\right)^2)}_{=0} \\
&\overset{(4.2)}{\leq} \inf_{\substack{\lambda\geq 0\\ \beta\in\mathbb{R}}} -\left(\lambda^*_\beta\varepsilon^2 + \int_{\mathbb{R}^d}\max_{y\in\mathbb{R}^d}\left\{V(y)+\rho(V(y)-\beta)^2 - \lambda^*_\beta\left\|y-\hat{x}\right\|^2\right\}\mathrm{d}\,\hat{\mu}(\hat{x})\right).
\end{aligned}$$

Combining this inequality with (4.9) we get the equality in (4.6).

REFERENCES

[1] M. Agueh and G. Carlier, *Barycenters in the wasserstein space*, SIAM Journal on Mathematical Analysis, 43 (2011), pp. 904–924.

[2] D. Alvarez-Melis, Y. Schiff, and Y. Mroueh, *Optimizing functionals on the space of probabilities with input convex neural networks*, arXiv preprint arXiv:2106.00774, (2021).

[3] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient flows: in metric spaces and in the space of probability measures*, Springer Science & Business Media, Basel, 2005.

[4] B. Amos, L. Xu, and J. Z. Kolter, *Input convex neural networks*, in International Conference on Machine Learning, PMLR, 2017, pp. 146–155.

[5] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, *Coherent measures of risk*, Mathematical finance, 9 (1999), pp. 203–228.

[6] F. Bach and L. Chizat, *Gradient descent on infinitely wide neural networks: Global convergence and generalization*, arXiv preprint arXiv:2110.08084, (2021).
[7] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces, 2011*, Springer, New York, NY, 2011.
[8] J. Blanchet and K. Murthy, *Quantifying distributional model risk via optimal transport*, Mathematics of Operations Research, 44 (2019), pp. 565–600.
[9] J. Blanchet, K. Murthy, and F. Zhang, *Optimal transport-based distributionally robust optimization: Structural properties and iterative schemes*, Mathematics of Operations Research, 47 (2022), pp. 1500–1529.
[10] M. Bongini, M. Fornasier, F. Rossi, and F. Solombrino, *Mean-field pontryagin maximum principle*, Journal of Optimization Theory and Applications, 175 (2017), pp. 1–38.
[11] B. Bonnet, *A pontryagin maximum principle in wasserstein spaces for constrained optimal control problems*, ESAIM: Control, Optimisation and Calculus of Variations, 25 (2019), p. 52.
[12] B. Bonnet and H. Frankowska, *Necessary optimality conditions for optimal control problems in wasserstein spaces*, Applied Mathematics & Optimization, 84 (2021), pp. 1281–1330.
[13] B. Bonnet and F. Rossi, *The pontryagin maximum principle in the wasserstein space*, Calculus of Variations and Partial Differential Equations, 58 (2019), pp. 1–36.
[14] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, *JAX: composable transformations of Python+NumPy programs*, 2018.
[15] Y. Brenier, *Polar factorization and monotone rearrangement of vector-valued functions*, Communications on pure and applied mathematics, 44 (1991), pp. 375–417.
[16] C. Bunne, L. Meng-Papaxanthos, A. Krause, and M. Cuturi, *Proximal optimal transport modeling of population dynamics*, in International Conference on Artificial Intelligence and Statistics (AISTATS), 2022.
[17] C. Bunne, S. G. Stark, G. Gut, J. S. Del Castillo, M. Levesque, K.-V. Lehmann, L. Pelkmans, A. Krause, and G. Rätsch, *Learning single-cell perturbation responses using neural optimal transport*, Nature Methods, (2023), pp. 1–10.
[18] G. Buttazzo, L. De Pascale, and P. Gori-Giorgi, *Optimal-transport formulation of electronic density-functional theory*, Physical Review A, 85 (2012), p. 062502.
[19] G. Carlier, *Optimal transportation and economic applications*, Lecture Notes, 18 (2012).
[20] V. Chenthamarakshan, P. Das, S. Hoffman, H. Strobelt, I. Padhi, K. W. Lim, B. Hoover, M. Manica, J. Born, T. Laino, et al., *Cogmol: Target-specific and selective drug design for covid-19 using deep generative models*, Advances in Neural Information Processing Systems, 33 (2020), pp. 4320–4332.
[21] S. Chewi, T. Maunu, P. Rigollet, and A. J. Stromme, *Gradient descent algorithms for bures-wasserstein barycenters*, in Conference on Learning Theory, 2020, pp. 1276–1304.
[22] L. Chizat and F. Bach, *On the global convergence of gradient descent for over-parameterized models using optimal transport*, Advances in neural information processing systems, 31 (2018).
[23] C. Chu, J. Blanchet, and P. Glynn, *Probability functional descent: A unifying perspective on gans, variational inference, and reinforcement learning*, in International Conference on Machine Learning, PMLR, 2019, pp. 1213–1222.
[24] J. C. Duchi, P. W. Glynn, and H. Namkoong, *Statistics of robust optimization: A generalized empirical likelihood approach*, Mathematics of Operations Research, 46 (2021), pp. 946–969.
[25] P. Embrechts, A. Schied, and R. Wang, *Robustness in the optimization of risk measures*, Operations Research, 70 (2022), pp. 95–110.
[26] P. M. Esfahani and D. Kuhn, *Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations*, arXiv preprint arXiv:1505.05116, (2015).
[27] A. Figalli and F. Glaudo, *An invitation to optimal transport, Wasserstein distances, and gradient flows*, European Mathematical Society, 2021.
[28] M. Fisher, J. Nocedal, Y. Trémolet, and S. J. Wright, *Data assimilation in weather forecasting: a case study in pde-constrained optimization*, Optimization and Engineering, 10 (2009), pp. 409–426.
[29] C. Frogner and T. Poggio, *Approximate Inference with Wasserstein Gradient Flows*, in International Conference on Artificial Intelligence and Statistics, vol. 108, 2020, pp. 2581–2590.
[30] W. Gangbo and A. Tudorascu, *On differentiability in the wasserstein space and well-posedness for hamilton–jacobi equations*, Journal de Mathématiques Pures et Appliquées, 125 (2019), pp. 119–174.

[31] R. Gao and A. Kleywegt, *Distributionally robust stochastic optimization with wasserstein distance*, Mathematics of Operations Research, 48 (2023), pp. 603–655.

[32] N. Gigli, *Second Order Analysis on $P_2(M, W_2)$*, American Mathematical Soc., 2012.

[33] A. A. Gosavi, S. K. Das, and S. L. Murray, *Beyond exponential utility functions: A variance-adjusted approach for risk-averse reinforcement learning*, in 2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), 2014, pp. 1–8.

[34] I. Greenberg, Y. Chow, M. Ghavamzadeh, and S. Mannor, *Efficient risk-averse reinforcement learning*, Advances in Neural Information Processing Systems, 35 (2022), pp. 32639–32652.

[35] A. Hakobyan and I. Yang, *Wasserstein distributionally robust control of partially observable linear systems: Tractable approximation and performance guarantee*, in 2022 IEEE 61st Conference on Decision and Control (CDC), 2022, pp. 4800–4807.

[36] E. Hazan, S. Kakade, K. Singh, and A. Van Soest, *Provably efficient maximum entropy exploration*, in International Conference on Machine Learning, PMLR, 2019, pp. 2681–2691.

[37] R. Jordan, D. Kinderlehrer, and F. Otto, *The variational formulation of the fokker–planck equation*, SIAM Journal on Mathematical Analysis, 29 (1998), pp. 1–17.

[38] JP Morgan, *Risk Metrics*, tech. report, JP Morgan, 1996.

[39] L. V. Kantorovich, *On the Translocation of Masses*, Journal of Mathematical Sciences, 133 (2006), pp. 1381–1382, https://doi.org/10.1007/s10958-006-0049-2.

[40] D. P. Kingma and M. Welling, *Auto-encoding variational bayes*, arXiv preprint arXiv:1312.6114, (2013).

[41] J. Knoblauch, J. Jewson, and T. Damoulas, *An optimization-centric view on bayes' rule: Reviewing and generalizing variational inference*, The Journal of Machine Learning Research, 23 (2022), pp. 5789–5897.

[42] P. Krokhmal, M. Zabarankin, and S. Uryasev, *Modeling and optimization of risk*, Surveys in operations research and management science, 16 (2011), pp. 49–66.

[43] 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, pp. 130–166.

[44] S. Kurcyusz, *On the existence and nonexistence of lagrange multipliers in banach spaces*, Journal of Optimization Theory and Applications, 20 (1976), pp. 81–110.

[45] G. Landrum et al., *Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling*, Greg Landrum, 8 (2013), p. 5281.

[46] N. Lanzetti, S. Bolognani, and F. Dörfler, *First-order conditions for optimization in the wasserstein space*, arXiv preprint arXiv:2209.12197, (2022).

[47] D. G. Luenberger, *Optimization by vector space methods*, John Wiley & Sons, New York, NY, 1997.

[48] H. Markowitz, *Portfolio Selection*, The Journal of Finance, 7 (1952), pp. 77–91.

[49] G. Monge, *Mémoire sur la théorie des déblais et de remblais*, in Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année, De l'Imprimerie Royale, 1781.

[50] B. S. Mordukhovich, *Variational analysis and applications*, vol. 30, Springer, 2018.

[51] H. Namkoong and J. C. Duchi, *Variance-based regularization with convex objectives*, Advances in neural information processing systems, 30 (2017).

[52] V. A. Nguyen, S. S. Abadeh, D. Filipović, and D. Kuhn, *Mean-covariance robust risk measurement*, arXiv preprint arXiv:2112.09959, (2021).

[53] V. M. Panaretos and Y. Zemel, *An invitation to statistics in Wasserstein space*, Springer Nature, Cham, 2020.

[54] J. Pilipovsky and P. Tsiotras, *Distributionally robust density control with wasserstein ambiguity sets*, arXiv preprint arXiv:2403.12378, (2024).

[55] D. Polykovskiy, A. Zhebrak, B. Sanchez-Lengeling, S. Golovanov, O. Tatanov, S. Belyaev, R. Kurbanov, A. Artamonov, V. Aladinskiy, M. Veselov, et al., *Molecular sets (moses): a benchmarking platform for molecular generation models*, Frontiers in pharmacology, 11 (2020), p. 565644.

[56] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*, John Wiley & Sons, New York, NY, 2014.

[57] H. Rahimian and S. Mehrotra, *Distributionally robust optimization: A review*, arXiv preprint arXiv:1908.05659, (2019).

[58] D. Reynolds, *Gaussian Mixture Models*, Springer US, Boston, MA, 2009, pp. 659–663, https://doi.org/10.1007/978-0-387-73003-5_196.

[59] P. Rigollet and J. Weed, *Entropic optimal transport is maximum-likelihood deconvolution*,

Comptes Rendus. Mathématique, 356 (2018), pp. 1228–1235.
[60] R. T. Rockafellar and R. J. B. Wets, *Variational analysis*, Springer Science & Business Media, Berlin, 2009.
[61] F. Santambrogio, *Optimal transport for applied mathematicians*, Birkäuser, NY, 55 (2015), p. 94.
[62] F. Santambrogio, {*Euclidean, metric, and Wasserstein*} *gradient flows: an overview*, Bulletin of Mathematical Sciences, 7 (2017), pp. 87–154.
[63] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, *Trust region policy optimization*, in International conference on machine learning, 2015, pp. 1889–1897.
[64] S. Shafieezadeh-Abadeh, L. Aolaritei, F. Dörfler, and D. Kuhn, *New perspectives on regularization and computation in optimal transport-based distributionally robust optimization*, arXiv preprint arXiv:2303.03900, (2023).
[65] A. Terpin, N. Lanzetti, and F. Dörfler, *Dynamic programming in probability spaces via optimal transport*, arXiv preprint arXiv:2302.13550, (2023).
[66] A. Terpin, N. Lanzetti, and F. Dörfler, *Learning diffusion at lightspeed*, arXiv preprint arXiv:todo.todo, (2024).
[67] A. Terpin, N. Lanzetti, B. Yardim, F. Dorfler, and G. Ramponi, *Trust region policy optimization with optimal transport discrepancies: Duality and algorithm for continuous actions*, Advances in Neural Information Processing Systems, 35 (2022), pp. 19786–19797.
[68] T. Uscidda and M. Cuturi, *The monge gap: A regularizer to learn all transport maps*, arXiv preprint arXiv:2302.04953, (2023).
[69] B. P. Van Parys, D. Kuhn, P. J. Goulart, and M. Morari, *Distributionally robust control of constrained stochastic systems*, IEEE Transactions on Automatic Control, 61 (2015), pp. 430–442.
[70] C. Villani, *Optimal transport: old and new*, Springer, Berlin, 2009.
[71] A. Wachi and Y. Sui, *Safe reinforcement learning in constrained markov decision processes*, in International Conference on Machine Learning, 2020, pp. 9797–9806.
[72] Z. Wang, P. W. Glynn, and Y. Ye, *Likelihood robust optimization for data-driven problems*, Computational Management Science, 13 (2016), pp. 241–261.
[73] D. Weininger, *Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules*, Journal of Chemical Information and Computer Sciences, 28 (1988), pp. 31–36, https://doi.org/10.1021/ci00057a005, https://doi.org/10.1021/ci00057a005.
[74] W. Wiesemann, D. Kuhn, and M. Sim, *Distributionally robust convex optimization*, Operations research, 62 (2014), pp. 1358–1376.
[75] Y. Yan, K. Wang, and P. Rigollet, *Learning gaussian mixtures using the wasserstein-fisher-rao gradient flow*, arXiv preprint arXiv:2301.01766, (2023).
[76] M.-C. Yue, D. Kuhn, and W. Wiesemann, *On linear optimization over wasserstein balls*, Mathematical Programming, 195 (2022), pp. 1107–1122.
[77] J. Zhang and K. Cho, *Query-efficient imitation learning for end-to-end autonomous driving*, arXiv preprint arXiv:1605.06450, (2016).
[78] R. Zhang, C. Chen, C. Li, and L. Carin, *Policy optimization as wasserstein gradient flows*, in 35th International Conference on Machine Learning, vol. 13, 2018, pp. 5737–5746.