Nonlinear Distributionally Robust Optimization

Mohammed Rayyan Sheriff and Peyman Mohajerin Esfahani

ABSTRACT. This article focuses on a class of distributionally robust optimization (DRO) problems where, unlike the growing body of the literature, the objective function is potentially non-linear in the distribution. Existing methods to optimize nonlinear functions in probability space use the Frechet derivatives, which present both theoretical and computational challenges. Motivated by this, we propose an alternative notion for the derivative and corresponding smoothness based on Gateaux (G)-derivative for generic risk measures. These concepts are explained via three running risk measure examples of variance, entropic risk, and risk on finite support sets. We then propose a G-derivative based Frank-Wolfe (FW) algorithm for generic non-linear optimization problems in probability spaces and establish its convergence under the proposed notion of smoothness in a completely norm-independent manner. We use the set-up of the FW algorithm to devise a methodology to compute a saddle point of the non-linear DRO problem. Finally, for the minimum variance portfolio selection problem we analyze the regularity conditions and compute the FW-oracle in various settings, and validate the theoretical results numerically.

Keywords. Gateaux derivative, norm-free-smoothness, Frank-Wolfe algorithm, saddle point

1. INTRODUCTION

Modern-day decision problems involve uncertainty in the form of a random variable ξ whose behavior is modeled via a probability distribution $\mathbb{P}_o$. A central object to formalize such decision-making problems under uncertainty is risk measures. The most popular risk measure is arguably the expected loss, yielding the classical decision-making problem of

$$\text{Stochastic Program:} \qquad \min_{x \in \mathcal{X}} \mathbb{E}_{\mathbb{P}_o}\big[\ell(x, \xi)\big], \tag{1}$$

where ℓ is the loss function of interest, and $\mathcal{X}$ being the set of feasible decisions. The paradigm of stochastic programming (SP) relies on the assumption that the distribution $\mathbb{P}_o$ is available (or at least up to sufficient statistics), thereby the expectation can be computed for every decision $x \in \mathcal{X}$. A common practical challenge is, however, that the complete information of $\mathbb{P}_o$ may not be available. Moreover, it might also be the case that the distribution is varying over a period of time which could be difficult to characterize. These limitations call for a more conservative risk measure to ameliorate the decision performance in such situations.

An alternative framework is Robust Optimization (RO) where the decision-maker has only access to the support of uncertainty and takes the most conservative approach:

$$\text{Robust Optimization:} \qquad \min_{x \in \mathcal{X}} \max_{\xi \in \Xi} \ell(x, \xi). \tag{2}$$

For many interesting examples, the RO min-max problem admits tractable reformulations that can be solved efficiently [2]. However, a generic RO problem is known to be computationally formidable as the inner maximization over ξ can be NP-hard. Moreover, if the support Ξ of the distribution is "large", the results of RO tend to be extremely conservative.

Distributionally Robust Optimization (DRO). The SP and RO decision models represent two extreme cases of having full or bare minimum distributional information, respectively. In practice, however, we often have more information about the ground truth distribution than just its support. A typical example is when we have sample realizations $\{\widehat{\xi}_i : i = 1, 2, \ldots, N\}$. Looking at such settings through the lens of SP, one may construct a *nominal distribution* $\widehat{\mathbb{P}}$ and use it as a proxy for $\mathbb{P}_o$ in the SP (1). A standard data-driven nominal distribution is the discrete distribution $\widehat{\mathbb{P}} = \sum_{i=1}^{N} \delta_{\widehat{\xi}_i}$. The SP decision when $\mathbb{P}_o = \widehat{\mathbb{P}}$ in (1) often admits a poor out-of-sample performance on a different dataset, a phenomenon that is also known as the "optimizer's curse" or "overfitting" [38]. On the other hand, the RO viewpoint in (2) completely disregards the statistical information of $\mathbb{P}_o$ available through the dataset $(\widehat{\xi}_i)_i$, or any other form of prior information.

An attempt to bridge the SP and RO modeling frameworks gives rise to the paradigm of *Distributionally Robust Optimization* (DRO), which dates back to the Scarf's seminal work on the ambiguity-averse newsvendor problem in 1958 [35]. The "ambiguity" set $\mathcal{P}$ is a family of distributions that are close in some sense to the nominal distribution $\widehat{\mathbb{P}}$, potentially including the true unknown distribution $\mathbb{P}_o$. With this in mind, the DRO problem is formulated as

$$\min_{x \in \mathcal{X}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(x, \xi)]. \tag{3}$$

If the ambiguity is *very big*, possibly including all possible distributions, the DRO problem (3) reduces to the RO problem (2). On the other hand, if the ambiguity set $\mathcal{P}$ is *very small*, potentially a singleton containing the nominal distribution, then the DRO problem reduces to the SP problem (1). In this light, the DRO framework (3) provides flexibility for the decision-maker to fill the gap between SP (1) and RO (2). The ambiguity set $\mathcal{P}$ in (3) is typically constructed either based on the moments information [9, 15, 42], or a neighborhood of $\widehat{\mathbb{P}}$ with respect to a notion of distance over probability distributions, e.g., Prohorov [11], Kullback-Leiber [18, 10], Wasserstein [18, 24, 20, 4, 13, 14], Sinkhoron [41], to name but a few; see also the survey [33] and the references therein.

Linear DRO problems: An important feature of (3) is the linearity of the objective function in the distribution $\mathbb{P}$, which is also shared among all the literature mentioned above. The simplicity of this linearity in the inner maximization of (3) is the underlying driving force to develop tractable convex reformulations and computational solutions for various combinations of ambiguity sets and cost functions [32, 44, 34, 37, 6]. This line of research effectively translates the original infinite-dimensional DRO problem (3) to a tractable finite-dimensional one; see also the general optimal transport framework of [43], and the case of mean-covariance risk measure [30] in a financial context.

Nonlinear DRO (NDRO): Our goal here is to generalize the linear setting (3) to

$$\min_{x \in \mathcal{X}} \sup_{\mathbb{P} \in \mathcal{P}} F(x, \mathbb{P}), \qquad (4)$$

where $F : \mathcal{X} \times \mathcal{P} \to \mathbb{R}$ is a generic, possibly nonlinear, function signifying a risk measure. We refer to the decision-making problem (4) as *Non-linear Distributionally Robust Optimization (NDRO).* There are several interesting NDRO examples including the variance [3], entropy [39], [7, Example 1, p. 14]. In this study, instead of focusing only on the outer decision x in (4), we aim to compute an approximate saddle point between the decision-maker and the nature presented by the distribution $\mathbb{P}$. A motivation supporting this effort is the fact that, unlike the worst-case distribution computed for a given decision, the saddle point distributions (also called Nash equilibria) naturally retain more realistic features [36]. From a computational perspective, the existing techniques deployed in linear DRO problems cannot be directly extended for the NDRO (4). The objective of this work is to precisely tackle this challenge, where we seek to devise a methodology along with mild regularity conditions under which a saddle point solution exists and can be computed.

Frank-Wolfe (FW) algorithm. It is a first order method that only uses the information of the gradient to solve a constrained optimization problem [17, 12, 8]. In a nutshell, each iteration of the FW algorithm optimizes the linear function given by the gradient of the objective function and then takes a step towards the optimizer of the linear problem. On the contrary, other first order methods like projected gradient descent require a quadratic function to be minimized at each iteration. Since solving a linear problem at each iteration is easier than a quadratic one, the iteration complexity of the FW algorithm is much simpler than other first order methods. This is crucial for optimization problems over probability distributions since the complexity of projecting onto the ambiguity set can be as challenging as solving the original NDRO problem. In comparison, optimizing a cost function that is linear in distributions admits strong duality and tractable finite-dimensional reformulations for many interesting examples as seen in linear DRO. Therefore, we seek to use the principles of the Frank-Wolfe algorithm in the context of NDRO problems and devise an iterative procedure to compute a saddle point of (4).

The FW algorithm for optimization problems over probability distributions has already been introduced in [19], which uses Frechet derivatives and the associated notion of smoothness to establish convergence. We would also like to highlight the work of [22], where the canonical gradient ascent-descent algorithm [25, 16] for min-max problems is extended to infinite dimensional spaces involving probability distributions using the Frechet derivatives and its smoothness. In a similar spirit, the recent work of [29] proposes a mirror-descent algorithm [26] for constrained nonlinear optimization problems over probability distributions using the Frechet derivative. However, Frechet derivatives are difficult to deal with in practice due to several prominent challenges including their (i) existence, (ii) finite representability, and (iii) norm consistency; see Section 3.3 for more details on this. Our focus in this study is to remedy this by proposing a FW algorithm based on an alternative G-derivative along with a completely norm-independent convergence analysis.

Contributions. The main contributions of this study are summarized as follows:

(i) **Norm free smoothness in probability spaces.** We propose a novel G-derivative based notion of derivative for non-linear risk measures (Definition 3.1), and the associated notion of

smoothness that is independent of the norm structure on the ambiguity set (Definition 3.6). Moreover, we also derive conditions on the function $F(x, \mathbb{P})$ such that the risk measure $\mathbb{P} \longmapsto \min_{x \in \mathcal{X}} F(x, \mathbb{P})$ is smooth in the sense of the proposed notion (Lemma 5.5).

(ii) **G-derivative based Frank-Wolfe algorithm.** We provide a Frank-Wolfe algorithm based on the proposed notion of derivative for optimizing generic non-linear risk measures. Moreover, the classical proof techniques for the FW algorithm carry forward under the proposed notion of derivative and smoothness resulting in apriori (Propositions 4.2) and aposteriori (Proposition 4.3) convergence guarantees that exist for finite-dimensional problems.

(iii) **Saddle point seeking algorithm for NDRO problems.** For the potentially infinite-dimensional min-max problem of NDRO, we propose a FW-based algorithm to compute a saddle point (Algorithm 1), and also quantify its convergence properties (Theorem 5.6).

(iv) **Minimum variance portfolio optimization problem.** We study the NDRO problem and our proposed algorithm for the minimum variance portfolio optimization problem in detail. To this end, we first establish the required convex regularity conditions for the algorithm (Lemma 6.2), and then provide a complete description of the corresponding FW-oracle under various settings (Lemmas 6.4 and 6.6). In the special case of $\Xi = \mathbb{R}^n$ and the type-2 Wasserstein ambiguity set, we slightly extend the results of [3] by providing a description of the saddle point of the minimum variance problem when the feasible portfolio set $\mathcal{X}$ is any arbitrary compact set (Proposition 6.5).

Organisation. In Section 2, we discuss a generic potentially non-linear optimization problem over distributions with relevant examples. In Section 3, we introduce the notion of a directional derivative and the associated notion of smoothness. In Section 4, the FW-algorithm and its convergence guarantees are discussed. In Section 5, we introduce the non-linear distributionally robust optimization problem, discuss a solution concept, and provide a FW-based algorithm to compute the solution. Finally, in Section 6 we discuss in detail, the NDRO problem in the context of minimum variance portfolio optimization under different settings, accompanied by some simulation results.

Notations. We employ standard notation throughout the article. The set of real-valued $n \times n$ matrices is denoted by $\mathbb{R}^{n \times n}$, the trace of a matrix M is denoted by $\mathrm{tr}(M)$. A matrix M being positive semi-definite is denoted by $M \succeq 0$. For a function r defined over a finite-dimensional space, its gradient at x is denoted by $\nabla r(x)$. If $\mathbb{P}, \mathbb{Q}$ are two probability distributions, then $\mathbb{E}_{\mathbb{P}}[\cdot]$ is the expectation with respect to the probability distribution $\mathbb{P}$, and $\mathbb{E}_{\mathbb{P}-\mathbb{Q}}[\cdot] := \mathbb{E}_{\mathbb{P}}[\cdot] - \mathbb{E}_{\mathbb{Q}}[\cdot]$. Besides these, any new definition or notation is defined whenever encountered.

2. Worst-case non-linear risk measures

In this section, we focus our attention to the potentially infinite-dimensional optimization problem:

$$R^* := \sup_{\mathbb{P} \in \mathcal{P}} R(\mathbb{P}), \tag{5}$$

where $R : \mathcal{P} \to \mathbb{R}$ is a desired, possibly *nonlinear*, concave risk measure, and $\mathcal{P}$ is the ambiguity set containing a family of probability distributions over $\Xi \subset \mathbb{R}^n$. Our main goal is to develop a framework with appropriate mathematical notions for a *Franke-Wolfe* (FW) like algorithm to solve problem (5) and investigate its convergence properties. Furthermore, since the FW algorithm for

(5) operates in an infinite dimensional setting, we also seek to derive its tractable finite-dimensional simplification to solve specific instances of (5).

Definition 2.1 (Regular risk (RR) measures). *A risk measure R in (5) is regular if it can be described as $R(\mathbb{P}) = r\big(\mathbb{E}_{\mathbb{P}}[L(\xi)]\big)$, for some functions $L : \Xi \to \mathbb{R}^m$ that is integrable for all $\mathbb{P} \in \mathcal{P}$ and $r : \mathbb{R}^m \to \mathbb{R}$ that is concave and differentiable.*

Example 2.2 (RR-examples). Throughout this study, we discuss three particular examples of the regular risk measures to showcase the concepts and our theoretical statement:

(a) *Variance*: A popular example of such a risk measure is the variance associated with the distribution. Formally, considering ξ to be a $\mathbb{P}$-distributed random variable for $\mathbb{P} \in \mathcal{P}$, the associated variance is

$$V(\mathbb{P}) := \mathbb{E}_{\mathbb{P}}\big[\|\xi - \mathbb{E}_{\mathbb{P}}[\xi]\|_2^2\big]. \tag{6}$$

Considering the functions $L : \Xi \to \mathbb{R}^{n \times n} \times \mathbb{R}^n$ and $r : \mathbb{R}^{n \times n} \times \mathbb{R}^n \to \mathbb{R}$

$$\begin{cases} L(\xi) = (\xi\xi^\top, \, \xi) \\ r(\Sigma, \mu) = \mathrm{tr}(\Sigma) - \|\mu\|_2^2, \end{cases} \tag{7}$$

one can observe that the variance is indeed an RR measure, since

$$V(\mathbb{P}) \;=\; \mathbb{E}_{\mathbb{P}}\big[\|\xi\|_2^2\big] - \|\mathbb{E}_{\mathbb{P}}[\xi]\|_2^2 \;=\; \mathrm{tr}\Big(\mathbb{E}_{\mathbb{P}}[\xi\xi^\top]\Big) - \|\mathbb{E}_{\mathbb{P}}[\xi]\|_2^2 \;=\; r\Big(\mathbb{E}_{\mathbb{P}}[L(\xi)]\Big).$$

(b) *Entropic risk*: Another interesting example of a non-linear RR measure is the entropic risk of a multi variate distribution [39, Section 5]. If $\mathbb{P} \in \mathcal{P}$ is a distribution with marginals $\mathbb{P}_j$ for $j = 1, 2, \ldots, n$, (i.e., the j-th component ξ_j is $\mathbb{P}_j$ distributed). The entropic risk $\mathcal{E}(\mathbb{P})$ associated with the distribution $\mathbb{P}$ is defined as

$$\mathcal{E}(\mathbb{P}) := \sum_{j=1}^{n} \frac{1}{\theta_j} \log\Big(\mathbb{E}_{\mathbb{P}_j}[e^{-\theta_j \xi_j}]\Big), \tag{8}$$

where $(\theta_j)_{j=1}^n$ is a collection of positive real numbers referred to as the *risk-aversion parameters*. The $\mathcal{E}$-measure is indeed an RR measure, which is seen by considering the respective functions: $L : \Xi \to \mathbb{R}^n$ and $r : (0, +\infty)^n \to \mathbb{R}$

$$\begin{cases} L(\xi) = \Big(e^{-\theta_1 \xi_1}, e^{-\theta_2 \xi_2}, \ldots, e^{-\theta_n \xi_n}\Big) \\ r(z) = \sum_{j=1}^{n} \frac{1}{\theta_j} \log(z_j). \end{cases} \tag{9}$$

(c) *Finite-support*: The final example is the case of Ξ being a finite set: $\Xi = \{\xi_i : i = 1, 2, \ldots, N\}$. In this case, the simplex Δ^N is the set of all probability distributions on Ξ, and the ambiguity set of distributions is a subset, $\mathcal{P} \subset \Delta^N$. It turns out that *any* arbitrary risk measure R can be characterized as an RR-measure in the sense of Definition 2.1 by introducing appropriate functions $r : \mathbb{R}^N \to \mathbb{R}$, and $L : \Xi \to \mathbb{R}^N$. To see this, we first observe that Ξ being a finite set gives rise to the enumerating bijection $b : \Xi \to \{1, 2, \ldots, N\}$ defined by $b(\xi) := \{i : \xi = \xi_i\}$; secondly, the matrix $M \in \mathbb{R}^{N \times N}$ given by $[M]_{ij} := (j/N)^i$ for $i, j = 1, 2, \ldots, N$, is invertible.

Then, considering the functions

$$\begin{cases} L(\xi) = \left(\left(b(\xi)/N \right), \left(b(\xi)/N \right)^2, \ldots, \left(b(\xi)/N \right)^N \right)^\top \\ r(z) = R\left(M^{-1} z \right) \quad \text{for} \quad z \in M(\mathcal{P}), \end{cases} \tag{10}$$

we have $\mathbb{E}_\mathbb{P}[L(\xi)] = M \cdot \mathbb{P}$ (viewing $\mathbb{P} \in \Delta^N$ as an element of $\mathbb{R}^N$), and as such, $R(\mathbb{P}) = r(\mathbb{E}_\mathbb{P}[L(\xi)])$ for all $\mathbb{P} \in \mathcal{P}$.

3. Derivatives of risk measures

A well-defined notion of the gradient is a fundamental quantity in developing iterative algorithms like that of FW to solve any optimization problem. For finite-dimensional convex problems, the FW algorithm optimizes the linear functional given by the gradient of the objective function over the feasible set at each iteration. Naturally, devising a similar algorithm for (5) requires at least a well-defined notion of *directional derivative* or G-derivative [1, (A.3), p. 152].

3.1. **Gateaux directional derivatives**

Definition 3.1 (G-derivative). *Given $\mathbb{P}, \mathbb{Q} \in \mathcal{P}$, the Gateaux (G)-derivative $dR(\mathbb{P}; \mathbb{Q})$ of the risk measure R at $\mathbb{P}$ in the direction $\mathbb{Q}$ is defined as*

$$dR(\mathbb{P}; \mathbb{Q}) := \lim_{\gamma \downarrow 0} \frac{1}{\gamma} \left(R(\mathbb{P}_\gamma) - R(\mathbb{P}) \right), \quad \text{where} \quad \mathbb{P}_\gamma = \mathbb{P} + \gamma(\mathbb{Q} - \mathbb{P}), \tag{11}$$

whenever the limit exists, and we say that the function R is $\mathbb{Q}$-directionally differentiable at $\mathbb{P}$. Moreover, we say that the risk measure R is directionally differentiable on $\mathcal{P}$ if it is $\mathbb{Q}$-directionally differentiable at every $\mathbb{P} \in \mathcal{P}$ and for all $\mathbb{Q} \in \mathcal{P}$.

We note that the G-directional derivative in Definition 3.1 does not rely on any metric underlying the space of probability distributions. This is in fact the main feature with respect to the alternative F-derivative (Frechet derivative) that will be discussed in Section 3.3. Figure 1 visualizes this directional derivative. Next lemma provides an explicit description of Definition 3.1 for regular risk measures.

FIGURE 1. Pictorial representation of the risk surface and its directional derivative.

Lemma 3.2 (Regular G-derivatives). *Suppose that the risk measure is regular, i.e., $R(\mathbb{P}) = r(\mathbb{E}_\mathbb{P}[L(\xi)])$ with ∇r denoting the gradient of function r. Then, for any $\mathbb{P}, \mathbb{Q} \in \mathcal{P}$, we have*

$$dR(\mathbb{P}; \mathbb{Q}) = \mathbb{E}_{\mathbb{Q}-\mathbb{P}} \left[\langle \nabla r(\mathbb{E}_\mathbb{P}[L(\xi)]), L(\xi) \rangle \right] = \langle \nabla r(\mathbb{E}_\mathbb{P}[L(\xi)]), \mathbb{E}_{\mathbb{Q}-\mathbb{P}}[L(\xi)] \rangle. \tag{12}$$

Lemma 3.2 indicates that the G-derivative of an RR-measure is essentially characterized by the finite-dimensional quantities $\mathbb{E}_\mathbb{P}[L(\xi)]$ and $\mathbb{E}_\mathbb{Q}[L(\xi)]$ (cf. (12)). Thus, algorithms that optimize an RR
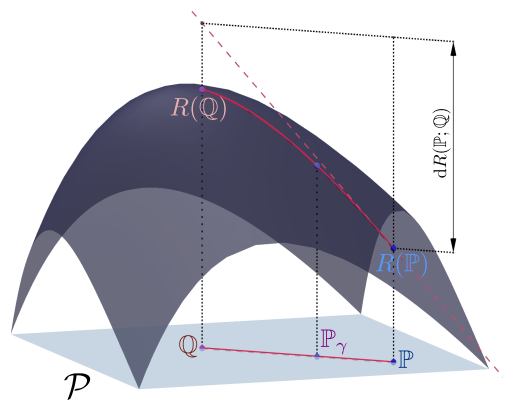
measure using their directional derivatives only need to track the evolution of these finite dimensional quantities, allowing them to be implemented tractably.

Example 3.3 (Regular G-derivatives). The directional derivative of an RR measure is completely characterized in terms of only a few finite dimensional quantities that depend on the moments of the distribution, and the functions r and L.

(a) *Variance*: With the underlying inner product $\langle(\Sigma, \mu), (\Sigma', \mu')\rangle := \operatorname{tr}(\Sigma^\top \Sigma') + \mu^\top \mu'$, we recall from (7) that $\nabla r(\Sigma, \mu) = \left(\mathbb{I}_n, -2\mu\right)$. Then the $\mathbb{Q}$-directional derivative of the variance V, at $\mathbb{P}$, is easily calculated from (12) as

$$dV(\mathbb{P}; \mathbb{Q}) = \operatorname{tr}\left(\Sigma_\mathbb{Q} - \Sigma_\mathbb{P}\right) - 2\mu_\mathbb{P}^\top(\mu_\mathbb{Q} - \mu_\mathbb{P}). \tag{13}$$

(b) $\mathcal{E}$-*measure*: With the canonical inner product $\langle z, z'\rangle = z^\top z'$, we recall from (9) that $\nabla r(z) = \left((\theta_1 z_1)^{-1}, (\theta_2 z_2)^{-1}, \ldots, (\theta_n z_n)^{-1}\right)$ at every $z \in \mathbb{R}_+^n$ for the $\mathcal{E}$-measure. Thus, its $\mathbb{Q}$-directional derivative at $\mathbb{P}$, is

$$d\mathcal{E}(\mathbb{P}; \mathbb{Q}) = \sum_{j=1}^n \frac{\mathbb{E}_{\mathbb{Q}_j - \mathbb{P}_j}[e^{-\theta_j \xi_j}]}{\theta_j \mathbb{E}_{\mathbb{P}_j}[e^{-\theta_j \xi_j}]}. \tag{14}$$

(c) *Finite support*: Recall from (10) that $\mathbb{E}_\mathbb{P}[L(\xi)] = M \cdot \mathbb{P}$ for all $\mathbb{P} \in \mathcal{P}$. Then with the canonical inner product $\langle z, z'\rangle = z^\top z'$, we have $\nabla r(z) = (M^{-1})^\top \nabla R(M^{-1} z)$ for all $z \in M(\mathcal{P})$. Substituting these quantities in (12) and simplifying, we get

$$dR(\mathbb{P}; \mathbb{Q}) = (\nabla R(\mathbb{P}))^\top (\mathbb{Q} - \mathbb{P}). \tag{15}$$

It is to be observed that the matrix M has no relevance in the G-derivative as one would expect since the risk measure R is defined independent of the matrix M.

The G-derivative enjoys inherent properties that will be helpful to devise computational solutions.

Proposition 3.4 (G-derivative: properties). *For any $\mathbb{P}, \mathbb{Q} \in \mathcal{P}$, and $\gamma \in [0, 1]$, let $\mathbb{P}_\gamma = \mathbb{P} + \gamma(\mathbb{Q} - \mathbb{P})$, then the directional derivative $dR(\mathbb{P}; \cdot)$ in Definition 3.1 satisfies*

$$\text{Positively homogeneous:} \quad dR(\mathbb{P}; \mathbb{P}_\gamma) = \gamma dR(\mathbb{P}; \mathbb{Q}) \tag{16a}$$

$$\text{Upper bound for concave risk measures:} \quad R(\mathbb{Q}) - R(\mathbb{P}) \leqslant dR(\mathbb{P}; \mathbb{Q}). \tag{16b}$$

Remark 3.5 (Optimality conditions). *Similar to KKT conditions, F-derivative based first order optimality conditions for both constrained and unconstrained versions of (5) are given in [21, Section 3]. With G-derivatives, the upper bound (16b) for the concave risk measure R immediately gives sufficient optimality conditions for (5). More precisely, if $\mathbb{P}^* \in \operatorname{argmax}_{\mathbb{Q} \in \mathcal{P}} dR(\mathbb{P}^*; \mathbb{Q})$, then applying (16b) for any $\mathbb{Q} \in \mathcal{P}$ yields*

$$R(\mathbb{Q}) \leqslant R(\mathbb{P}^*) + dR(\mathbb{P}^*; \mathbb{Q}) \leqslant R(\mathbb{P}^*) + dR(\mathbb{P}^*; \mathbb{P}^*) = R(\mathbb{P}^*),$$

which establishes the optimality $\mathbb{P}^* \in \operatorname{argmax}_{\mathbb{Q} \in \mathcal{P}} R(\mathbb{Q})$.

3.2. Smoothness

The proposed approach to solve (5) builds on the Frank-Wolfe (FW) algorithm for finite dimensional convex optimization problems [12]. It is a well known fact that the primal sub-optimality in FW algorithm converges at a sub-linear rate $O(1/k)$ for optimization of "smooth" objective functions over compact feasible sets in finite dimensional problems. A convex function is said to be smooth if it has Lipschitz continuous gradients with respect to some norm. The choice of norm in an infinite dimensional setting can be problematic, as all the norms are not equivalent (unlike the finite dimensional setting). To extend such convergence attributes for a FW like algorithm in the infinite dimensional setting of (5), we first propose an appropriate notion of smoothness in terms of the directional derivatives (as in Definition 3.1) that is also norm independent.

Definition 3.6 (G-smoothness). *The risk measure R is C-smooth if there exists a constant $C \geqslant 0$ such that for all $\mathbb{P}, \mathbb{Q} \in \mathcal{P}$ and $\gamma \in [0, 1]$, we have the inequality*

$$\mathrm{d}R\big(\mathbb{P}_\gamma; \mathbb{P}\big) + \mathrm{d}R\big(\mathbb{P}; \mathbb{P}_\gamma\big) \;\leqslant\; \gamma^2 C \quad \textit{where } \mathbb{P}_\gamma = \mathbb{P} + \gamma\big(\mathbb{Q} - \mathbb{P}\big) \;. \tag{17}$$

Connection to the existing notions of smoothness. The notion of smoothness in Definition 3.6 is a generalization of the canonical smoothness condition of Lipschitz continuous gradients [27, Section 2.1.5]. Formally, a function $f : \mathcal{D} \to \mathbb{R}$ with a finite dimensional domain $(\mathcal{D}, \|\cdot\|)$ and gradients ∇f is said to be β-smooth if

$$\|\nabla f(x) - \nabla f(y)\|_* \;\leqslant\; \beta \|x - y\| \quad \text{holds for all } x, y \in \mathcal{D}. \tag{18}$$

It is further generalized (or relaxed) by the notion of Holder-smoothness (particularly, 1-Holder smooth) where it is required that

$$\big\langle \nabla f(x) - \nabla f(y), y - x \big\rangle \;\leqslant\; \beta \|x - y\|^2 \quad \text{holds for all } x, y \in \mathcal{D}. \tag{19}$$

Furthermore, if the set $\mathcal{D}$ is $\|\cdot\|$-bounded in addition, then the notion of Holder-smoothness (19) is sufficient to the requirement that there exists some constant $C \geqslant 0$ such that for every $x, y \in \mathcal{D}$, $\gamma \in [0, 1]$, the function f satisfies the inequality

$$\big\langle \nabla f(x) - \nabla f(x_\gamma), x_\gamma - x \big\rangle \;\leqslant\; \gamma^2 C \quad \text{where} \;\; x_\gamma := x + \gamma(y - x). \tag{20}$$

Expanding the left-hand side of (20) as $\big\langle \nabla f(x) - \nabla f(x_\gamma), x_\gamma - x \big\rangle = \big\langle \nabla f(x), x_\gamma - x \big\rangle + \big\langle \nabla f(x_\gamma), x - x_\gamma \big\rangle$, we see that the individual terms are simply the directional derivatives of f; Definition 3.6 of the proposed notion of smoothness becomes apparent at once. We also highlight that the notion of smoothness in Definition 3.6 is closely related to the notion of "curvature coefficient" used in [17, (3)]. In fact, it can be easily shown that a function has a finite curvature coefficient if it is smooth in the sense of Definition 3.6.

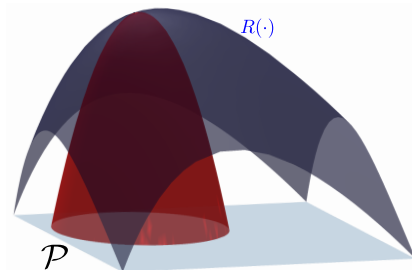FIGURE 2. Concave-quadratic lower bound via smoothness.

Concave quadratic lower bound via smoothness. It is well established that in a finite-dimensional setting, smoothness conditions like (18) and (19) give rise to a global convex (resp. concave) quadratic upper bound (resp. lower bound). For the risk measure R specifically, a sample representation of such a quadratic lower bound is shown in Figure

2. The existence of such bounds guarantees that the curvature of the function is at most that of the quadratic bounds, which is crucial in concluding the convergence of the FW algorithm. In other words, if a function f is β-canonically smooth (as in (18)), then the following inequality holds:

$$-\frac{\beta}{2}\|y-x\|^2 \leqslant f(y) - f(x) - \langle \nabla f(x),\, y-x\rangle \leqslant \frac{\beta}{2}\|y-x\|^2 \quad \text{for all } x,y \in \mathcal{D}.$$

The notion of smoothness in Definition 3.6 imposes similar quadratic bounds, but without using any norm. This is done by enforcing "quadratic-like" bounds to hold uniformly over all directions. The following Lemma establishes the smoothness of the class of RR measures.

Lemma 3.7 (Regular G-smoothness). *If a regular risk measure $R(\mathbb{P}) = r(\mathbb{E}_{\mathbb{P}}[L(\xi)])$, satisfies*

(i) *bounded diameter in expectations:* $d := \sup_{\mathbb{P},\mathbb{Q}\in\mathcal{P}} \|\mathbb{E}_{\mathbb{P}}[L(\xi)] - \mathbb{E}_{\mathbb{Q}}[L(\xi)]\| < +\infty$

(ii) *smooth r in the canonical sense* (18):

$$\|\nabla r(u) - \nabla r(v)\|_* \leqslant \beta \|u - v\| \quad \text{holds for all } u,v \in \{\mathbb{E}_{\mathbb{P}}[L(\xi)] : \mathbb{P} \in \mathcal{P}\}.$$

Then, it is (βd^2)-smooth in the sense of Definition 3.6.

Example 3.8 (Regular G-smoothness). For the RR measures in Examples 2.2 and 3.3 with the same underlying norms, the smoothness constant β of the respective r-function is as follows:

(a) *Variance*: Recalling the function r from (7), and that $\nabla r(\Sigma, \mu) = (\mathbb{I}_n, -2\mu)$, we get $\beta = 2$.
(b) *Entropic risk*: Assume that there exists $b > 0$ such that $b \leqslant \mathbb{E}_{\mathbb{P}_j}[e^{-\theta_j \xi_j}]$ for all $j = 1, 2, \ldots, n$, and $\mathbb{P} = (\mathbb{P}_j)_j \in \mathcal{P}$, and letting $\theta_{\min} := \min_{j \leqslant n} \theta_j$, we see that the function $r : [b, +\infty)^n \to \mathbb{R}$ as given in (9) has β-canonically smooth for $\beta = (b^2 \theta_{\min})^{-1}$.
(c) *Finite support*: Assuming that the risk measure R has β'-Lipschitz continuous gradients on Δ^N with respect to some norm $\|\cdot\|$. It follows from (10), that $\beta = \beta' \|M^{-1}\|_o^2$, where, $\|M^{-1}\|_o = \sup_{\|u\|\leqslant 1} \|M^{-1}u\|$.

One of the advantages of the proposed notion of smoothness is that it allows us to employ inequalities and bound sets in a finite-dimensional space to guarantee the smoothness of the risk measures. Since all norms on finite-dimensional spaces are equivalent, establishing that $\beta, d < +\infty$ with respect to the same norm is not restrictive, even though this may give rise to dimension-dependent constants β and d.

3.3. Frechet derivatives

An important observation to be made is that the $\mathbb{Q}$-directional derivative $dR(\mathbb{P};\mathbb{Q})$ of regular-risk measures is affine in $\mathbb{Q}$.[1] In principle, without any further assumptions on the risk measure R, its $\mathbb{Q}$-directional derivative need not be affine in $\mathbb{Q}$. Counter examples of functions that have well defined G-derivatives in all directions but that are non-linear in the direction exist even among functions defined on $\mathbb{R}^2$, let alone the infinite dimensional setting of $\mathcal{P}$. A sufficient condition for the directional derivative to be affine in $\mathbb{Q}$ is the existence of a stronger notion of derivative called the *Frechet-derivative* or *F-derivative*.

---

[1] A function $f : \mathcal{S} \to \mathbb{R}$ is said to affine if $f(x + \theta(y - x)) = f(x) + \theta(f(y) - f(x))$ for every $x, y \in \mathcal{S}$ and $\theta \in [0, 1]$.

Definition 3.9 (Frechet-derivative)**.** *The Frechet(F)-derivative of the risk measure R at $\mathbb{P} \in \mathcal{P}$ associated with a given norm $\|\cdot\|_{\mathcal{P}}$ on $\mathcal{P}$, is a function $\ell_{\mathbb{P}} : \Xi \to \mathbb{R}$ such that the mapping $\mathcal{P} \ni \mathbb{P}' \longmapsto \mathbb{E}_{\mathbb{P}'}[\ell_{\mathbb{P}}(\xi)]$ is continuous w.r.t. $\|\cdot\|_{\mathcal{P}}$ and satisfies*

$$0 = \lim_{\|\mathbb{P}'-\mathbb{P}\|_{\mathcal{P}} \downarrow 0} \frac{R(\mathbb{P}') - R(\mathbb{P}) - \mathbb{E}_{\mathbb{P}'-\mathbb{P}}[\ell_{\mathbb{P}}(\xi)]}{\|\mathbb{P}'-\mathbb{P}\|_{\mathcal{P}}}. \tag{21}$$

Smoothness with F-derivatives. If the F-derivative $\ell_{\mathbb{P}}$ were to exist at every $\mathbb{P} \in \mathcal{P}$, the canonical notion of smoothness (37) can be naturally extended into the infinite dimensional setting. The risk measure R is said to be *Frechet-smooth* if there exists some $\beta \geqslant 0$ such that its F-derivative $\ell_{\mathbb{P}}$ satisfies

$$\text{Frechet-Smoothness:} \quad \|\ell_{\mathbb{P}} - \ell_{\mathbb{P}'}\|_{\mathcal{P}*} \leqslant \beta \|\mathbb{P} - \mathbb{P}'\|_{\mathcal{P}} \quad \text{for all } \mathbb{P}, \mathbb{P}' \in \mathcal{P}, \tag{22}$$

where $\|\cdot\|_{\mathcal{P}*}$ is the dual norm of $\|\cdot\|_{\mathcal{P}}$. If the risk measure in (5) is Frechet-smooth, most of the convergence analysis due to smoothness in finite-dimensional convex problems simply carries forward to the infinite-dimensional setting right away. This provides a natural recipe to extend the FW algorithm into the infinite dimensional setting of probability spaces and establish their convergence under F-smoothness. F-derivative based FW-algorithms in probability spaces have already been studied in the literature [19] with a slightly more general notion of smoothness than (22), and for a slightly more general class of risk measures than the usual concavity assumption. Our approach differs from [19] in the fact that we only make use of G-derivatives in both the development of the FW-algorithm and also in establishing its convergence based on only G-derivative based regularity conditions, which are simpler to deal with than F-derivatives.

Comparison with G-derivatives. It is not necessary for a function to have F-derivatives even if it has affine directional derivatives in all directions, such counterexamples exist even in a finite-dimensional setting. On the contrary, if the risk measure R has a well-defined F-derivative $\ell_{\mathbb{P}}$, then it can be shown that its $\mathbb{Q}$-directional derivatives also exist in all directions $\mathbb{Q} \in \mathcal{P}$. This is easily seen by considering $\mathbb{P}' = \mathbb{P} + \gamma(\mathbb{Q} - \mathbb{P})$ for $\gamma \in [0, 1]$ in the definition (21) and observing that $\mathbb{E}_{\mathbb{P}'-\mathbb{P}}[\ell_{\mathbb{P}}(\xi)] = \gamma \mathbb{E}_{\mathbb{Q}-\mathbb{P}}[\ell_{\mathbb{P}}(\xi)]$. When $\mathbb{Q} \neq \mathbb{P}$, the limit $\|\mathbb{P}' - \mathbb{P}\|_{\mathcal{P}} \downarrow 0$ is achieved if and only if $\gamma \downarrow 0$. Then it follows from (21) that

$$0 = \lim_{\gamma \downarrow 0} \frac{R(\mathbb{P}_{\varepsilon}) - R(\mathbb{P}) - \gamma \mathbb{E}_{\mathbb{Q}-\mathbb{P}}[\ell_{\mathbb{P}}(\xi)]}{\gamma \|\mathbb{Q} - \mathbb{P}\|_{\mathcal{P}}} = \frac{1}{\|\mathbb{Q} - \mathbb{P}\|_{\mathcal{P}}} \Big(\mathrm{d}R(\mathbb{P}; \mathbb{Q}) - \mathbb{E}_{\mathbb{Q}-\mathbb{P}}[\ell_{\mathbb{P}}(\xi)] \Big).$$

Since $\mathbb{Q} \neq \mathbb{P}$, we conclude that the $\mathbb{Q}$-directional derivative exists and $\mathrm{d}R(\mathbb{P}; \mathbb{Q}) = \mathbb{E}_{\mathbb{Q}-\mathbb{P}}[\ell_{\mathbb{P}}(\xi)]$. Moreover, this equality holds even if $\mathbb{Q} = \mathbb{P}$ since $\mathrm{d}R(\mathbb{P}; \mathbb{P}) = 0 = \mathbb{E}_{\mathbb{P}-\mathbb{P}}[\ell_{\mathbb{P}}(\xi)]$.

The notion of F-derivative relies heavily on the underlying metric structure on $\mathcal{P}$, whereas, the notion (Definition 3.1) of G-derivatives is independent of it. To compare, for a G-derivative to exist along a given direction, it is only required for the limit in (11) to exist. Whereas, the existence of an F-derivative requires that the limit in (11) is achieved uniformly over all possible directions.

Challenges with F-derivatives. Even though the notion of Frechet-smoothness (22) is a natural extension of canonical smoothness (37) in an infinite dimensional setting; working with F-derivatives is potentially challenging due to the following reasons:

(i) **Existence.** It is a stronger requirement that the limit in (22) converges uniformly (w.r.t $\|\cdot\|_{\mathcal{P}}$) in all directions, which often implies that an F-derivative might not even exist.

(ii) **Finite representability** An F-derivative is a function $\ell_{\mathbb{P}} : \Xi \to \mathbb{R}$ which is an infinite dimensional object, so apriori, it is not clear as to whether it can be characterized in terms of a few finite dimensional quantities.

(iii) **Norm consistency.** Most importantly, it is often very difficult to establish the smoothness condition (22) of the F-derivatives in a specific norm. To elaborate further, we know that the FW algorithm in a finite-dimensional convex problem converges sub-linearly, if the feasible set is bounded and the objective function is smooth. Since all norms on finite dimensional vector spaces are equivalent, the choices of norms for establishing the smoothness of the objective function, and the boundedness of the feasibility set are irrelevant (even though this could potentially give rise to dimensionally dependent constants). However, since no such equivalence exists between norms on an infinite dimensional space, it becomes then necessary that the risk measure is Frechet-smooth w.r.t. the same norm under which the ambiguity set is bounded, which is a much stronger condition to expect.

The distinction between the notion of smoothness in Definition 3.6 and the Frechet-smoothness (22) is further illustrated by considering the example of regular risk measures. As discussed, it suffices to have Lipschitz continuous gradients of the function r and the set $\{\mathbb{E}_{\mathbb{P}}[L(\xi)] : \mathbb{P} \in \mathcal{P}\}$ bounded for the regular risk to be G-smooth. On the contrary, being F-smooth requires the mapping $\mathbb{P}' \longmapsto \mathbb{E}_{\mathbb{P}'}[\ell_{\mathbb{P}}(\xi)]$ to be continuous, which is a stronger requirement for the F-derivative existence.

4. The Frank-Wolfe Algorithm

Given the notion of G-derivative as in Definition 3.1, the Frank-Wolfe (FW) algorithm for (5) is an iterative procedure that involves solving the optimization problem

$$\sup_{\mathbb{Q} \in \mathcal{P}} \mathrm{d}R(\mathbb{P}; \mathbb{Q}), \tag{23}$$

for a given $\mathbb{P} \in \mathcal{P}$, at each iteration. The FW-problem (23) is linear if and only if the G-derivative $\mathrm{d}R(\mathbb{P}; \mathbb{Q})$ is affine in $\mathbb{Q}$, for every $\mathbb{P} \in \mathcal{P}$, translating the problem into the linear DRO class in (3). This is indeed the case for many interesting risk measures as seen in Example 3.3. Moreover, in all finite dimensional optimization problems, the corresponding FW-problem (23) is always linear, which need not be the case for a generic infinite-dimensional optimization problem like (5).

Implementation of the FW algorithm only requires a well-defined notion of the G-derivative, and as seen in Lemma 3.2, such objects can be computed by means of a few finite-dimensional quantities in several problems of interest. Moreover, even though the *FW-problem* (23) is infinite-dimensional in nature, it admits tractable finite-dimensional convex reformulations for many relevant applications similar to linear DROs [20].This is a compelling reason to investigate FW methods for optimization problems over probability distributions, particularly in the setting of (5) with a non-linear risk measure.

4.1. **Frank-Wolfe oracle**

The FW-oracle is a set-valued mapping $\mathcal{F} : \mathcal{P} \times [0, +\infty) \to 2^{\mathcal{P}}$ defined as

$$\mathcal{F}(\mathbb{P}, \gamma) := \left\{ \mathbb{Q}' \in \mathcal{P} \text{ such that } \sup_{\mathbb{Q} \in \mathcal{P}} \mathrm{d}R(\mathbb{P}; \mathbb{Q}) \leqslant \gamma \delta C + \mathrm{d}R(\mathbb{P}; \mathbb{Q}') \right\}. \tag{24}$$

Given the current iterate $\mathbb{P} \in \mathcal{P}$, the FW algorithm involves solving the FW-problem (23) to obtain its approximate solution $\mathbb{Q}' \in \mathcal{P}$, then the current iterate $\mathbb{P}$ is updated by moving it towards $\mathbb{Q}'$ as shown in Figure 3.

Additive accuracy of the oracle. The parameter $\delta > 0$ is an arbitrary positive number signifying the accuracy of the FW oracle, and C is the smoothness constant of the risk measure R as in Definition 3.6. It must be observed that in an iterative scheme to solve (5), it is typical that the step-size sequence (γ_k) is monotonically decreasing and converges to 0. Therefore, it is also required that the FW-oracle solves the sub-problems (23) up to a greater precision as the iterations progress.

The Frank-wolfe-gap. We refer to the quantity $\sup_{\mathbb{Q} \in \mathcal{P}} \mathrm{d}R(\mathbb{P}; \mathbb{Q})$ as the *Frank-Wolfe(FW)-gap* at $\mathbb{P}$; and it is crucial in defining aposteriori stopping criteria for the FW-algorithm. Along with the distribution $\mathbb{Q}'$, we shall assume that the FW-oracle also provides access to the quantity $\mathrm{d}R(\mathbb{P}; \mathbb{Q}')$, which is an approximate value of the FW-gap at $\mathbb{P}$.
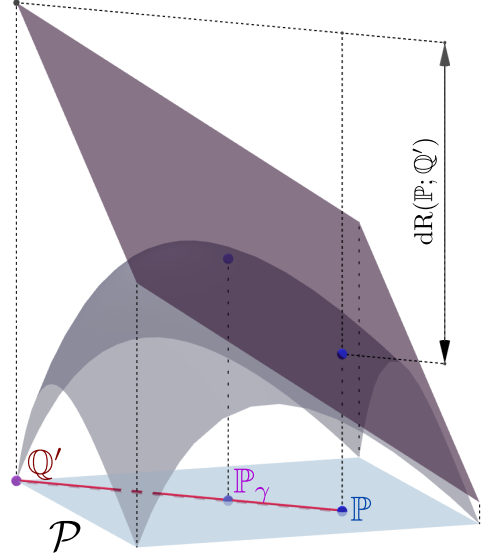

FIGURE 3. FW-oracle.

Lemma 4.1 (FW-one-step-bounds)**.** *Consider* (5) *with a risk measure that is C-smooth in the sense of Definition 3.6 for some $C \geqslant 0$, and let R^* be its optimal value. Let $\mathcal{F}$ be the corresponding FW-oracle as in* (24) *with an arbitrary accuracy parameter $\delta \geqslant 0$. For any $\mathbb{P} \in \mathcal{P}$, $\gamma \in [0, 1]$, and $\mathbb{Q}' \in \mathcal{F}(\mathbb{P}, \gamma)$, let $\mathbb{P}_\gamma = \mathbb{P} + \gamma(\mathbb{Q}' - \mathbb{P})$ be the one-step-ahead FW update from $\mathbb{P}$ with a step-size of γ. Then we have*

$$R^* - R(\mathbb{P}) \leqslant \sup_{\mathbb{Q} \in \mathcal{P}} \mathrm{d}R(\mathbb{P}; \mathbb{Q}) \leqslant \frac{1}{\gamma} \Big(R(\mathbb{P}_\gamma) - R(\mathbb{P}) \Big) + \gamma C(1 + \delta). \tag{25}$$

Rearranging the inequality (25), the one-step improvement in sub-optimality is seen to be

$$R^* - R(\mathbb{P}_\gamma) \leqslant \gamma^2 C(1 + \delta) + (1 - \gamma)\big(R^* - R(\mathbb{P})\big). \tag{26}$$

4.2. **Frank-Wolfe algorithm and convergence guarantees**

The FW-algorithm seeks to solve the optimization problem (5) by iteratively solving the FW-problem (23) using a FW-oracle as in (24). More precisely, given a step size sequence $(\gamma_k)_k \subset [0, 1]$ and a distribution $\mathbb{P}_0 \in \mathcal{P}$, the FW-algorithm generates a sequence of distributions $(\mathbb{P}_k)_k$ such that

$$\text{FW-algorithm:} \quad \mathbb{P}_{k+1} = \mathbb{P}_k + \gamma_k(\mathbb{Q}_k - \mathbb{P}_k), \quad \text{where} \quad \mathbb{Q}_k \in \mathcal{F}(\mathbb{P}_k, \gamma_k). \tag{27}$$

It must be observed that the implementation of the FW algorithm (27) does not depend on the choice of a norm on the ambiguity set $\mathcal{P}$. Therefore, it is desirable to have a *norm-independent* analysis of the FW algorithm. To this end, we take inspiration from [17], which has a similar analysis for the convergence of FW-algorithm for finite-dimensional problems by working with the notion of *curvature co-efficient* instead of the canonical smoothness (18) used in [12]. It turns out that the notion of smoothness defined in Definition 3.6 is also amenable to a similar analysis of the FW-algorithm; with the advantage that (17) is more in-line as a generalized notion of smoothness from (18) and (19).

Proposition 4.2 (Apriori bounds). *Consider* (5) *with a risk measure that is C-smooth in the sense of Definition 3.6 for some $C \in [0, +\infty)$, and let R^* be its optimal value. For $k = 1, 2, \ldots,$ let $(\mathbb{P}_k)_k$ be the sequence of distributions obtained from the FW-algorithm* (27) *with a step size sequence $\gamma_k = \frac{2}{k+2}$ and some $\mathbb{P}_0 \in \mathcal{P}$. Then*

$$R^* - R(\mathbb{P}_k) \ \leqslant \ \frac{4C}{k+2}(1+\delta) \quad \text{for all } k \geqslant 1. \tag{28}$$

4.3. **FW-gap based termination and aposteriori bounds**

Suppose that we know some $C' \geqslant 0$ that satisfies the smoothness condition (17). Then for any given $\varepsilon > 0$, if the FW algorithm is run for $K \geqslant \left\lceil \frac{4C'(1+\delta)}{\varepsilon} \right\rceil - 2$ iterations with the step-size sequence $\gamma_k = \frac{2}{k+2}$, then the last iterate $\mathbb{P}_K$ is guaranteed to be ε sub-optimal in objective value. Thus, in principle, it suffices to only know some upper bound $C' \geqslant C$ for the smoothness constant. However, if finding C exactly is challenging, and the known upper bound C' is not tight; the theoretically guaranteed number of iterations required for ε-sub-optimality may not be practical. In such a setting, it turns out that the optimal value of the FW-problem (23) called the *FW-gap* provides a good measure to define aposteriori stopping criteria. Moreover, we shall also see later for NDRO problems that terminating the FW algorithm when the FW-gap is small provides worst-case performance bounds in the context of DRO problems. We follow the analysis of [17, 8] by considering the FW algorithm under two regimes of step-size sequence to obtain provable upper bounds on the FW-gap towards later iterations.

Since the oracle employed to solve the linear minimization sub-problems at each iteration is only accurate to some specified precision, the actual value of the FW-gap: $\sup_{\mathbb{Q} \in \mathcal{P}} \mathrm{d}R(\mathbb{P}; \mathbb{Q})$ is never known exactly. However, at each iteration k, the FW-oracle does provide its approximate estimate $\mathrm{d}R(\mathbb{P}; \mathbb{Q}_k)$ which satisfies the inequality $\sup_{\mathbb{Q} \in \mathcal{P}} \mathrm{d}R(\mathbb{P}; \mathbb{Q}) \leqslant \delta\gamma_k C + \mathrm{d}R(\mathbb{P}; \mathbb{Q}_k)$. So, for a given value of $\varepsilon > 0$ we terminate the FW procedure by examining the quantity $\mathrm{d}R(\mathbb{P}; \mathbb{Q}_k)$ such that the desired FW-gap inequality: $\sup_{\mathbb{Q} \in \mathcal{P}} \mathrm{d}R(\mathbb{P}; \mathbb{Q}) \leqslant \varepsilon$, is satisfied.

Proposition 4.3 (Aposteriori bounds). *Consider* (5) *with a risk measure that is C-smooth in the sense of Definition 3.6 for some $C \geqslant 0$, and let R^* be its optimal value. Let $K \geqslant 1$ and let $(\mathbb{P}_k)_k$ be the sequence obtained from the FW-algorithm* (27) *using a diminishing step-size $\gamma_k := \frac{2}{k+2}$ for $k = 0, 1, \ldots, K-1$, and then a constant step-size $\gamma_k = \frac{2}{K+2}$ for $k = K, K+1, \ldots, 2K+1$. Finally, let $g_k = \mathrm{d}R(\mathbb{P}; \mathbb{Q}_k)$ for $k = 1, 2, \ldots, 2K+1$ be the sequence of approximate FW-gaps. There exists*

$\widehat{k} \in \{K, K+1, \ldots, 2K+1\}$ *such that*

$$R^* - R(\mathbb{P}_{\widehat{k}}) \;\leqslant\; \sup_{\mathbb{Q} \in \mathcal{P}} \mathrm{d}R(\mathbb{P}_{\widehat{k}}; \mathbb{Q}) \;\leqslant\; \frac{2C(2+3\delta)}{K+2}, \tag{29}$$

and every such $\widehat{k}$ is recognised by verifying the inequality: $g_{\widehat{k}} \leqslant \frac{4C}{K+2}(1+\delta)$.

Remark 4.4 (Explicit error bounds). *Consider the setting of Proposition 4.3 with $K = K(\varepsilon) := \left\lceil \frac{2C}{\varepsilon}(2+3\delta) \right\rceil - 2$, for any given $\varepsilon > 0$. Then there exists $\widehat{k} \in \{K(\varepsilon), K(\varepsilon)+1, \ldots, 2K(\varepsilon)+1\}$ such that*

$$\sup_{\mathbb{Q} \in \mathcal{P}} \mathrm{d}R(\mathbb{P}_{\widehat{k}}; \mathbb{Q}) \leqslant \varepsilon,$$

and every such $\widehat{k}$ is recognised by verifying the inequality $g_{\widehat{k}} \leqslant \varepsilon \frac{2+2\delta}{2+3\delta}$.

Remark 4.5 (Two-regimes step-size). *The two-regimes for γ_k in Proposition 4.3 turns out to be crucial to obtain provable guarantees that the FW-gap is bounded above in the later iterations. Even though such certificates are of independent interest in their own right, having such upper-bounds is also essential in the context of DRO problems. An upper bound on the FW-gap at iteration k ensures that the performance of the decision x_k for the worst-case distribution is not "too-bad".*

5. Non-linear Distributionally Robust Optimization

Let us recall that a generic DRO problem is formulated as the min-max problem

$$F^* := \inf_{x \in \mathcal{X}} \sup_{\mathbb{P} \in \mathcal{P}} F(x, \mathbb{P}), \tag{30}$$

where $\mathcal{X} \subset \mathbb{R}^n$ is a closed convex set, denoting the set of feasible decisions, and $\mathcal{P}$ denotes a given ambiguity set of distributions. A DRO problem is said to be *feasible* if $F^* < +\infty$, which happens if and only if there exists some $x \in \mathcal{X}$ such that $\sup_{\mathbb{P} \in \mathcal{P}} F(x, \mathbb{P}) < +\infty$. Our objective is to develop a framework to solve a generic DRO problem (30). Particularly, with emphasis on the case when $F(x, \mathbb{P})$ is non-linear in $\mathbb{P}$ for every $x \in \mathcal{X}$; in which case, we refer to (30) as a *non-linear distributionally robust optimization* (NDRO) problem.

Example 5.1 (Regular NDRO examples). Similar to the examples of RR measures given in Example 2.2, we present two analogous examples of the function F that are considered in the context of the NDRO problem:

(a) *Variance*: The variance $V(x, \mathbb{P})$ associated with the distribution $\mathbb{P}$ and a decision x, is

$$V(x, \mathbb{P}) := x^\top \big(\Sigma_{\mathbb{P}} - \mu_{\mathbb{P}} \mu_{\mathbb{P}}^\top \big) x, \quad \text{where } \Sigma_{\mathbb{P}} := \mathbb{E}_{\mathbb{P}}[\xi \xi^\top] \text{ and } \mu_{\mathbb{P}} = \mathbb{E}_{\mathbb{P}}[\xi]. \tag{31}$$

(b) *Entropic risk measure*: Another example is the entropic risk measure $\mathcal{E}(x, \mathbb{P})$ of a multi variate distribution $\mathbb{P}$ and decision x. If $\mathbb{P}$ is a distribution with marginals $\mathbb{P}_j$ for $j = 1, 2, \ldots, n$, (i.e., the j-th component ξ_j is $\mathbb{P}_j$ distributed). For a given set of *risk-aversion parameters* $(\theta_j)_{j=1}^n \subset (0, +\infty)$, the associated $\mathcal{E}$-measure is defined as

$$\mathcal{E}(x, \mathbb{P}) := \sum_{j=1}^n \frac{1}{\theta_j} \log \Big(\mathbb{E}_{\mathbb{P}_j}[e^{-\theta_j x_j \xi_j}] \Big). \tag{32}$$

NDRO Dual problem. Associated with the DRO problem (30) is its dual problem:

$$\text{Dual problem:} \quad F_* := \sup_{\mathbb{P} \in \mathcal{P}} \inf_{x \in \mathcal{X}} F(x, \mathbb{P}). \tag{33}$$

In general, we have *weak-duality* $F_* \leqslant F^*$ relating the optimal values of the primal problem (30) and its dual (33). If $F_* = F^*$ specifically, we say that *strong-duality* holds between (30) and (33). Moreover, suppose the DRO problem (30) and its dual (33) admit the solutions x^* and $\mathbb{P}^*$, i.e.,

$$x^* \in \operatorname*{argmin}_{x \in \mathcal{X}} \sup_{\mathbb{P} \in \mathcal{P}} F(x, \mathbb{P}) \quad \text{and} \quad \mathbb{P}^* \in \operatorname*{argmax}_{\mathbb{P} \in \mathcal{P}} \inf_{x \in \mathcal{X}} F(x, \mathbb{P}),$$

Then, the pair $(x^*, \mathbb{P}^*)$ is said to be a *saddle point* solution to the problems (30) and (33), which is also characterised by the inequalities

$$\max_{\mathbb{P} \in \mathcal{P}} F(x^*, \mathbb{P}) \;=\; F(x^*, \mathbb{P}^*) \;=\; \min_{x \in \mathcal{X}} F(x, \mathbb{P}^*).$$

Existence of a saddle point is sufficient for strong-duality to hold, however, it is not necessary. Therefore, whenever strong-duality holds, we consider the slightly relaxed notion of an ε-sub-optimal saddle points as a solution concept for (30).

Definition 5.2 (ε-saddle point). Given $\varepsilon \geqslant 0$, a pair $(x_\varepsilon, \mathbb{P}_\varepsilon) \in \mathcal{X} \times \mathcal{P}$ is an ε-saddle point of the DRO problem (30), if it satisfies

$$-\varepsilon + \sup_{\mathbb{P} \in \mathcal{P}} F(x_\varepsilon, \mathbb{P}) \;\leqslant\; F(x_\varepsilon, \mathbb{P}_\varepsilon) \;\leqslant\; \varepsilon + \inf_{x \in \mathcal{X}} F(x, \mathbb{P}_\varepsilon). \tag{34}$$

It is worth noting that if $(x_\varepsilon, \mathbb{P}_\varepsilon)$ is an ε-saddle point, then we have the inequalities:

$$\sup_{\mathbb{Q} \in \mathcal{P}} F(x_\varepsilon, \mathbb{Q}) \;\leqslant\; 2\varepsilon + \inf_{y \in \mathcal{X}} F(y, \mathbb{P}_\varepsilon) \;\leqslant\; 2\varepsilon + \inf_{y \in \mathcal{X}} \sup_{\mathbb{Q} \in \mathcal{P}} F(y, \mathbb{Q}) \qquad = 2\varepsilon + F^*, \quad \text{and}$$

$$\inf_{y \in \mathcal{X}} F(y, \mathbb{P}_\varepsilon) \;\geqslant\; -2\varepsilon + \sup_{\mathbb{Q} \in \mathcal{P}} F(x_\varepsilon, \mathbb{Q}) \;\geqslant\; -2\varepsilon + \sup_{\mathbb{Q} \in \mathcal{P}} \inf_{y \in \mathcal{X}} F(y, \mathbb{Q}) \quad = -2\varepsilon + F_*.$$

In other words, if $(x_\varepsilon, \mathbb{P}_\varepsilon)$ is an ε-saddle point, then both x_ε and $\mathbb{P}_\varepsilon$ are at most 2ε-sub-optimal to the DRO problem (30) and its dual (33) respectively. Consequently, the decision x_ε is guaranteed to be at most 2ε worse from the best decision that could have been made in a DRO framework.

Solving the minimization over x in the dual-problem (33) results in a maximization problem (potentially non-linear) over the distributions

$$\sup_{\mathbb{P} \in \mathcal{P}} R(\mathbb{P}), \quad \text{where } R(\mathbb{P}) := \inf_{x \in \mathcal{X}} F(x, \mathbb{P}). \tag{35}$$

Denoting $x(\mathbb{P}) := \operatorname{argmin}_{x \in \mathcal{X}} F(x, \mathbb{P})$, for every $\mathbb{P} \in \mathcal{P}$, the proposed method to compute an ε-saddle point of the DRO problem generates a sequence $(x_k, \mathbb{P}_k)_k$, where $x_k \in x(\mathbb{P}_k)$ for each k, and the sequence of distributions $(\mathbb{P}_k)_k$ is obtained by applying the FW algorithm to the maximization problem (35). If a pair $(x', \mathbb{P}')$ satisfies $\mathbb{P}' \in \operatorname{argmax}_{\mathbb{P} \in \mathcal{P}} R(\mathbb{P})$ and $x' \in x(\mathbb{P}')$, then it is not guaranteed that $x' \in \operatorname{argmin}_{y \in \mathcal{X}} \sup_{\mathbb{P} \in \mathcal{P}} F(y, \mathbb{P})$ unless $x(\mathbb{P}')$ is unique. It so turns out that the regularity assumptions on F and $\mathcal{P}$, required for the algorithm convergence, also ensure uniqueness.

5.1. **NDRO: derivatives and smoothness**

Our proposed method to solve the NDRO problem (30) by applying Frank-Wolfe algorithm to (35) requires that the risk measure R therein has well defined G-derivatives that are also smooth in the sense of Definition 3.6. This is not guaranteed a priori. In the following, we impose some regularity assumptions on the function F that guarantee the required smoothness of the risk measure R.

Assumption 5.3 (NDRO smoothness)**.** *Let $F_x(\,\cdot\,) \coloneqq F(x,\,\cdot\,)$, for every $x \in \mathcal{X}$. We assume that there exists positive constants α, C_1, C_2 such that*

(i) Strong convexity*: The function $F(x, \mathbb{P})$ is α-strongly convex in x w.r.t. the norm $\|\cdot\|$, uniformly over all $\mathbb{P} \in \mathcal{P}$. In other words, we have the inequality*

$$\frac{\alpha}{2} \|x - y\|^2 \;\leqslant\; F(y, \mathbb{P}) - F(x, \mathbb{P}) - \langle \nabla_1 F(x, \mathbb{P}), \, y - x \rangle, \qquad \forall\, x, y \in \mathcal{X}, \quad \forall\, \mathbb{P}, \mathbb{Q} \in \mathcal{P}.^2 \quad (36)$$

(ii) Smoothness*: The function $F_x(\mathbb{P})$ is directionally differentiable on $\mathcal{P}$, and C_2-smooth in the sense of Definition 3.6, uniformly over $x \in \mathcal{X}$.*

(iii) Continuous derivatives*: The G-derivative $\mathrm{d}F_x(\mathbb{P}; \mathbb{Q})$ is C_1-Lipschitz continuous in x uniformly over $\mathbb{P}, \mathbb{Q} \in \mathcal{P}$. In other words, the following inequality holds:*

$$\mathrm{d}F_x(\mathbb{P}; \mathbb{Q}) - \mathrm{d}F_y(\mathbb{P}; \mathbb{Q}) \;\leqslant\; C_1 \|x - y\|, \qquad \forall\, x, y \in \mathcal{X} \quad \forall\, \mathbb{P}, \mathbb{Q} \in \mathcal{P}. \quad (37)$$

Remark 5.4 (Choice of norm on $\mathcal{X}$)**.** *It must be noted that the norm $\|\cdot\|$ considered in the strong convexity assumption (36) and the continuity assumption (37) is identical. Considering an identical norm is not restrictive since all norms on $\mathcal{X}$ are equivalent. However, using such equivalence often makes the resulting constants α, C_1 to be dimension dependent (of $\mathcal{X}$). We emphasise here that the smoothness constant given in Lemma 5.5, requires that the constants α, and C_1 that satisfy conditions (36) and (37), to satisfy with an identical norm.*

For now, we shall assume that these conditions for the abstract problem (30) are satisfied. However, for specific problems like minimum variance portfolio optimization, we shall determine verifiable conditions whenever possible so that Assumption 5.3 is indeed satisfied, (see Lemma 6.2). Informally, condition (37) together with the smoothness condition is akin to saying that the directional derivatives of $F_x(\mathbb{P})$ are similar to being "Lipschitz continuous" with respect to both x and $\mathbb{P}$. Moreover, the strong convexity assumption implies that the mapping $\mathbb{P} \longmapsto x(\mathbb{P})$ is also similar to being "Lipschitz continuous". These two consequences together, yield the smoothness of R. The following lemma formally establishes this deduction in a norm independent (in $\mathbb{P}$) analysis.

Lemma 5.5 (NDRO-derivative properties)**.** *Let the function F satisfy Assumption 5.3 with constants $\alpha > 0$, and $C_1, C_2 \geqslant 0$, then the following holds for the risk measure R as defined in (35)*

(i) Danskin's theorem*: The risk measure R is directionally differentiable on $\mathcal{P}$, and for any $\mathbb{P}, \mathbb{Q} \in \mathcal{P}$ its $\mathbb{Q}$-directional derivative $\mathrm{d}R(\mathbb{P}; \mathbb{Q})$ at $\mathbb{P}$, is given by*

$$\mathrm{d}R(\mathbb{P}; \mathbb{Q}) = \mathrm{d}F_{x(\mathbb{P})}(\mathbb{P}; \mathbb{Q}). \quad (38)$$

(ii) Smoothness*: The risk measure R is C-smooth in the sense of Definition 3.6 for*

$$C = C_2 + \frac{C_1}{2\alpha}\Big(C_1 + \sqrt{C_1^2 + 4\alpha C_2}\Big).$$

---

[2]$\nabla_1 F(x, \mathbb{P}) \coloneqq (\partial F / \partial x)(x, \mathbb{P})$ denotes the partial derivative of F w.r.t. x evaluated at $(x, \mathbb{P})$.

5.2. Frank-Wolfe based algorithm for the NDRO problem.

Let $\mathbb{P}_k$ for $k = 0, 1, 2, \ldots$, be the sequence of iterates generated by the FW-algorithm (27) for R as defined in (35). Assume that the FW oracle is δ-accurate, for some arbitrary $\delta > 0$. So, in order to produce a solution x_ε to the NDRO problem (30) for a given $\varepsilon > 0$; one must now decide the number of iterations $K(\varepsilon)$, the step-size sequence γ_k for $k = 0, 1, \ldots, K(\varepsilon)$, and the stopping criteria for the FW-algorithm. We now present the main result of the article that provides a solution to the NDRO problem (30).

Algorithm 1: Max-min algorithm for NDRO problem (30)

Input: A distribution $\mathbb{P} \in \mathcal{P}$, positive real numbers ε and C, and access to a FW-oracle corresponding to the function F.

Output: The final decision variable x_ε and worst case distribution $\mathbb{P}_\varepsilon$.

1 **Initialisation**: $\mathbb{P}_0 := \mathbb{P}$

first stepsize regime

2 **for**: $k = 0, 1, 2, \ldots, K(\varepsilon) := \left\lceil \frac{2C}{\varepsilon}(2 + 3\delta) \right\rceil - 2,$

    stepsize: $\gamma_k := \frac{2}{k+2}$

    find $(x_k, \mathbb{Q}_k) \in \mathcal{X} \times \mathcal{P}$ such that

$$F\big(x_k, \mathbb{P}_k\big) = \min_{x \in \mathcal{X}} F\big(x, \mathbb{P}_k\big) \quad \text{and} \quad \sup_{\mathbb{Q} \in \mathcal{P}} \mathrm{d}F_{x_k}(\mathbb{P}_k; \mathbb{Q}) \leqslant \delta\gamma_k C + \mathrm{d}F_{x_k}(\mathbb{P}_k; \mathbb{Q}_k) \tag{39}$$

    $g_k = \mathrm{d}F_{x_k}(\mathbb{P}_k; \mathbb{Q}_k)$

    $\mathbb{P}_{k+1} := \mathbb{P}_k + \gamma_k(\mathbb{Q}_k - \mathbb{P}_k)$

3 **end for**

second stepsize regime

4 **for** $k = K(\varepsilon) + 1, \ldots, 2K(\varepsilon) + 1$

    stepsize: $\gamma_k := \frac{2}{K(\varepsilon)+2}$

    if $g_k > \varepsilon \frac{2+2\delta}{2+3\delta}$, **do**

      find $(x_k, \mathbb{Q}_k) \in \mathcal{X} \times \mathcal{P}$ such that

$$F\big(x_k, \mathbb{P}_k\big) = \min_{x \in \mathcal{X}} F\big(x, \mathbb{P}_k\big) \quad \text{and} \quad \sup_{\mathbb{Q} \in \mathcal{P}} \mathrm{d}F_{x_k}(\mathbb{P}_k; \mathbb{Q}) \leqslant \delta\gamma_k C + \mathrm{d}F_{x_k}(\mathbb{P}_k; \mathbb{Q}_k) \tag{40}$$

      $g_k = \mathrm{d}F_{x_k}(\mathbb{P}_k; \mathbb{Q}_k)$

      $\mathbb{P}_{k+1} := \mathbb{P}_k + \gamma_k(\mathbb{Q}_k - \mathbb{P}_k)$

    else

      Output: $x_\varepsilon := x_k$ and $\mathbb{P}_\varepsilon := \mathbb{P}_k$ and **end for**.

5 **end for**

Theorem 5.6 (NDRO solution). *Consider the Non-linear DRO problem* (30) *under the setting of Assumption 5.3. Then the following holds*

(i) Strong duality:

$$F^* = \min_{x \in \mathcal{X}} \sup_{\mathbb{P} \in \mathcal{P}} F\big(x, \mathbb{P}\big) = \sup_{\mathbb{P} \in \mathcal{P}} \min_{x \in \mathcal{X}} F\big(x, \mathbb{P}\big) = F_* \tag{41}$$

(ii) *Given any $\varepsilon > 0$, the pair $(x_\varepsilon, \mathbb{P}_\varepsilon)$ computed from Algorithm 1 is an ε-saddle point in the sense of Definition 5.2.*

5.3. **Slower convergence without strong-convexity**

In this case, we assume that the function F satisfies conditions (ii) and (iii) of Assumption 5.3. However, it may not be necessarily strongly convex in x. For example, the variance risk measure is strongly convex if and only if the smallest eigenvalue of the matrix $(\Sigma_\mathbb{P} - \mu_\mathbb{P}\mu_\mathbb{P}^\top)$ is bounded away from 0 uniformly over $\mathbb{P} \in \mathcal{P}$, which might not be the case. Even in such a setting, we desire to develop methods that compute an ε-saddle point of F using the setup of Algorithm 1 for any $\varepsilon > 0$. We take inspiration from the smoothing techniques in the optimization literature [28] for smoothing a non-smooth convex function and devise similar techniques that work with a suitable strongly-convex approximation F_ε, of F, and still use Algorithm 1 to compute an ε-saddle point of F. To this end, we assume that the set $\mathcal{X}$ is also bounded in addition to being closed, thus, compact. Many common examples of $\mathcal{X}$ like the simplex Δ^n, satisfy the compactness assumption.

For any given an $\varepsilon > 0$, $x \in \mathcal{X}$, and $\mathbb{P} \in \mathcal{P}$, let $F_\varepsilon(x, \mathbb{P}) := F(x, \mathbb{P}) + (\varepsilon/B_x^2) \|x\|^2$. We propose to solve the following min-max problem in place of (30)

$$\min_{x \in \mathcal{X}} \sup_{\mathbb{P} \in \mathcal{P}} F_\varepsilon(x, \mathbb{P}), \tag{42}$$

where $B_x := \max_{x \in X} \|x\|$. It is apparent at once that $F_\varepsilon(x, \mathbb{P})$ is $(2\varepsilon/B_x^2)$-strongly convex in x, uniformly over $\mathbb{P} \in \mathcal{P}$, and consequently satisfies all the conditions in Assumption 5.3. Thus, employing Algorithm 1 with F_ε computes a pair $(x', \mathbb{P}')$ that satisfies the inequalities

$$-\varepsilon + \sup_{\mathbb{Q} \in \mathcal{P}} F_\varepsilon(x', \mathbb{Q}) \;\leqslant\; F_\varepsilon(x', \mathbb{P}') \;\leqslant\; \varepsilon + \inf_{y \in \mathcal{X}} F_\varepsilon(y, \mathbb{P}'). \tag{43}$$

It turns out that such a pair $(x', \mathbb{P}')$ is an ε-saddle point of F as well. This is easily seen by observing that on the one hand, we have

$$\begin{aligned}(\varepsilon/B_x^2) \|x'\|^2 + \sup_{\mathbb{Q} \in \mathcal{P}} F(x', \mathbb{Q}) \;&=\; \sup_{\mathbb{Q} \in \mathcal{P}} F_\varepsilon(x', \mathbb{Q}) \;\leqslant\; \varepsilon + F_\varepsilon(x', \mathbb{P}') \quad \text{(from (43))}\\ &=\; \varepsilon + (\varepsilon/B_x^2) \|x'\|^2 + F(x', \mathbb{P}').\end{aligned}$$

Thus, we immediately get $\sup_{\mathbb{Q} \in \mathcal{P}} F(x', \mathbb{Q}) \leqslant \varepsilon + F(x', \mathbb{P}')$. On the other hand, since the minimization over x is solved exactly in Algorithm 1, we have

$$\begin{aligned}F_\varepsilon(x', \mathbb{P}') \;&=\; \min_{y \in \mathcal{X}} F_\varepsilon(y, \mathbb{P}') \;=\; \min_{y \in \mathcal{X}} F(y, \mathbb{P}') + (\varepsilon/B_x^2) \|y\|^2\\ &\leqslant\; \varepsilon + \min_{y \in \mathcal{X}} F(y, \mathbb{P}') \quad \text{since } \|y\| \leqslant B_x \text{ for all } y \in \mathcal{X} .\end{aligned}$$

Thus, it is apparent that the pair $(x', \mathbb{P}')$ is an ε-saddle point of the function F as well.

Corollary 5.7 (Slower convergence). *Consider function $F(x, \mathbb{P})$ that is not necessarily strongly convex in x. Then for any desired precision $\varepsilon > 0$, an ε-saddle point of F can be computed by applying Algorithm 1 with $K(\varepsilon) = \left\lceil \frac{2C_\varepsilon}{\varepsilon}(2 + 3\delta) \right\rceil - 2$, to the strongly convex approximate function F^ε , where $C_\varepsilon = C_2 + \frac{C_1 B_x^2}{4\varepsilon}\left(C_1 + \sqrt{C_1^2 + \frac{8\varepsilon}{B_x^2} C_2}\right)$.*

Since the strong convexity parameter of F_ε itself depends on ε, we conclude from Lemma 5.5 that the risk measure $R_\varepsilon(\mathbb{P}) := \min_{y \in \mathcal{X}} F_\varepsilon(y, \mathbb{P})$, also has an ε dependent smoothness constant $C_\varepsilon = C_2 + \frac{C_1 B_x^2}{4\varepsilon}\left(C_1 + \sqrt{C_1^2 + \frac{8\varepsilon}{B_x^2} C_2}\right)$. Now, Algorithm 1 terminates in $O(K(\varepsilon))$ iterations, where $K(\varepsilon) = O(C_\varepsilon/\varepsilon)$. Since $C_\varepsilon = O(1/\varepsilon)$, it is easily seen that for non-strongly convex functions F, applying Algorithm 1 to its regularized strongly-convex approximation F_ε takes $O(1/\varepsilon^2)$ iterations to compute an ε-saddle point of F. To compare, recall that for a strongly-convex function F, Algorithm 1 takes $O(1/\varepsilon)$ iterations to compute an ε-saddle point. This trade-off between speed of convergence and precision in the approximation is a typical occurrence in standard smoothing techniques as well.

6. Minimum variance portfolio optimization problem

One of the textbook examples of NDRO is the minimum variance portfolio selection problem [23, 31, 3, 30]. To this end, let $\widehat{\xi}_i$: $i = 1, 2, \ldots, N$ be i.i.d. samples drawn from some unknown underlying distribution $\mathbb{P}_o$, and let $\widehat{\mathbb{P}}$ be the nominal distribution that is uniformly distributed over the samples $(\widehat{\xi}_i)_i$. Let $\mathcal{P} = \mathcal{W}_m(\widehat{\mathbb{P}}, \rho) := \{\mathbb{Q} : \mathsf{WD}_m(\widehat{\mathbb{P}}, \mathbb{Q}) \leqslant \rho\}$, where WD_m is the m-th order Wasserstein distance between the distributions, induced by the transportation cost coming from a norm $\|\cdot\|$ on Ξ. More specifically, we have

$$
\mathsf{WD}_m(\mathbb{P}, \mathbb{Q}) := \begin{cases} \sup_\pi & \left(\mathbb{E}_\pi[\|u - v\|^m]\right)^{1/m} \quad \text{where } (u, v) \text{ are } \pi\text{-jointly distributed} \\ \text{subject to} & \mathbb{P}(u) = \int_v \pi(u, v) dv \quad \text{and} \quad \mathbb{Q}(v) = \int_u \pi(u, v) du. \end{cases}
$$

Associated with the distribution $\mathbb{P}$, let $\Sigma_\mathbb{P} := \mathbb{E}_\mathbb{P}[\xi \xi^\top]$ and $\mu_\mathbb{P} := \mathbb{E}_\mathbb{P}[\xi]$ denote its first and second moments respectively. Finally, let $\mathcal{X} \subset \Delta^n$ denote the set of feasible actions, then we seek to investigate the min-sup problem:

$$
\min_{x \in \mathcal{X}} \sup_{\mathbb{P} \in \mathcal{W}_m(\widehat{\mathbb{P}}, \rho)} V(x, \mathbb{P}) := x^\top \left(\Sigma_\mathbb{P} - \mu_\mathbb{P} \mu_\mathbb{P}^\top\right) x. \tag{44}
$$

We shall study in detail the notion of the directional derivative, smoothness, and the resulting FW-oracle with its tractable formulations for (44) under different conditions.

6.1. Derivatives and convex regularity.

Using the short-hand notation $V_x(\cdot) := V(x, \cdot)$ for every x, we recall that the mapping $\mathbb{P} \longmapsto V_x(\mathbb{P})$ is an RR measure in the sense of Definition 2.1 with maps L and r given by

$$
L(\xi) = (\xi \xi^\top, \xi) \quad \text{and} \quad r(\Sigma, \mu) = x^\top \Sigma x - \left(x^\top \mu\right)^2,
$$

Thus, in view of Lemma 3.2, the directional derivatives of V_x are given by

$$
dV_x(\mathbb{P}; \mathbb{Q}) = \mathbb{E}_{\mathbb{Q}-\mathbb{P}}\left[x^\top \left((\xi - 2\mu_\mathbb{P}) \xi^\top\right) x\right]. \quad \text{for every } \mathbb{P}, \mathbb{Q} \in \mathcal{W}_m(\widehat{\mathbb{P}}, \rho), \tag{45}
$$

Assumption 6.1 (Min-variance moments bound). *There exists constants $B_\Sigma, B_\mu \geqslant 0$ such that*

$$
\|\Sigma_\mathbb{Q} - \Sigma_\mathbb{P}\|_o \leqslant B_\Sigma \quad \text{and} \quad \|\mu_\mathbb{Q} - \mu_\mathbb{P}\|_2 \leqslant B_\mu \quad \text{for all} \quad \mathbb{P}, \mathbb{Q} \in \mathcal{W}_m(\widehat{\mathbb{P}}, \rho), \tag{46}
$$

where $\Sigma_\mathbb{P} = \mathbb{E}_\mathbb{P}[\xi \xi^\top]$ and $\mu_\mathbb{P} = \mathbb{E}_\mathbb{P}[\xi]$.

Lemma 6.2 (Min-variance regularity conditions)**.** *Consider the minimum variance portfolio optimization problem* (44), *and suppose that Assumption* 6.1 *holds with constants* B_Σ, B_μ. *Let* $V_x(\,\cdot\,) :=$ $V(x,\,\cdot\,)$ *for every* $x \in \mathcal{X}$, *then the following assertions hold*

(i) Smoothness: *The risk measure* V_x *is* $\left(2B_\mu^2\right)$*-smooth in the sense of Definition* 3.6, *uniformly over* $x \in \mathcal{X}$.

(ii) Continuous partial derivatives: *For* $C_1 = 2\left(B_\Sigma + 2\left(B_\mu + \|\widehat{\mu}\|_2\right)^2 + B_\mu^2\right)$, *the directional derivatives* $\mathrm{d}V_x(\mathbb{P};\mathbb{Q})$, *satisfy*

$$\mathrm{d}V_x(\mathbb{P};\mathbb{Q}) - \mathrm{d}V_y(\mathbb{Q};\mathbb{P}) \leqslant C_1 \left\|x - y\right\|_2, \qquad \forall\, x,y \in \mathcal{X} \quad \forall\, \mathbb{P},\mathbb{Q} \in \mathcal{P}.$$

If Assumption 6.1 holds, then Lemma 6.2 ensures that the *smoothness* and *continuity* conditions of Assumption 5.3 are satisfied for the minimum variance portfolio optimization problem (44). However, the uniform strong-convexity assumption need not hold in general. For instance, if the ambiguity set is large (i.e., if ρ is large), then all the data points ξ_i can be perturbed within their respective ρ-neighborhoods such that the variance corresponding to the perturbed points is rank deficient. Thus, the strong-convexity condition in Assumption 5.3 is not satisfied at the distribution supported over the perturbed points. In such cases, adding an explicit strongly-convex regularizer to apply Algorithm 1 (as discussed in Section 5.3) not only provides theoretical guarantees for convergence but also improves the speed of the algorithm in our observation; see the numerical simulations concerning Figure 5.

6.2. **The Frank-Wolfe algorithm**

The FW-Oracle. Let us consider the FW-problem arising in the NDRO problem of minimum variance portfolio selection (44),

$$\sup_{\mathbb{Q} \in \mathcal{W}_m(\widehat{\mathbb{P}},\rho)} \mathrm{d}V_x(\mathbb{P};\mathbb{Q}). \tag{47}$$

Since the directional derivatives $\mathrm{d}V_x(\mathbb{P};\mathbb{Q})$ are affine in $\mathbb{Q}$ for every pair $(x,\mathbb{P})$, the corresponding FW-problem is linear. However, the existence and characterization results of the solution to the FW-problem (47) change depending on the interplay of (i) the Wasserstein distance type m, (ii) the transportation cost $\|\cdot\|$, and (iii) the support set Ξ (unbounded, or compact). Therefore, the specific settings for which the corresponding FW-oracle is easy to describe are discussed later in the section. To this end, we simplify (47), and study its dual problem which is used later in the characterization of a solution to (47).

It is a simple algebraic exercise to verify that

$$\begin{aligned}
\mathrm{d}V_x(\mathbb{P};\mathbb{Q}) &= \mathbb{E}_{\mathbb{Q}-\mathbb{P}}\left[x^\top\left((\xi - 2\mu_\mathbb{P})\xi^\top\right)x\right] \\
&= \mathbb{E}_\mathbb{Q}\left[(\xi - \mu_\mathbb{P})^\top\left(xx^\top\right)(\xi - \mu_\mathbb{P})\right] - x^\top\left(\Sigma_\mathbb{P} - \mu_\mathbb{P}\mu_\mathbb{P}^\top\right)x.
\end{aligned}$$

Since the distribution $\mathbb{P}$ is constant in (47), eliminating terms that only depend on $\mathbb{P}$ does not affect the set of maximizers. Thus, we have

$$\operatorname*{argmax}_{\mathbb{Q} \in \mathcal{W}_m(\widehat{\mathbb{P}},\rho)} \mathrm{d}V_x(\mathbb{P};\mathbb{Q}) = \operatorname*{argmax}_{\mathbb{Q} \in \mathcal{W}_m(\widehat{\mathbb{P}},\rho)} \mathbb{E}_\mathbb{Q}\left[(\xi - \mu_\mathbb{P})^\top\left(xx^\top\right)(\xi - \mu_\mathbb{P})\right]. \tag{48}$$

We now focus on a generic version of the linear worst-case distribution problem:

$$\sup_{\mathbb{Q}\in\mathcal{W}_m(\widehat{\mathbb{P}},\rho)} \mathbb{E}_{\mathbb{Q}}[(\xi-v)^\top(xx^\top)(\xi-v)], \tag{49}$$

for any given $x, v \in \mathbb{R}^n$. We known that [20, Theorem 7] (49) admits an equivalent dual formulation given by

$$\inf_{\eta\geqslant 0} \eta\rho^m + \frac{1}{N}\sum_{i=1}^N \begin{cases} \sup_{q_i} & (q_i+\xi_i-v)^\top(xx^\top)(q_i+\xi_i-v) - \eta\|q_i\|^m \\ \text{subject to} & q_i+\xi_i \in \Xi \quad \text{for all} \quad i=1,2,\ldots,N. \end{cases} \tag{50}$$

For $\eta \geqslant 0$ and $\xi \in \Xi$, let

$$q(\eta,\xi,v) := \begin{cases} \underset{q}{\text{argmax}} & \left((x^\top q) + x^\top(\xi-v)^\top\right)^2 - \eta\|q\|^m \\ \text{subject to} & q+\xi \in \Xi, \end{cases} \tag{51}$$

whenever an optimal solution exists. If the dual problem (50) admits an optimal solution η_x, and $q(\eta_x,\xi_i,v)$ exists for each $i=1,2,\ldots,N$; then for any collection $q_i' \in q(\eta_x,\xi_i,v)$, $i=1,2,\ldots,N$, the discrete distribution $\mathbb{Q}_x(\xi) := \frac{1}{N}\sum_{i=1}^N \delta(\xi-\xi_i-q_i')$ is a maximizer for the linear worst-case distribution problem (49).

Tractable one step FW-update. We observe that in the minimum variance problem (44), the only information needed pertaining to a distribution $\mathbb{Q} \in \mathcal{W}_m(\widehat{\mathbb{P}},\rho)$ is its first and second order moments $\mu_{\mathbb{Q}} = \mathbb{E}_{\mathbb{Q}}[\xi]$, $\Sigma_{\mathbb{Q}} = \mathbb{E}_{\mathbb{Q}}[\xi\xi^\top]$ respectively. Thus, in order to solve the DRO problem (44) via the FW-algorithm (Algorithm 1), it is apparent that it suffices to track the evolution of these moments rather than the entire distribution (which gets increasingly difficult with the iterations). To this end, a single FW-update (39), in Algorithm 1, for the minimum-variance problem (44) can be summarised in terms of the finite dimensional quantities:

$$\begin{cases} \text{Solve for } x &: \quad x_k \in \underset{x\in\mathcal{X}}{\text{argmin}}\ x^\top(\Sigma_k - \mu_k\mu_k^\top)x \\[2mm] \text{FW-oracle} &: \quad \begin{cases} \eta_k = \eta_{x_k} \quad \text{and} \quad q_i \in q(\eta_k,\xi_i,\mu_k) \quad \text{for all} \quad i=1,2,\ldots,N, \\ \mu_k' = \dfrac{1}{N}\sum_{i=1}^N(\xi_i+q_i) \quad \text{and} \quad \Sigma_k' = \dfrac{1}{N}\sum_{i=1}^N(\xi_i+q_i)(\xi_i+q_i)^\top, \end{cases} \\[4mm] \text{FW-update} &: \quad \mu_{k+1} = \mu_k + \gamma_k(\mu_k'-\mu_k), \qquad \Sigma_{k+1} = \Sigma_k + \gamma_k(\Sigma_k'-\Sigma_k). \end{cases} \tag{52}$$

We now focus on specific settings of (44) for which the FW-oracle is easily characterised.

Case 1: unconstrained support. We consider the setting: $\Xi = \mathbb{R}^n$, $\mathcal{X} \in \Delta^n$ being any compact subset, and the transportation cost $\|\cdot\|$ to define the Wasserstein ambiguity set (and let $\|\cdot\|_*$ be the associated dual norm).

Lemma 6.3. *Consider the maximization problem (51) for $\Xi = \mathbb{R}^n$, and $\eta \geqslant 0$, $x,v,\xi \in \mathbb{R}^n$, and let J denote its optimal value. Then the following assertions hold*

(i) *(Unbounded)– $J = +\infty$, and no optimal solution exists for (51) in the following cases: $(m < 2)$, $(m > 2, \eta = 0)$, $(m = 2, \eta < \|x\|_*^2)$, and $(m = 2, \eta = \|x\|_*^2, x^\top(\xi-v) \neq 0)$.*

(ii) *(Bounded)– $J < +\infty$, and the optimal solution $q(\eta,\xi,v)$ exists for (51) in all other cases, and in particular,*

(a) *$(m > 2, \eta > 0)$, both $J < +\infty$, and $q(\eta,\xi,v)$ exists.*

(b) $(m = 2, \eta > \|x\|_*^2)$, then we have

$$J = \frac{\eta \left| x^\top (\xi - v) \right|^2}{\eta - \|x\|_*^2}, \quad q(\eta, \xi, v) = \frac{\|x\|_* \left(x^\top (\xi - v) \right)}{\eta - \|x\|_*^2} \bar{q}_x \ for \ \bar{q}_x \in \underset{\|\bar{q}\| \leqslant 1}{\operatorname{argmin}} \ x^\top \bar{q}.$$

(c) $(m = 2, \eta = \|x\|_*^2, x^\top (\xi - v) = 0)$, then $J = 0$ and $q(\eta, \xi, v) = \{s\bar{q}_x : s \geqslant 0\}$.

Lemma 6.4 (Min-variance FW-oracle: unconstrained support). *Consider the linear worst-case-distribution problem* (49) *and its dual* (50)*, under the setting,* $\Xi = \mathbb{R}^n$*,* $m = 2$*, and any* $x \in \mathcal{X}$*. The solutions* $(\mathbb{Q}_x, \eta_x)$ *to* (49) *and its dual* (50) *are given by*

$$\begin{cases} \eta_x = \|x\|_*^2 + \dfrac{\|x\|_*}{\rho} \sqrt{\dfrac{1}{N} \sum_{i=1}^{N} |x^\top (\xi_i - v)|^2}, \\[4mm] \mathbb{Q}_x(\xi) = \dfrac{1}{N} \sum_{i=1}^{N} \delta\big(\xi - (\xi_i + q_i')\big) \quad for \ any \ q_i' \in q(\eta_x, \xi_i, v), \ i = 1, 2, \dots, N. \end{cases} \tag{53}$$

It turns out that for the special case of $m = 2$, the problem (44) with unconstrained support reduces to a simple empirical risk minimization problem. This was already discovered in [3] specifically for the case when $\mathcal{X} = \mathcal{X}(\bar{\alpha}) := \{x \in \Delta^n : \mathbb{E}_\mathbb{Q}[x^\top \xi] \geqslant \bar{\alpha}, \text{ for all } \mathbb{Q} \in \mathcal{P}\}$, for any $\bar{\alpha}$. We generalize it slightly by showing that a similar conclusion holds for any compact set $\mathcal{X} \subset \Delta^n$. Moreover, we provide an alternate proof completely based on first order optimality conditions and the FW-oracle, for the min-max problem (44). We emphasize here that since it is shown explicitly that the min-max problem (44) reduces to an empirical minimization problem, a solution can be computed without the need to run the FW-algorithm (52) iteratively. More importantly, this also alleviates the need to ensure that the convex regularity assumptions hold for this special case of (44).

Proposition 6.5 (Min-variance saddle point). *Consider the minimum variance portfolio optimization problem* (44) *in the setting of* $m = 2$*, and* $\Xi = \mathbb{R}^n$*. Let*

$$x^* \in \underset{x \in \mathcal{X}}{\operatorname{argmin}} \sqrt{\left\langle x, (\widehat{\Sigma} - \widehat{\mu}\widehat{\mu}^\top)x \right\rangle} + \rho \|x\|_*, \tag{54}$$

then there exists some $\bar{q}_{x^*} \in \arg \max_{\|\bar{q}\| \leqslant 1} \langle x^*, \bar{q} \rangle$ *such that for*

$$\mathbb{P}^* := \frac{1}{N} \sum_{i=1}^{N} \delta\big(\xi - (\xi_i + q_i^*)\big), \quad where \quad q_i^* = \frac{\rho \langle x^*, \xi_i - \widehat{\mu} \rangle}{\sqrt{\left\langle x^*, (\widehat{\Sigma} - \widehat{\mu}\widehat{\mu}^\top)x^* \right\rangle}} \bar{q}_{x^*} \quad \forall i \leqslant N, \tag{55}$$

the pair $(x^*, \mathbb{P}^*)$ *is a saddle point to* (44).

Case 2: ellipsoidal support. We consider the setting: $\Xi = \mathcal{E}_M := \{\xi : \langle \xi, M\xi \rangle \leqslant 1\}$ for some $M \succeq 0$, $m = 2$, and the underlying transportation cost defining the Wasserstein distance to be $\| \cdot \| = \| \cdot \|_2$. Under this setting, the FW problem in (39) involves maximizing a quadratic function subject to convex ellipsoidal constraints. Even if this problem is non-convex in general, it admits a tractable reformulation as an SDP via the celebrated *S-procedure* [5, Appendix B], [40].

Lemma 6.6 (Min-variance FW-oracle: ellipsoidal support)**.** *Consider the minimum variance portfolio optimization problem* (44) *under the setting of the ellipsoidal support. The corresponding dual problem* (50) *of the linear worst-case-distribution problem* (49) *is equivalent to the SDP*

$$
\begin{cases}
\min\limits_{\eta\in\mathbb{R},\ \lambda,\theta\in\mathbb{R}^N} \quad \eta\rho^2 - \dfrac{1}{N}\sum_{i=1}^{N}\theta_i \\[2mm]
\text{subject to} \quad
\begin{cases}
\eta \geqslant 0,\ \ and\ \lambda_i \geqslant 0\ \forall i \leqslant N, \\[2mm]
\begin{bmatrix}
\eta\mathbb{I}_n - xx^\top + \lambda_i M & (xx^\top)v - \eta\xi_i \\
(xx^\top)v^\top - \eta\xi_i^\top & \eta\|\xi_i\|_2^2 - (x^\top v)^2 - \lambda_i - \theta_i
\end{bmatrix} \succeq 0,\ \ \forall i \leqslant N.
\end{cases}
\end{cases}
\tag{56}
$$

Moreover, for any solution $(\eta^*, \lambda^*, \theta^*)$, *to the SDP* (56), *the pair* $(\mathbb{Q}_x, \eta_x)$ *given by*

$$
\eta_x = \eta^*, \quad and \quad \mathbb{Q}_x(\xi) = \frac{1}{N}\sum_{i=1}^{N}\delta\Big(\xi - \big(\eta^*\mathbb{I}_n - xx^\top + \lambda_i^* M\big)^{-1}\big(\eta^*\xi_i - xx^\top v\big)\Big),
\tag{57}
$$

is a solution to (49) *and its dual* (50) *respectively.*

6.3. **Numerical results**

We validate the convergence attributes of Algorithm 1 for the NDRO problem of the minimum variance portfolio selection (44) with ellipsoidal support (i.e., $\Xi = \mathcal{E}_M \subset \mathbb{R}^n$). The positive definite matrix M that characterizes the support $\mathcal{E}_M$, is generated randomly and it is ensured to be reasonably well conditioned. The samples ξ_i, for $i = 1, 2, \ldots, N = 2*n$, are drawn randomly from $\mathcal{E}_M$, and the distribution $\widehat{\mathbb{P}}$ is uniformly distributed over the samples. Then we solve

$$
V^* := \min_{x\in\Delta^n} \sup_{\mathbb{P}\in\mathcal{W}_2(\widehat{\mathbb{P}},\rho)} V(x, \mathbb{P}),
\tag{58}
$$

for $n = 25$ and various values of ρ (the radius of the ambiguity set). The proposed FW based method (Algorithm 1 with the simplified FW-update (52)) is implemented in MATLAB, wherein the SDP (56) corresponding to the FW-oracle and the minimization problem: $\min_{x\in\Delta^n} V(x, \mathbb{P}_k)$ are solved using the cvx solver.

First, Figure 4 shows the convergence of Algorithm 1 for $K(\varepsilon) = 75$ iterations, and for three distinct values of $\rho = 0.1, 0.5$, and 1. The upper plots in the figure show the evolution of the primal and dual functions: $\min_{x\in\Delta^n} V(x, \mathbb{P}_k)$ and $\sup_{\mathbb{P}\in\mathcal{W}_2(\widehat{\mathbb{P}},\rho)} V(x_k, \mathbb{P})$ respectively, w.r.t. the iteration k of the algorithm. Since Algorithm 1 explicitly solves the minimization over x in each iteration (see (39)), the primal function is readily available. However, this is not the case with the dual function which needs to be computed independently at each iteration for the current iterate x_k. We compute it by running the FW-update (52) for several iterations ($\sim K(\varepsilon)$) independently with x_k held fixed, this amounts to applying the FW-algorithm (27) for the risk measure $V_{x_k}(\cdot)$. In the bottom plots, we show the evolution w.r.t. the iteration k, of the two sub-optimality metrics:

$$
\begin{cases}
\text{primal sub-optimality} \quad : \quad V^* - \min\limits_{x\in\Delta^n} V(x, \mathbb{P}_k) \quad \text{and} \\[3mm]
\text{duality gap} \quad : \quad \sup\limits_{\mathbb{P}\in\mathcal{W}_2(\widehat{\mathbb{P}},\rho)} V(x_k, \mathbb{P}) - \min\limits_{x\in\Delta^n} V(x, \mathbb{P}_k).
\end{cases}
$$

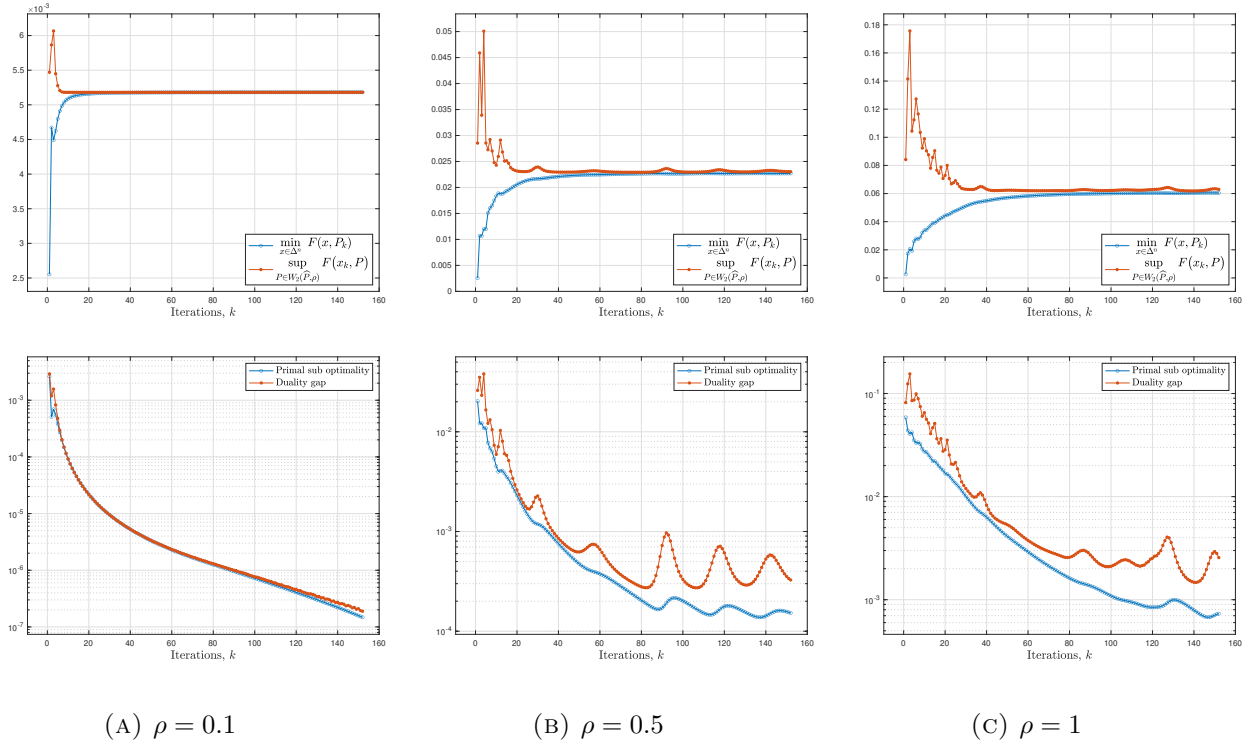(A) $\rho = 0.1$ (B) $\rho = 0.5$ (C) $\rho = 1$

FIGURE 4. Convergence plots for Algorithm 1 with $K(\varepsilon) = 75$, applied to (58).

With explicit regularization of x. We observe in Figure 4 that, as the value of ρ increases, the convergence of the algorithm becomes less smooth, and also slower. Perhaps, this can be attributed to the fact that the strong-convexity assumption in Assumption 5.3 fails. This is so because, a larger value of ρ allows the samples $(\xi_i)_i$ to be perturbed in such a way that the variance matrix of the perturbed points is rank deficient. Thus, the function $V(x, \mathbb{P})$ is not strongly convex in x for $\mathbb{P}$ corresponding to the perturbed points.

To remedy this, as suggested in Section 5.3, we compute an approximate saddle point of (58) by applying the FW-algorithm to the explicitly regularized min-sup problem

$$\min_{x \in \Delta^n} \sup_{\mathbb{P} \in \mathcal{W}_2(\widehat{\mathbb{P}}, \rho)} \frac{\alpha}{2} \|x\|_2^2 + V(x, \mathbb{P}). \tag{59}$$

Since $\|x\|_2 \leqslant \|x\|_1 = 1$ for all $x \in \Delta^n$, we know that an ε-saddle point of (58) can be computed for any $\varepsilon > 0$ by applying Algorithm 1 to (59) with $\alpha = 2\varepsilon$. We emphasize that both the primal function: $\min_{x \in \Delta^n} V(x, \mathbb{P}_k)$ and the dual function: $\sup_{\mathbb{P} \in \mathcal{W}_2(\widehat{\mathbb{P}}, \rho)} V(x_k, \mathbb{P})$ are not accessible in the implementation of Algorithm 1 for the explicitly regularized objective function. Thus, these quantities are computed separately at each iteration of the algorithm to record the primal sub-optimality and the duality gap.

In Figure 5, we show the convergence plots of Algorithm 1 when applied to the explicitly regularized problem (59). We consider the same data set from Figure 4 but with a slightly larger value of $\rho = 1.5$, making the conditioning of the problem even worse than that for $\rho = 1$.

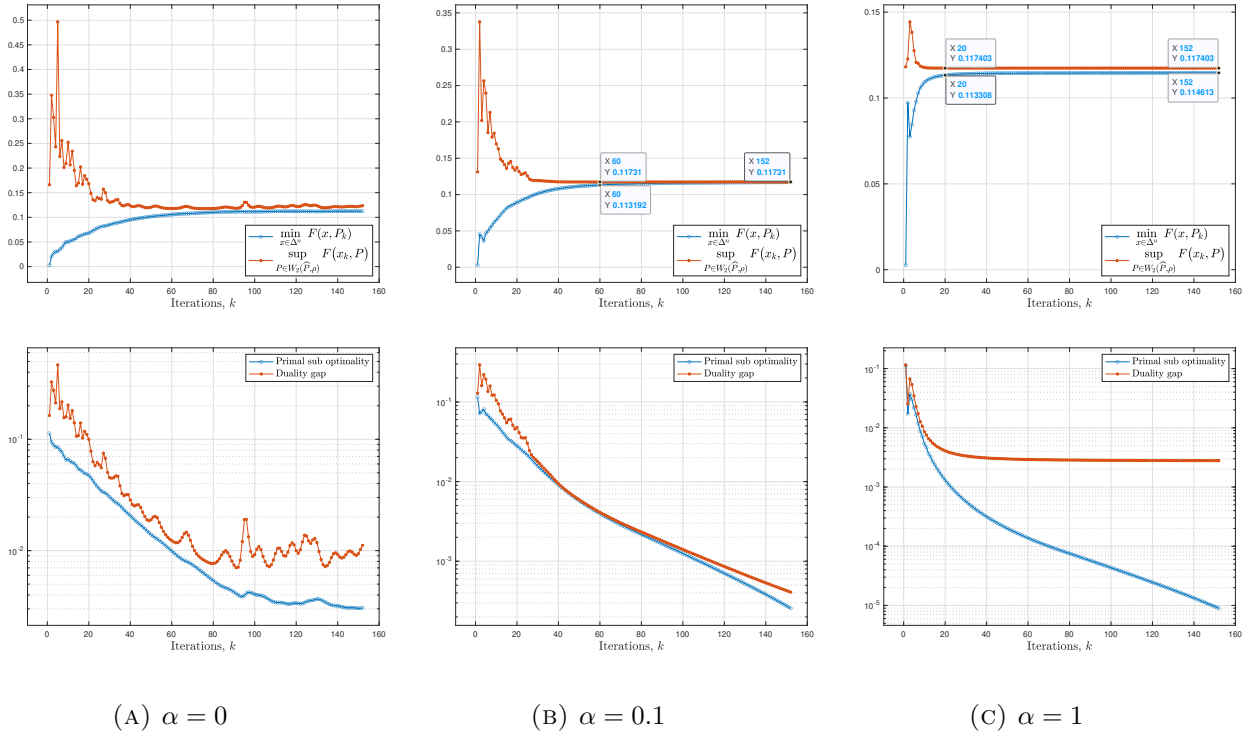(A) $\alpha = 0$ (B) $\alpha = 0.1$ (C) $\alpha = 1$

FIGURE 5. Convergence plots for Algorithm 1 with $K(\varepsilon) = 75$, applied to the explicitly regularized approximation (59).

We show the convergence plots of the algorithm for three different values of $\alpha = 0, 0.1, 1$. Clearly, for $\alpha = 0$, the problem is not regularized, and the convergence is bad (even worse than that for $\rho = 1$ from Figure 4c). Then the effect of explicit regularization and how it improves the regularity and convergence can be clearly seen in Figure 5. For $\alpha = 0.1$ and 1, we see that the solution computed is an ε-saddle point for the respective value of ε (i.e., $\varepsilon = 0.05$ for $\alpha = 0.1$ and $\varepsilon = 0.5$ for $\alpha = 1$). In addition, Figures 5b and 5c highlight the trade-off between the speed of convergence and worst-case performance metrics. For the larger value of $\alpha = 1$, due to better regularity, the algorithm is faster which can be seen in Figure 5c that both the curves of primal (blue) and dual (red) functions achieve within one percent of their final value in less than 20 iterations. Whereas, with $\alpha = 0.1$, even after 60 iterations, the primal function (blue curve) is only within 3.5 percent of its final value. However, this improved speed comes at a cost as clearly evident from Figures 5b and 5c that the worst-case cost for $\alpha = 0.1$ (0.11731) is smaller than that for $\alpha = 1$ (0.117403). Furthermore, even though the solution computed with both $\alpha = 0, 1$ and 1, is within the respective sub-optimality levels, the duality gap however, actually goes to zero for $\alpha = 0.1$, which is not the case for $\alpha = 1$.

7. TECHNICAL PROOFS

7.1. **Proofs related to derivatives and smoothness.**

In this part we cover the technical proofs of the theoretical statements in Section 3.

Proof of Lemma 3.2. From the definition 3.1, we see that

$$
\begin{aligned}
\mathrm{d}R(\mathbb{P}; \mathbb{Q}) &= \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \left(R(\mathbb{P} + \varepsilon(\mathbb{Q} - \mathbb{P})) - R(\mathbb{P}) \right) \\
&= \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \left(r\left(\mathbb{E}_{\mathbb{P}}[L(\xi)] + \varepsilon\left(\mathbb{E}_{\mathbb{Q}-\mathbb{P}}[L(\xi)] \right) \right) - r\left(\mathbb{E}_{\mathbb{P}}[\xi] \right) \right) \\
&= \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \left(r\left(\mathbb{E}_{\mathbb{P}}[\xi] \right) + \varepsilon \left\langle \nabla r(\mathbb{E}_{\mathbb{P}}[L(\xi)]), \ \mathbb{E}_{\mathbb{Q}-\mathbb{P}}[L(\xi)] \right\rangle + o(\varepsilon^2) - r\left(\mathbb{E}_{\mathbb{P}}[\xi] \right) \right) \\
&= \left\langle \nabla r(\mathbb{E}_{\mathbb{P}}[L(\xi)]), \ \mathbb{E}_{\mathbb{Q}-\mathbb{P}}[L(\xi)] \right\rangle \\
&= \mathbb{E}_{\mathbb{Q}-\mathbb{P}} \left[\left\langle \nabla r(\mathbb{E}_{\mathbb{P}}[L(\xi)]), \ L(\xi) \right\rangle \right].
\end{aligned}
$$

□

Proof of Proposition 3.4. From Definition 3.1, we have

$$
\begin{aligned}
\mathrm{d}R(\mathbb{P}; \mathbb{P}_\gamma) &= \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \Big(R(\mathbb{P} + \varepsilon(\mathbb{P}_\gamma - \mathbb{P})) - R(\mathbb{P}) \Big) = \lim_{\varepsilon \downarrow 0} \frac{\gamma}{(\varepsilon\gamma)} \Big(R(\mathbb{P} + \varepsilon\gamma(\mathbb{Q} - \mathbb{P})) - R(\mathbb{P}) \Big) \\
&= \gamma \lim_{\varepsilon' \downarrow 0} \frac{1}{\varepsilon'} \Big(R(\mathbb{P} + \varepsilon'(\mathbb{Q} - \mathbb{P})) - R(\mathbb{P}) \Big) \\
&= \gamma \mathrm{d}R(\mathbb{P}; \mathbb{Q}),
\end{aligned}
$$

which establishes (16a). To establish (16b), we see that

$$
\begin{aligned}
\mathrm{d}R(\mathbb{P}; \mathbb{Q}) &= \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \Big(R(\mathbb{P} + \varepsilon(\mathbb{Q} - \mathbb{P})) - R(\mathbb{P}) \Big) \geqslant \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \Big(R(\mathbb{P}) + \varepsilon(R(\mathbb{Q}) - R(\mathbb{P})) - R(\mathbb{P}) \Big) \\
&\geqslant R(\mathbb{Q}) - R(\mathbb{P}),
\end{aligned}
$$

where the first inequality is due to concavity of $\mathbb{P} \longmapsto R(\mathbb{P})$. This completes the proof.   □

Proof of Lemma 3.7. For any $\mathbb{P}, \mathbb{Q} \in \mathcal{P}$ with $\mathbb{P}_\gamma = \mathbb{P} + \gamma(\mathbb{Q} - \mathbb{P})$ and $\gamma \in [0, 1]$ we see that

$$
\begin{aligned}
\mathrm{d}R&(\mathbb{P}_\gamma; \mathbb{P}) + \mathrm{d}R(\mathbb{P}; \mathbb{P}_\gamma) \\
&= \left\langle \nabla r(\mathbb{E}_{\mathbb{P}}[L(\xi)]), \ \mathbb{E}_{\mathbb{P}_\gamma}[L(\xi)] - \mathbb{E}_{\mathbb{P}}[L(\xi)] \right\rangle + \left\langle \nabla r(\mathbb{E}_{\mathbb{P}_\gamma}[L(\xi)]), \ \mathbb{E}_{\mathbb{P}}[L(\xi)] - \mathbb{E}_{\mathbb{P}_\gamma}[L(\xi)] \right\rangle \\
&= \left\langle \nabla r(\mathbb{E}_{\mathbb{P}}[L(\xi)]) - \nabla r(\mathbb{E}_{\mathbb{P}_\gamma}[L(\xi)]), \ \mathbb{E}_{\mathbb{P}_\gamma}[L(\xi)] - \mathbb{E}_{\mathbb{P}}[L(\xi)] \right\rangle \\
&\leqslant \beta \left\| \mathbb{E}_{\mathbb{P}_\gamma}[L(\xi)] - \mathbb{E}_{\mathbb{P}}[L(\xi)] \right\|^2 \leqslant \gamma^2 \beta \left\| \mathbb{E}_{\mathbb{Q}}[L(\xi)] - \mathbb{E}_{\mathbb{P}}[L(\xi)] \right\|^2 \leqslant \gamma^2 (\beta d^2).
\end{aligned}
$$

□

7.2. **Proofs for FW-algorithm.**

In this part we cover the technical proofs of the theoretical statements in Section 4.

Proof for Lemma 4.1. For $\mathbb{P}$ and $\mathbb{P}_\gamma$, we see that the inequalities hold.

$$
\begin{aligned}
R(\mathbb{P}) - R(\mathbb{P}_\gamma) &\leqslant \mathrm{d}R(\mathbb{P}_\gamma; \mathbb{P}) && \text{from (16b)} \\
&= \mathrm{d}R(\mathbb{P}_\gamma; \mathbb{P}) + \mathrm{d}R(\mathbb{P}; \mathbb{P}_\gamma) - \mathrm{d}R(\mathbb{P}; \mathbb{P}_\gamma) \\
&\leqslant \gamma^2 C - \gamma \mathrm{d}R(\mathbb{P}; \mathbb{Q}') && \text{from smoothness (17), and (16a)} \\
&\leqslant \gamma^2 C(1 + \delta) - \gamma \, \mathrm{d}R(\mathbb{P}; \mathbb{Q}) && \text{from (24).}
\end{aligned}
$$

Rearranging terms and using (16b), we get

$$R(\mathbb{Q}) - R(\mathbb{P}) \quad \leqslant \quad \mathrm{d}R\big(\mathbb{P};\mathbb{Q}\big) \quad \leqslant \quad \frac{1}{\gamma}\Big(R(\mathbb{P}_\gamma) - R(\mathbb{P})\Big) + \gamma C(1 + \delta),$$

wherein, by considering the supremum over $\mathbb{Q} \in \mathcal{P}$, the inequalities in (25) follow at once. $\qquad\square$

Proof for Proposition 4.2. Let $(\mathbb{P}_k)_k$ be the sequence of Frank-Wolfe iterates updated as in (27). From the one-step inequality (26), we have

$$R^* - R\big(\mathbb{P}_{k+1}\big) \leqslant \gamma_k^2 C(1 + \delta) + (1 - \gamma_k)\big(R^* - R(\mathbb{P}_k)\big) \quad \text{for every } k \geqslant 0. \tag{60}$$

Using (60), we shall prove the primal convergence of FW-algorithm following the classical approach [17, 8] by the method of induction.

Base case: For $k = 1$, since $\mathbb{P}_1 = \mathbb{P}_0 + \gamma_0(\mathbb{Q}_0 - \mathbb{P}_0)$ we deduce from (60) using $\gamma = \gamma_0 = 1$ that

$$R^* - R(\mathbb{P}_1) \leqslant C(1 + \delta) \; < \; \frac{4C}{3}(1 + \delta).$$

Thus, (28) holds for $k = 1$.

Induction step: Now, assuming that (28) holds for some $k \geqslant 1$, we deduce from (60) that

$$\begin{aligned}
R^* - R\big(\mathbb{P}_{k+1}\big) \;&\leqslant\; \gamma_k^2 C(1 + \delta) + (1 - \gamma_k)\big(R^* - R(\mathbb{P}_k)\big) \\
&\leqslant\; \frac{4C(1 + \delta)}{(k + 2)^2} + \frac{k}{k + 2}\frac{4C(1 + \delta)}{k + 2} = \frac{4C(1 + \delta)}{k + 2}\frac{1 + k}{k + 2} \\
&\leqslant\; \frac{4C(1 + \delta)}{k + 3}, \quad \text{since } (k + 2)^2 \leqslant (k + 1)(k + 3).
\end{aligned}$$

Thus, we conclude that (28) indeed holds for every $k \geqslant 1$. The proof is now complete. $\qquad\square$

Proof for Proposition 4.3. Following the ideas of [8, 17], we first prove by the method of contradiction that there exists $\widehat{k}$ such that

$$\sup_{\mathbb{Q} \in \mathcal{P}} \mathrm{d}R\big(\mathbb{P}_{\widehat{k}};\mathbb{Q}\big) \leqslant \frac{4C(1 + \delta)}{K + 2} \quad \text{for some } \widehat{k} \in \{K, \ldots, 2K + 1\}. \tag{61}$$

Suppose that $\sup_{\mathbb{Q} \in \mathcal{P}} \mathrm{d}R(\mathbb{P}_k;\mathbb{Q}) > \frac{4C(1 + \delta)}{K + 2}$ for every $k \in \{K, \ldots, 2K + 1\}$, then by considering $\gamma = \frac{2}{K + 2}$ in (25) and rearranging terms, we see that for every $k \in \{K, \ldots, 2K + 1\}$, we have the inequality

$$\begin{aligned}
R\big(\mathbb{P}_k\big) - R\big(\mathbb{P}_{k+1}\big) \;&\leqslant\; \frac{4}{(K + 2)^2}C(1 + \delta) \; - \; \frac{2}{K + 2}\sup_{\mathbb{Q} \in \mathcal{P}} \mathrm{d}R\big(\mathbb{P}_k;\mathbb{Q}\big) \\
&<\; \frac{4}{(K + 2)^2}C(1 + \delta) \; - \; \frac{8}{(K + 2)^2}C(1 + \delta) \\
&<\; -\frac{4}{K + 2}C(1 + \delta).
\end{aligned}$$

Summing these inequalities for $k = K, \ldots, 2K+1$, we get $R\big(\mathbb{P}_K\big) - R\big(\mathbb{P}_{2K+2}\big) \; < \; -\big(4/K+2\big)C(1+\delta)$. Combining this with the sub-optimality of $R\big(\mathbb{P}_K\big)$ from (28) finally gives

$$R^* - R\big(\mathbb{P}_{2K+2}\big) \; < \; -\frac{4}{K + 2}C(1 + \delta) \; + \; \frac{4}{K + 2}C(1 + \delta) \; < \; 0,$$

which is a contradiction. Therefore, there must exist some $\widehat{k} \in \{K, \ldots, 2K + 1\}$, such that (61) holds.

For every $\widehat{k}$ such that (61) holds, we also necessarily have

$$g_{\widehat{k}} = \mathrm{d}R\big(\mathbb{P}_{\widehat{k}}; \mathbb{Q}_{\widehat{k}}\big) \leqslant \sup_{\mathbb{Q} \in \mathcal{P}} \ \mathrm{d}R\big(\mathbb{P}_{\widehat{k}}; \mathbb{Q}\big) \leqslant \frac{4C(1 + \delta)}{K + 2}. \tag{62}$$

We emphasise that at each iteration k, the FW-oracle solves the FW-problems upto an additive accuracy of $\delta\gamma_{\widehat{k}}C$. Therefore, we do not have access to the exact value of $\sup_{\mathbb{Q} \in \mathcal{P}} \ \mathrm{d}R\big(\mathbb{P}_k; \mathbb{Q}\big)$, but only its approximation g_k. Consequently, it is only possible to verify whether the upper bound (62) for $g_{\widehat{k}}$ is satisfied for some $\widehat{k}$, and impossible to verify whether (61) holds. Moreover, even if (62) holds for some $\widehat{k}$, it is not necessary that (61) also holds since $g_{\widehat{k}} \leqslant \sup_{\mathbb{Q} \in \mathcal{P}} \mathrm{d}R\big(\mathbb{P}_{\widehat{k}}; \mathbb{Q}\big)$. However, since $g_{\widehat{k}}$ is approximately equal to the FW-gap, an upper bound on $g_{\widehat{k}}$ gives the following slightly worse upper bound on the FW-gap

$$\sup_{\mathbb{Q} \in \mathcal{P}} \ \mathrm{d}R\big(\mathbb{P}_{\widehat{k}}; \mathbb{Q}\big) \leqslant \delta\gamma_{\widehat{k}}C + \mathrm{d}R\big(\mathbb{P}_{\widehat{k}}; \mathbb{Q}_{\widehat{k}}\big) \leqslant \frac{2\delta C}{K + 2} + \frac{4C(1 + \delta)}{K + 2} \leqslant \frac{2C(2C + 3\delta)}{K + 2}.$$

Finally, since $\mathbb{R}^{\mathbb{Q}} - \mathbb{R}^{\mathbb{P}} \leqslant \mathrm{d}R\big(\mathbb{P}; \mathbb{Q}\big)$ for every $\mathbb{Q} \in \mathcal{P}$ (follows from (16b)), taking supremum over $\mathbb{Q} \in \mathcal{P}$ immediately gives the inequality

$$R^* - R(\mathbb{P}) \leqslant \sup_{\mathbb{Q} \in \mathcal{P}} \ \mathrm{d}R\big(\mathbb{P}_{\widehat{k}}; \mathbb{Q}\big).$$

This completes the proof of the proposition. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

7.3. **Proofs for the NDRO problem.**

In this part we cover the technical proofs of the theoretical statements in Section 5.

Proof of Lemma 5.5 . First recall that $x(\mathbb{P}) = \mathrm{argmin}_{x \in \mathcal{X}} \ F\big(x, \mathbb{P}\big)$ for any $\mathbb{P} \in \mathcal{P}$.

Proof of Danskin's theorem. For any given $\mathbb{P}, \mathbb{Q} \in \mathcal{P}$, let us define the mappings

$$\begin{cases} \mathcal{X} \times [0, 1] \ni (x, \gamma) \longmapsto f(x, \gamma) \ := \ F\big(x, \mathbb{P}_\gamma\big) \\ \qquad\qquad [0, 1] \ni \gamma \longmapsto g(\gamma) \ := \ \min_{x \in \mathcal{X}} \ f(x, \gamma).^3 \end{cases}$$

It is easily seen that $g(\gamma) = R\big(\mathbb{P}_\gamma\big)$. Moreover, for the one sided derivatives of g at $\gamma \in [0, 1)$ defined: $dg(\gamma; 1) := \lim_{\delta \downarrow 0} \frac{1}{\delta}\big(g(\gamma + \delta) - g(\gamma)\big)$, we easily verify that $\mathrm{d}R\big(\mathbb{P}; \mathbb{Q}\big) = dg(0; 1)$. Using the short-hand notation $f_x(\cdot) := f(x, \cdot)$, we also verify similarly that $df_x(0; 1) = dF_x(\mathbb{P}, \mathbb{Q})$, for every $x \in \mathcal{X}$.

For every $\gamma \in [0, 1]$, let $x_\gamma := \mathrm{argmin}_{x \in \mathcal{X}} \ f(x, \gamma)$, we know from the Danskin's theorem [1, (A.3), p. 152] that the one sided derivatives $dg(\gamma; 1)$ at γ, are given by

$$dg(\gamma; 1) = df_{x_\gamma}(\gamma; 1) = \lim_{\delta \downarrow 0} \frac{1}{\delta}\big(f(x_\gamma, \gamma + \delta) - f(x_\gamma, \gamma)\big).$$

Collecting everything, we have $\mathrm{d}R\big(\mathbb{P}; \mathbb{Q}\big) = dg(0; 1) = df_{x_0}(0; 1) = dF_{x(\mathbb{P})}(\mathbb{P}; \mathbb{Q})$.

Proof of smoothness. Recall that $\nabla_1 F(x, \mathbb{P}) := \big(\partial F/\partial x\big)(x, \mathbb{P})$ denotes the partial derivative of F w.r.t. x evaluated at $(x, \mathbb{P})$. Since $x(\mathbb{P}) = \mathrm{argmin}_{x \in \mathcal{X}} F\big(x, \mathbb{P}\big)$, the first order optimality conditions give

$$\big\langle \nabla_1 F\big(x(\mathbb{P}), \mathbb{P}\big), \ y - x(\mathbb{P}) \big\rangle \ \geqslant \ 0 \quad \text{ for all } y \in \mathcal{X}. \tag{63}$$

Due to α-strong-convexity of $F(y, \mathbb{P})$ in y, we have

$$\frac{\alpha}{2} \|x(\mathbb{P}) - x(\mathbb{Q})\|^2 \leqslant F(x(\mathbb{Q}), \mathbb{P}) - F(x(\mathbb{P}), \mathbb{P}) - \langle \nabla_1 F(x(\mathbb{P}), \mathbb{P}), \, x(\mathbb{Q}) - x(\mathbb{P}) \rangle$$

$$\leqslant F(x(\mathbb{Q}), \mathbb{P}) - F(x(\mathbb{P}), \mathbb{P}) \quad \text{from (63) with } y = x(\mathbb{Q}).$$

Similarly, α-strong-convexity of $F(\cdot, \mathbb{Q})$ gives us

$$\frac{\alpha}{2} \|x(\mathbb{Q}) - x(\mathbb{P})\|^2 \leqslant F(x(\mathbb{P}), \mathbb{Q}) - F(x(\mathbb{Q}), \mathbb{Q}).$$

Combining the two inequalities, we infer that the inequality

$$\begin{aligned} \alpha \|x(\mathbb{P}) - x(\mathbb{Q})\|^2 &\leqslant F(x(\mathbb{Q}), \mathbb{P}) - F(x(\mathbb{P}), \mathbb{P}) + F(x(\mathbb{P}), \mathbb{Q}) - F(x(\mathbb{Q}), \mathbb{Q}) \\ &= F(x(\mathbb{Q}), \mathbb{P}) - F(x(\mathbb{Q}), \mathbb{Q}) + F(x(\mathbb{P}), \mathbb{Q}) - F(x(\mathbb{P}), \mathbb{P}) \\ &\leqslant dF_{x(\mathbb{Q})}(\mathbb{Q}; \mathbb{P}) + dF_{x(\mathbb{P})}(\mathbb{P}; \mathbb{Q}) \quad \text{from (16b)} \\ &= dR(\mathbb{Q}; \mathbb{P}) + dR(\mathbb{P}; \mathbb{Q}) \quad \text{from (38),} \end{aligned} \tag{64}$$

holds for every $\mathbb{P}, \mathbb{Q} \in \mathcal{P}$. On the one hand, for $\mathbb{P}_\gamma = \mathbb{P} + \gamma(\mathbb{Q} - \mathbb{P})$, $\gamma \in [0, 1]$, we have

$$\begin{aligned} dR(\mathbb{P}; \mathbb{P}_\gamma) &+ dR(\mathbb{P}_\gamma; \mathbb{P}) \\ &= dF_{x(\mathbb{P})}(\mathbb{P}; \mathbb{P}_\gamma) + dF_{x(\mathbb{P}_\gamma)}(\mathbb{P}_\gamma; \mathbb{P}) \quad \text{from (38)} \\ &= dF_{x(\mathbb{P})}(\mathbb{P}; \mathbb{P}_\gamma) - dF_{x(\mathbb{P}_\gamma)}(\mathbb{P}; \mathbb{P}_\gamma) + dF_{x(\mathbb{P}_\gamma)}(\mathbb{P}; \mathbb{P}_\gamma) + dF_{x(\mathbb{P}_\gamma)}(\mathbb{P}_\gamma; \mathbb{P}) \\ &= \gamma\Big(dF_{x(\mathbb{P})}(\mathbb{P}; \mathbb{Q}) - dF_{x(\mathbb{P}_\gamma)}(\mathbb{P}; \mathbb{Q})\Big) + \Big(dF_{x(\mathbb{P}_\gamma)}(\mathbb{P}; \mathbb{P}_\gamma) + dF_{x(\mathbb{P}_\gamma)}(\mathbb{P}_\gamma; \mathbb{P})\Big) \quad \text{from (16a)} \\ &\leqslant \gamma C_1 \|x(\mathbb{P}_\gamma) - x(\mathbb{P})\| + \gamma^2 C_2, \end{aligned} \tag{65}$$

where the last inequality is due to (37) and the smoothness condition of Assumption 5.3. On the other hand, considering $\mathbb{Q} = \mathbb{P}_\gamma$ in (64), we have

$$\alpha \|x(\mathbb{P}_\gamma) - x(\mathbb{P})\|^2 \leqslant RPgrad\mathbb{P}_\gamma\mathbb{P} + dR(\mathbb{P}; \mathbb{P}_\gamma). \tag{66}$$

Collecting (65) and (66) together, we see that

$$\alpha \|x(\mathbb{P}) - x(\mathbb{P}_\gamma)\|^2 \leqslant \gamma C_1 \|x(\mathbb{P}) - x(\mathbb{P}_\gamma)\| + \gamma^2 C_2.$$

On rearranging and simplifying terms, it is now easily verified that

$$\left(\|x(\mathbb{P}) - x(\mathbb{P}_\gamma)\| - \frac{\gamma}{2\alpha}\Big(C_1 - \sqrt{C_1^2 + 4\alpha C_2}\Big) \right) \left(\|x(\mathbb{P}) - x(\mathbb{P}_\gamma)\| - \frac{\gamma}{2\alpha}\Big(C_1 + \sqrt{C_1^2 + 4\alpha C_2}\Big) \right) \leqslant 0,$$

which is only true if

$$\frac{\gamma}{2\alpha}\Big(C_1 - \sqrt{C_1^2 + 4\alpha C_2}\Big) \leqslant \|x(\mathbb{P}) - x(\mathbb{P}_\gamma)\| \leqslant \frac{\gamma}{2\alpha}\Big(C_1 + \sqrt{C_1^2 + 4\alpha C_2}\Big).$$

The lower bound is irrelevant since it is negative. However, the upper bound is non-trivial, and employing it in (65) finally gives

$$\begin{aligned} dR(\mathbb{P}; \mathbb{P}_\gamma) + dR(\mathbb{P}_\gamma; \mathbb{P}) &\leqslant \gamma^2 \frac{C_1}{2\alpha}\Big(C_1 + \sqrt{C_1^2 + 4\alpha C_2}\Big) + \gamma^2 C_2 \\ &\leqslant \gamma^2 \left(C_2 + \frac{C_1}{2\alpha}\Big(C_1 + \sqrt{C_1^2 + 4\alpha C_2}\Big) \right). \end{aligned}$$

Since $C_2 + \frac{C_1}{2\alpha}\left(C_1 + \sqrt{C_1^2 + 4\alpha C_2}\right) < +\infty$, we conclude that the risk measure R is C-smooth in the sense of Definition 3.6 for every $C \geqslant C_2 + \frac{C_1}{2\alpha}\left(C_1 + \sqrt{C_1^2 + 4\alpha C_2}\right)$. The proof is now complete. □

Proof for Theorem 5.6. First, let us assume that assertion (ii) holds, then we see that

$$\min_{x \in \mathcal{X}} \sup_{\mathbb{Q} \in \mathcal{P}} F(x, \mathbb{Q}) \leqslant sup_{\mathbb{Q} \in \mathcal{P}} F(x_\varepsilon, \mathbb{Q}) \leqslant \varepsilon + \min_{x \in \mathcal{X}} F(x, \mathbb{P}_\varepsilon) \leqslant \varepsilon + \sup_{\mathbb{Q} \in \mathcal{P}} \min_{x \in \mathcal{X}} F(x, \mathbb{Q}),$$

holds for every $\varepsilon > 0$, and thus, we have $\min_{x \in \mathcal{X}} \sup_{\mathbb{Q} \in \mathcal{P}} F(x, \mathbb{Q}) \leqslant \sup_{\mathbb{Q} \in \mathcal{P}} \min_{x \in \mathcal{X}} F(x, \mathbb{Q})$. This, together with weak duality: $\sup_{\mathbb{Q} \in \mathcal{P}} \min_{x \in \mathcal{X}} F(x, \mathbb{Q}) \leqslant \min_{x \in \mathcal{X}} \sup_{\mathbb{Q} \in \mathcal{P}} F(x, \mathbb{Q})$ proves assertion (i).

To establish assertion (ii), let us recall that

$$R(\mathbb{P}) = \min_{x \in \mathcal{X}} F(x, \mathbb{P}) \quad \text{and} \quad x(\mathbb{P}) = \operatorname*{argmin}_{x \in \mathcal{X}} F(x, \mathbb{P}).$$

Since $dR(\mathbb{P}; \mathbb{Q}) = dF_{x(\mathbb{P})}(\mathbb{P}_k; \mathbb{Q})$ in view of (38), the iterates $(\mathbb{P}_k)_k$ obtained from Algorithm 1 can be equivalently regarded as the ones obtained from the FW-algorithm (27) for the maximization problem: $\sup_{\mathbb{P} \in \mathcal{P}} R(\mathbb{P})$ under the setting of Proposition 4.3. Therefore, we conclude from Proposition 4.3, and more specifically from Remark 4.4, we know that there exists a $K(\varepsilon) \leqslant \widehat{k} \leqslant 2K(\varepsilon) + 1$ such that $dR(\mathbb{P}_{\widehat{k}}; \mathbb{Q}_{\widehat{k}}) \leqslant \varepsilon \frac{2+2\delta}{2+3\delta}$.

We also conclude from Remark 4.4 that $\sup_{\mathbb{Q} \in \mathcal{P}} dR(\mathbb{P}_{\widehat{k}}; \mathbb{Q}) \leqslant \varepsilon$ for *any* $K(\varepsilon) \leqslant \widehat{k} \leqslant 2K(\varepsilon) + 1$ satisfying $dR(\mathbb{P}_{\widehat{k}}; \mathbb{Q}_{\widehat{k}}) \leqslant \varepsilon \frac{2+2\delta}{2+3\delta}$. In particular, for $(x_\varepsilon, \mathbb{P}_\varepsilon)$ to be the output of Algorithm 1, we know that

$$\sup_{\mathbb{Q} \in \mathcal{P}} dF_{x_\varepsilon}(\mathbb{P}_\varepsilon; \mathbb{Q}) = \sup_{\mathbb{Q} \in \mathcal{P}} dR(\mathbb{P}_\varepsilon; \mathbb{Q}) \leqslant \varepsilon.$$

Consequently, for any $\mathbb{Q} \in \mathcal{P}$, we have

$$F(x_\varepsilon, \mathbb{Q}) - F(x_\varepsilon, \mathbb{P}_\varepsilon) \leqslant dF_{x_\varepsilon}(\mathbb{P}_\varepsilon; \mathbb{Q}) \leqslant \varepsilon,$$

where the first inequality follows from (16b) for $F(x_\varepsilon, \cdot)$. Finally, taking the supremum over $\mathbb{Q} \in \mathcal{P}$ we conclude $\sup_{\mathbb{Q} \in \mathcal{P}} F(x_\varepsilon, \mathbb{Q}) \leqslant \varepsilon + F(x_\varepsilon, \mathbb{P}_\varepsilon)$; which together with the fact that $F(x_\varepsilon, \mathbb{P}_\varepsilon) = \min_{x \in \mathcal{X}} F(x, \mathbb{P}_\varepsilon)$ implies $(x_\varepsilon, \mathbb{P}_\varepsilon)$ being indeed an ε-saddle point in the sense of Definition 5.2. The proof is now complete. □

7.4. **Proofs of the minimum variance portfolio selection problem.**

In this part we cover the technical proofs of the theoretical statements in Section 6.

Proofs related to smoothness in the min-variance portfolio selection problem.

Proof of Lemma 6.2. For any $x \in \mathcal{X}$, since V_x is an RR measure, we simplify (45) to get

$$dV_x(\mathbb{P}; \mathbb{Q}) = x^\top(\Sigma_\mathbb{Q} - \Sigma_\mathbb{P})x - 2(x^\top \mu_\mathbb{P})(x^\top(\mu_\mathbb{Q} - \mu_\mathbb{P})) \tag{67a}$$

$$= x^\top(\Sigma_\mathbb{Q} - \Sigma_\mathbb{P}) \cdot x + (x^\top \mu_\mathbb{P})^2 + (x^\top(\mu_\mathbb{Q} - \mu_\mathbb{P}))^2 - (x^\top \mu_\mathbb{Q})^2. \tag{67b}$$

Proof for smooth partial G-derivatives. Consider any $x \in \mathcal{X}$, $\mathbb{P}, \mathbb{Q} \in \mathcal{W}_m(\widehat{\mathbb{P}}, \rho)$, and let $\mathbb{P}_\gamma := \mathbb{P} + \gamma(\mathbb{Q} - \mathbb{P})$ for $\gamma \in [0, 1]$. Using $\Sigma_{\mathbb{P}_\gamma} = \Sigma_\mathbb{P} + \gamma(\Sigma_\mathbb{Q} - \Sigma_\mathbb{P})$ and $\mu_{\mathbb{P}_\gamma} = \mu_\mathbb{P} + \gamma(\mu_\mathbb{Q} - \mu_\mathbb{P})$, we

conclude from (67a), that

$$\mathrm{d}V_x(\mathbb{P};\mathbb{P}_\gamma) = x^\top(\Sigma_{\mathbb{P}_\gamma} - \Sigma_\mathbb{P})x \; + \; 2\big(x^\top\mu_\mathbb{P}\big)\big(x^\top(\mu_{\mathbb{P}_\gamma} - \mu_\mathbb{P})\big)$$
$$= \; \gamma\Big(x^\top(\Sigma_\mathbb{Q} - \Sigma_\mathbb{P})x \; - \; 2\big(x^\top\mu_\mathbb{P}\big)\big(x^\top(\mu_\mathbb{Q} - \mu_\mathbb{P})\big)\Big).$$

Similarly, we also have

$$\mathrm{d}V_x(\mathbb{P}_\gamma;\mathbb{P}) = x^\top(\Sigma_\mathbb{P} - \Sigma_{\mathbb{P}_\gamma})x \; + \; 2\big(x^\top\mu_{\mathbb{P}_\gamma}\big)\big(x^\top(\mu_\mathbb{P} - \mu_{\mathbb{P}_\gamma})\big)$$
$$= 2\gamma^2\big(x^\top(\mu_\mathbb{Q} - \mu_\mathbb{P})\big)^2 + \; \gamma\Big(2\big(x^\top(\mu_\mathbb{Q} - \mu_\mathbb{P})\big)\big(x^\top\mu_\mathbb{P}\big) \; - \; x^\top(\Sigma_\mathbb{Q} - \Sigma_\mathbb{P})x\Big).$$

Combining the two equalities, we get

$$\mathrm{d}V_x(\mathbb{P};\mathbb{P}_\gamma) \; + \; \mathrm{d}V_x(\mathbb{P}_\gamma;\mathbb{P}) = 2\gamma^2\big(x^\top(\mu_\mathbb{Q} - \mu_\mathbb{P})\big)^2 \leqslant 2\gamma^2 \, \|x\|_2^2 \, \|\mu_\mathbb{Q} - \mu_\mathbb{P}\|_2^2 \leqslant \gamma^2\big(2B_\mu^2\big).$$

The last inequality follows from (46) and the fact that $\|x\|_2 \leqslant \|x\|_1 = 1$, for every $x \in \Delta^n$. Thus, the risk measure V_x is $\big(2B_\mu^2\big)$-smooth, uniformly over $x \in \mathcal{X}$. This proves assertion (ii) of the Lemma.

Proof for continuity of G-derivatives. Firstly, we begin by showing that the inequality

$$x^\top M x - y^\top M y \leqslant 2 \, \|M\|_o \, \|x - y\|_2 \quad \text{holds for any } x, y \in \Delta^n \text{ and } M \in \mathbb{S}^{n \times n}, \tag{68}$$

where, $\|M\|_o := \max_{\|v\|_2 = 1} \, |\langle v, \, Mv\rangle| \,$, is the operator norm. This follows since

$$\begin{aligned} x^\top M x - y^\top M y \; = \; & (x+y)^\top M(x-y) \\ \leqslant \; & \|x - y\|_2 \, \|M(x+y)\|_o \quad \text{from the Cauchy-Schwartz inequality} \\ \leqslant \; & \|x - y\|_2 \, \|x + y\|_2 \, \|M\|_o \quad \text{from the definition of } \|\cdot\|_o \\ \leqslant \; & 2 \, \|M\|_o \, \|x - y\|_2 \qquad \text{since } \|z\|_2 \leqslant \|z\|_1 = 1 \text{ for all } z \in \Delta^n. \end{aligned}$$

Now, for any $x, y \in \Delta^n$, we conclude from (67b) that

$$\begin{aligned} \mathrm{d}V_x(\mathbb{P};\mathbb{Q}) - \mathrm{d}V_y(\mathbb{P};\mathbb{Q}) \quad = \quad & x^\top(\Sigma_\mathbb{Q} - \Sigma_\mathbb{P})x \; - \; y^\top(\Sigma_\mathbb{Q} - \Sigma_\mathbb{P})y \\ & + \; (x^\top\mu_\mathbb{P})^2 \; + \; (x^\top(\mu_\mathbb{Q} - \mu_\mathbb{P}))^2 \; - \; (x^\top\mu_\mathbb{Q})^2 \\ & - \; (y^\top\mu_\mathbb{P})^2 \; - \; (y^\top(\mu_\mathbb{Q} - \mu_\mathbb{P}))^2 \; + \; (y^\top\mu_\mathbb{Q})^2. \end{aligned}$$

Employing (68) for $M = (\Sigma_\mathbb{Q} - \Sigma_\mathbb{P})$, $\mu_\mathbb{P}\mu_\mathbb{P}^\top$, $\mu_\mathbb{Q}\mu_\mathbb{Q}^\top$, and $(\mu_\mathbb{P} - \mu_\mathbb{Q})(\mu_\mathbb{P} - \mu_\mathbb{Q})^\top$; we have the inequalities

$$\begin{cases} x^\top(\Sigma_\mathbb{Q} - \Sigma_\mathbb{P})x \; - \; y^\top(\Sigma_\mathbb{Q} - \Sigma_\mathbb{P})y \; \leqslant \; 2 \, \|\Sigma_\mathbb{Q} - \Sigma_\mathbb{P}\|_o \, \|x - y\|_2 \\ (x^\top\mu_\mathbb{P})^2 - (y^\top\mu_\mathbb{P})^2 \; \leqslant \; 2 \, \|\mu_\mathbb{P}\|_2^2 \, \|x - y\|_2 \\ (y^\top\mu_\mathbb{Q})^2 - (y^\top\mu_\mathbb{Q})^2 \; \leqslant \; 2 \, \|\mu_\mathbb{Q}\|_2^2 \, \|x - y\|_2 \\ (x^\top(\mu_\mathbb{Q} - \mu_\mathbb{P}))^2 - (y^\top(\mu_\mathbb{Q} - \mu_\mathbb{P}))^2 \; \leqslant \; 2 \, \|\mu_\mathbb{Q} - \mu_\mathbb{P}\|_2^2 \, \|x - y\|_2, \end{cases} \tag{69}$$

where we have used the fact that $\|vv^\top\|_o = \|v\|_2^2$ for any v. These inequalities give us the upper bound

$$\mathrm{d}V_x(\mathbb{P};\mathbb{Q}) - \mathrm{d}V_y(\mathbb{P};\mathbb{Q}) \; \leqslant \; 2\Big(\|\Sigma_\mathbb{Q} - \Sigma_\mathbb{P}\|_o + \|\mu_\mathbb{P}\|_2^2 + \|\mu_\mathbb{Q}\|_2^2 + \|\mu_\mathbb{Q} - \mu_\mathbb{P}\|_2^2 \Big) \, \|x - y\|_2$$

Since $\|\mu_\mathbb{P}\|_2 \leqslant \|\mu_\mathbb{P} - \widehat{\mu}\|_2 + \|\widehat{\mu}\|_2$, and $\mathbb{P}, \widehat{\mathbb{P}} \in \mathcal{P}$, we conclude from (46) that $\|\mu_\mathbb{P}\|_2 \leqslant B_\mu + \|\widehat{\mu}\|_2$, (and similarly for $\mu_\mathbb{Q}$). This together with the condition $\|\Sigma_\mathbb{Q} - \Sigma_\mathbb{P}\|_o \leqslant B_\Sigma$ from (46), finally gives

$$\mathrm{d}V_x(\mathbb{P};\mathbb{Q}) - \mathrm{d}V_y(\mathbb{P};\mathbb{Q}) \; \leqslant \; 2\Big(B_\Sigma + 2\big(B_\mu + \|\widehat{\mu}\|_2\big)^2 + B_\mu^2 \Big) \, \|x - y\|_2$$

The constant C_1 in the assertion (ii) of the Lemma is immediately picked as $C_1 = 2\Big(B_\Sigma + 2\big(B_\mu + \|\widehat{\mu}\|_2\big)^2 + B_\mu^2\Big)$. The proof is now complete. $\square$

Proofs for the case of unconstrained support, $(\Xi = \mathbb{R}^n)$.

Proof of Lemma 6.3. Substituting $q = s\bar{q}$ for $s \geqslant 0$ and $\|\bar{q}\| = 1$ in (51), it is equivalently reformulated as

$$\begin{cases} \max_{s \geqslant 0, \bar{q} \in \mathbb{R}^n} & \big(s(x^\top \bar{q}) + x^\top(\xi - v)\big)^2 - \eta s^m \\ \text{subject to} & \|\bar{q}\| = 1. \end{cases} \tag{70}$$

Observe that $\big(s(x^\top \bar{q}) + x^\top(\xi - v)\big)^2 \leqslant \big(s\,|x^\top \bar{q}| + |x^\top(\xi - v)|\big)^2$, where, equality holds if and only if $\mathrm{sgn}(x^\top \bar{q}) = \mathrm{sgn}(x^\top(\xi - v))$. Moreover, applying the Holder's inequality $|x^\top \bar{q}| \leqslant \|\bar{q}\|\, \|x\|_* = \|x\|_*$ yields

$$\big(s(x^\top \bar{q}) + x^\top(\xi - v)\big)^2 \leqslant \big(s\,\|x\|_* + \big|x^\top(\xi - v)\big|\big)^2. \tag{71}$$

The upper bound (71) is achieved if and only if $\bar{q} = \mathrm{sgn}(x^\top(\xi-v))\bar{q}_x$, for any $\bar{q}_x \in \mathrm{argmax}_{\{\|\bar{q}\| \leqslant 1\}}\ x^\top \bar{q}$. Thus, every such $\bar{q}$ is an optimal solution to (70), irrespective of s. Simplifying the optimization over $\bar{q}$ in (70), the problem reduces to

$$\max_{s \geqslant 0}\ \big(s\,\|x\|_* + \big|x^\top(\xi - v)\big|\big)^2 - \eta s^m. \tag{72}$$

Now, an optimal solution to (51) exists if and only if (72) admits an optimal solution s^*; in which case, we have $q(\eta, \xi) = s^*\,\mathrm{sgn}(x^\top(\xi - v))\bar{q}_x$.

Optimal solution to (72) under different settings. If $m < 2$, it is straight forward to see that the optimal value of (72) is unbounded irrespective of η. Consequently, no optimal solution exists for the maximization problem (51) in this setting. On the contrary, if $m > 2$, it is also easily seen that the objective function of (72) is coercive if and only if $\eta > 0$. Therefore, the maximal value of (72) (and Consequently (51)), is bounded and achieved.

For $m = 2$, the objective function of (72) can be simplified to

$$-s^2\big(\eta - \|x\|_*^2\big) + 2s\,\|x\|_*\,\big|x^\top(\xi - v)\big| + \big|x^\top(\xi - v)\big|^2.$$

It is clear that the optimal value of (72) is unbounded if $\eta < \|x\|_*^2$. Whereas, if $\eta > \|x\|_*^2$, it is bounded, in which case, equating the derivative w.r.t. s equal to 0, gives that the optimal value of (72) (and consequently, also $J(\eta, \xi)$) is equal to $\frac{\eta}{\eta - \|x\|_*^2}\,\big|x^\top(\xi - v)\big|^2$, which is achieved at $s^* = \frac{\|x\|_*\,|x^\top(\xi-v)|}{\eta - \|x\|_*^2}$. Finally, if $\eta = \|x\|_*^2$, the optimal value of (72) is unbounded if $x^\top(\xi_i - v) \neq 0$; otherwise, it is bounded and equal to 0 which is achieved for any $s \geqslant 0$. $\square$

Proof of Lemma 6.4. On the one hand, if $\big|x^\top(\xi_i - v)\big| = 0$ for all $i = 1, 2, \ldots, N$, we now see that the dual problem (50) reduces to

$$\inf_{\eta \geqslant \|x\|_*^2}\ \eta \rho^2,$$

which admits the optimal solution $\eta^* = \|x\|_*^2$ with an optimal value $\|x\|_*^2 \rho^2$. On the other hand, if $\left|x^\top(\xi_i - v)\right| > 0$ for at least some $i \in \{1, 2, \ldots, N\}$, the dual problem (50) reduces to

$$\inf_{\eta > \|x\|_*^2} \quad \eta\rho^2 + \frac{\eta}{\eta - \|x\|_*^2}\left(\frac{1}{N}\sum_{i=1}^{N}\left|x^\top(\xi_i - v)\right|^2\right). \tag{73}$$

It is easily verified that the first order optimality conditions for (73) are satisfied at

$$\eta_x := \|x\|_*^2 + \frac{\|x\|_*}{\rho}\sqrt{\frac{1}{N}\sum_{i=1}^{N}|x^\top(\xi_i - v)|^2}.$$

Thus, η_x is the unique optimal solution to (73), and consequently, to the dual problem (50). Moreover, due to strong duality of (50), we also know that $\mathbb{Q}_x(\xi) = \frac{1}{N}\sum_{i=1}^{N}\delta\big(\xi - (\xi_i + q_i')\big)$ is an optimal solution to the linear worst case distribution problem (49) for any collection $q_i' \in q(\eta, \xi_i)$, $i = 1, 2, \ldots, N$. $\qquad\square$

Proof of Proposition 6.5. We first see that if $x^* \in \operatorname{argmin}_{x\in\mathcal{X}}\left(\rho\|x\|_* + \sqrt{\langle x, Vx\rangle}\right)$, then the first order necessary optimality conditions imply that there exists a sub-gradient g of the function $\left(\rho\|x\|_* + \sqrt{\langle x, Vx\rangle}\right)$ at x^* for which the inclusion $x^* \in \operatorname{argmin}_{y\in\mathcal{X}}\langle g, y\rangle$ holds. A quick look reveals that g must be of the form $\rho\bar{q}_{x^*} + \left(1/\sqrt{\langle x^*, Vx^*\rangle}\right)Vx^*$, where $\bar{q}_{x^*} \in \operatorname{argmax}_{\{\|\bar{q}\|\leqslant 1\}}\langle x^*, \bar{q}\rangle$. Therefore, there exists some $\bar{q}_{x^*} \in \operatorname{argmax}_{\{\|\bar{q}\|\leqslant 1\}}\langle x^*, \bar{q}\rangle$ such that x^* satisfies the inclusion

$$x^* \in \operatorname*{argmin}_{y\in\mathcal{X}}\left(\rho\langle y, \bar{q}_{x^*}\rangle + \frac{\langle x^*, Vy\rangle}{\sqrt{\langle x^*, Vx^*\rangle}}\right). \tag{74}$$

Selecting such a $\bar{q}_{x^*}$ to define $\mathbb{P}^*$ in (55), we also verify that

$$\mu_{\mathbb{P}^*} = \frac{1}{N}\sum_{i=1}^{N}(\xi_i + q_i^*) = \widehat{\mu} + \frac{\rho\frac{1}{N}\sum_{i=1}^{N}\langle x^*, \xi_i - \widehat{\mu}\rangle}{\sqrt{\left\langle x^*, (\widehat{\Sigma} - \widehat{\mu}\widehat{\mu}^\top)x^*\right\rangle}}\bar{q}_{x^*} = \widehat{\mu}.$$

Now, we shall establish that the pair $(x^*, \mathbb{P}^*)$ is a saddle point by proving that both x^* and $\mathbb{P}^*$ are optimal solutions to their respective problems while the other is held fixed.

Optimality of x^* for the minimization condition. We begin by showing that the inclusion $x^* \in \operatorname{argmin}_{y\in\mathcal{X}}V(y, \mathbb{P}^*)$ holds, by showing that the corresponding first order optimality condition

$$x^* \in \operatorname*{argmin}_{y\in\mathcal{X}}\langle\nabla_1 V(x^*, \mathbb{P}^*), y\rangle = \operatorname*{argmin}_{y\in\mathcal{X}}\left\langle\left(\Sigma_{\mathbb{P}^*} - \widehat{\mu}\widehat{\mu}^\top\right)x^*, y\right\rangle, \tag{75}$$

is satisfied. This condition is also sufficient for optimality since $V(y, \mathbb{P}^*)$ is convex in y. Simplifying the cost function in (75) yields

$$\left\langle\left(\Sigma_{\mathbb{P}^*} - \widehat{\mu}\widehat{\mu}^\top\right)x^*, y\right\rangle$$

$$= \frac{1}{N}\sum_{i=1}^{N}\langle x^*, \xi_i + q_i^*\rangle\langle y, \xi_i + q_i^*\rangle - \left\langle x^*, (\widehat{\mu}\widehat{\mu}^\top)y\right\rangle$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left\langle x^*, (\xi_i\xi_i^\top - \widehat{\mu}\widehat{\mu}^\top)y\right\rangle + \langle x^*, \xi_i\rangle\langle y, q_i^*\rangle + \langle x^*, q_i^*\rangle\langle y, \xi_i\rangle + \langle x^*, q_i^*\rangle\langle y, q_i^*\rangle.$$

Letting $V := \widehat{\Sigma} - \widehat{\mu}\widehat{\mu}^\top$, we see that $V = \frac{1}{N} \sum_{i=1}^N (\xi_i - \widehat{\mu})\xi_i^\top$, and $\frac{1}{N} \sum_{i=1}^N |\langle x^*, \xi_i - \widehat{\mu} \rangle|^2 = \langle x^*, V x^* \rangle$, from which we simplify

$$
\begin{cases}
\dfrac{1}{N} \sum_{i=1}^N \langle x^*, \xi_i \rangle \langle y, q_i^* \rangle \;=\; \rho \langle y, \bar{q}_{x^*} \rangle \dfrac{\frac{1}{N} \sum_{i=1}^N \langle x^*, \xi_i - \widehat{\mu} \rangle \langle \xi_i, x^* \rangle}{\sqrt{\langle x^*, V x^* \rangle}} & = \rho \langle y, \bar{q}_{x^*} \rangle \sqrt{\langle x^*, V x^* \rangle}, \\[3ex]
\dfrac{1}{N} \sum_{i=1}^N \langle x^*, q_i^* \rangle \langle y, \xi_i \rangle \;=\; \rho \|x^*\|_* \dfrac{\frac{1}{N} \sum_{i=1}^N \langle x^*, \xi_i - \widehat{\mu} \rangle \langle \xi_i, y \rangle}{\sqrt{\langle x^*, V x^* \rangle}} & = \rho \|x^*\|_* \dfrac{\langle x^*, V y \rangle}{\sqrt{\langle x^*, V x^* \rangle}}, \\[3ex]
\dfrac{1}{N} \sum_{i=1}^N \langle x^*, q_i^* \rangle \langle y, q_i^* \rangle \;=\; \rho^2 \|x^*\|_* \langle y, \bar{q}_{x^*} \rangle \dfrac{\frac{1}{N} \sum_{i=1}^N |\langle x^*, \xi_i - \widehat{\mu} \rangle|^2}{\langle x^*, V x^* \rangle} & = \rho^2 \|x^*\|_* \langle y, \bar{q}_{x^*} \rangle.
\end{cases}
$$

Employing the above relations, we see that the objective function in (74) simplifies to

$$
\left\langle (\Sigma_{\mathbb{P}^*} - \widehat{\mu}\widehat{\mu}^\top) x^*, y \right\rangle = \langle x^*, V y \rangle + \rho \|x^*\|_* \frac{\langle x^*, V y \rangle}{\sqrt{\langle x^*, V x^* \rangle}} + \rho \langle y, \bar{q}_{x^*} \rangle \sqrt{\langle x^*, V x^* \rangle} + \rho^2 \|x^*\|_* \langle y, \bar{q}_{x^*} \rangle
$$

$$
= \left(\sqrt{\langle x^*, V x^* \rangle} + \rho \|x^*\|_* \right) \left(\frac{\langle x^*, V y \rangle}{\sqrt{\langle x^*, V x^* \rangle}} + \rho \langle y, \bar{q}_{x^*} \rangle \right).
$$

In view of the inclusion (74), the sufficient optimality condition (75) follows immediately. Consequently, the inclusion $x^* \in \operatorname{argmin}_{x \in \mathcal{X}} V(x, \mathbb{P}^*)$ also holds.

Optimality of $\mathbb{P}^*$ for the maximization condition. Similar to the proof of the minimization condition, we establish the maximization condition of the saddle point by showing that the first order optimality conditions are satisfied. We first recall from (53) (with $v = \mu_{\mathbb{P}^*} = \widehat{\mu}$) that

$$
\eta_{x^*} \;=\; \|x^*\|_*^2 + \frac{\|x^*\|_*}{\rho} \sqrt{\left\langle x^*, (\widehat{\Sigma} - \widehat{\mu}\widehat{\mu}^\top) x^* \right\rangle}.
$$

On the one hand, if $\left\langle x^*, (\widehat{\Sigma} - \widehat{\mu}\widehat{\mu}^\top) x^* \right\rangle = 0$, we have $\eta_{x^*} = \|x^*\|_*^2$, in which case, we conclude from assertion (ii-c) of Lemma 6.3 that $0 \in q(\eta_{x^*}, \xi_i, \widehat{\mu})$ for all $i = 1, 2, \ldots, N$. Moreover, $\left\langle x^*, (\widehat{\Sigma} - \widehat{\mu}\widehat{\mu}^\top) x^* \right\rangle = 0$ also implies that $\langle x^*, \xi_i - \widehat{\mu} \rangle = 0$, and hence $q_i^* = 0$ for all $i = 1, 2, \ldots, N$. Thus, we have $q_i^* \in q(\eta_{x^*}, \xi_i, \mu_{\mathbb{P}^*})$ for all $i = 1, 2, \ldots, N$.

On the other hand, if $\left\langle x^*, (\widehat{\Sigma} - \widehat{\mu}\widehat{\mu}^\top) x^* \right\rangle \neq 0$, we have $\eta_{x^*} > \|x^*\|_*^2$, in which case, we see

$$
q_i^* = \frac{\rho \langle x^*, \xi_i - \widehat{\mu} \rangle}{\sqrt{\left\langle x^*, (\widehat{\Sigma} - \widehat{\mu}\widehat{\mu}^\top) x^* \right\rangle}} \bar{q}_{x^*} = \frac{\|x^*\|_* \langle x^*, \xi_i - \widehat{\mu} \rangle}{\eta_{x^*} - \|x^*\|_*^2} \bar{q}_{x^*} \;\in\; q(\eta_{x^*}, \xi_i, \widehat{\mu}) \quad \text{for all } i = 1, 2, \ldots, N,
$$

where the last inclusion follows from assertion (ii-b) of Lemma 6.3 since $\bar{q}_{x^*} \in \operatorname{argmax}_{\|\bar{q}\| \leqslant 1} \langle x^*, \bar{q} \rangle$. Finally, noting that $\widehat{\mu} = \mu_{\mathbb{P}^*}$, we have $q_i^* \in q(\eta_{x^*}, \xi_i, \mu_{\mathbb{P}^*})$ for all $i \leqslant N$. Since $\mathbb{P}^* = \frac{1}{N} \sum_{i=1}^N \delta(\xi - \xi_i - q_i^*)$, we conclude from Lemma 6.4 that $\mathbb{P}^* \in \operatorname{argmax}_{\mathbb{Q} \in \mathcal{P}} \mathrm{d}V_{x^*}(\mathbb{P}^*; \mathbb{Q})$. Thus, $\mathbb{P}^*$ satisfies the first order optimality conditions which are also sufficient. Hence, in view of Remark 3.5, we conclude that the inclusion $\mathbb{P}^* \in \operatorname{argmax}_{\mathbb{P} \in \mathcal{P}} V(x^*, \mathbb{P})$ also holds. Therefore, $(x^*, \mathbb{P}^*)$ is indeed a saddle point of (44). The proof is now complete. $\qquad \square$

Proofs for the case of ellipsoidal support, $(\Xi = \mathcal{E}_M)$.

Proof of Lemma 6.6. Letting $q' := q + \xi$, the maximization problem (51) can be equivalently written in terms of q' as:

$$\sup_{\{q':\langle q', Mq'\rangle \leqslant 1\}} \left\langle q', \left(xx^\top - \eta \mathbb{I}_n\right)q' \right\rangle + 2\left\langle q', \eta\xi - (xx^\top)v \right\rangle + \left((x^\top v)^2 - \eta \|\xi\|_2^2\right). \tag{76}$$

Even though $M \succeq 0$, which makes the feasible set of (76) convex; it is to be observed that the objective function is concave if and only if $\eta \geqslant \|x\|_2^2$. Therefore, (76) is not a convex problem in general. However, we observe that (76) is a quadratically constrained quadratic program, that is strictly feasible. For such problems, the *S-procedure* guarantees an equivalent reformulation as a tractable SDP

$$\begin{cases} \min_{\lambda \in [0,+\infty),\ \theta \in \mathbb{R}} & -\theta \\ \text{subject to} & \begin{bmatrix} \eta \mathbb{I}_n - xx^\top + \lambda M & (xx^\top)v - \eta\xi \\ (xx^\top)v^\top - \eta\xi^\top & \eta \|\xi\|_2^2 - (x^\top v)^2 - \lambda - \theta \end{bmatrix} \succeq 0, \end{cases} \tag{77}$$

with the optimal values of (77) and (76) being equal. Moreover, if (θ^*, λ^*) is a solution to the SDP (77), we also conclude from the S-procedure that $q' = \left(\eta \mathbb{I}_n - xx^\top + \lambda^* M\right)^{-1}\left(\eta\xi - xx^\top v\right)$, is an optimal solution to the maximization problem (76). Consequently, for every $\eta \geqslant 0$, we conclude

$$q(\eta, \xi) = -\xi + \left(\eta \mathbb{I}_n - xx^\top + \lambda^* M\right)^{-1}\left(\eta\xi - xx^\top v\right),$$

is a solution to the maximization problem (76). Substituting for each $i = 1, 2, \ldots, N$, the maximization problem over q_i in (50) with its equivalent SDP (77), we immediately arrive at (56). Now, suppose $(\eta^*, \lambda^*, \theta^*)$ is a solution to the SDP (56). Then, the pair $(\mathbb{Q}_x, \eta_x)$ as given by (57) is a solution to the FW problem (49) and its dual (50), respectively. This concludes the proof. $\qquad\square$

References

[1] D. P. BERTSEKAS, *Control of uncertain systems with a set-membership description of the uncertainty.*, PhD thesis, Massachusetts Institute of Technology, 1971.

[2] D. BERTSIMAS, D. B. BROWN, AND C. CARAMANIS, *Theory and applications of robust optimization*, SIAM review, 53 (2011), pp. 464–501.

[3] J. BLANCHET, L. CHEN, AND X. Y. ZHOU, *Distributionally robust mean-variance portfolio selection with Wasserstein distances*, Management Science, 68 (2022), pp. 6382–6410.

[4] 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.

[5] S. P. BOYD AND L. VANDENBERGHE, *Convex optimization*, Cambridge university press, 2004.

[6] J. CAI, J. LI, AND T. MAO, *Distributionally robust optimization under distorted expectations*, Available at SSRN 3566708, (2020).

[7] Y. CHEN AND Y. HU, *Multivariate coherent risk measures induced by multivariate convex risk measures*, Positivity, 24 (2020), pp. 711–727.

[8] K. L. CLARKSON, *Coresets, sparse greedy approximation, and the frank-wolfe algorithm*, ACM Transactions on Algorithms (TALG), 6 (2010), pp. 1–30.

[9] E. DELAGE AND Y. YE, *Distributionally robust optimization under moment uncertainty with application to data-driven problems*, Operations research, 58 (2010), pp. 595–612.

[10] J. C. DUCHI AND H. NAMKOONG, *Learning models with uniform performance via distributionally robust optimization*, The Annals of Statistics, 49 (2021), pp. 1378–1406.

[11] E. ERDOĞAN AND G. IYENGAR, *Ambiguous chance constrained problems and robust optimization*, Mathematical Programming, 107 (2006), pp. 37–61.

[12] M. FRANK AND P. WOLFE, *An algorithm for quadratic programming*, Naval research logistics quarterly, 3 (1956), pp. 95–110.

[13] R. GAO, *Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality*, Operations Research, (2022).

[14] R. GAO AND A. KLEYWEGT, *Distributionally robust stochastic optimization with Wasserstein distance*, Mathematics of Operations Research, (2022).

[15] J. GOH AND M. SIM, *Distributionally robust optimization and its tractable approximations*, Operations research, 58 (2010), pp. 902–917.

[16] E. Y. HAMEDANI, A. JALILZADEH, N. S. AYBAT, AND U. V. SHANBHAG, *Iteration complexity of randomized primal-dual methods for convex-concave saddle point problems*, arXiv preprint arXiv:1806.04118, 5 (2018).

[17] M. JAGGI, *Revisiting frank-wolfe: Projection-free sparse convex optimization*, in International Conference on Machine Learning, PMLR, 2013, pp. 427–435.

[18] R. JIANG AND Y. GUAN, *Data-driven chance constrained stochastic program*, Mathematical Programming, 158 (2016), pp. 291–327.

[19] C. KENT, J. BLANCHET, AND P. GLYNN, *Frank-wolfe methods in probability space*, preprint available at arXiv:2105.05352, (2021).

[20] D. KUHN, P. MOHAJERIN ESFAHANI, V. A. NGUYEN, AND S. A. S., *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.

[21] N. LANZETTI, S. BOLOGNANI, AND F. DÖRFLER, *First-order conditions for optimization in the Wasserstein space*, arXiv preprint arXiv:2209.12197, (2022).

[22] L. LIU, Y. ZHANG, Z. YANG, R. BABANEZHAD, AND Z. WANG, *Infinite-dimensional optimization for zero-sum games via variational transport*, in International Conference on Machine Learning, PMLR, 2021, pp. 7033–7044.

[23] H. M. MARKOWITZ, *Portfolio selection, the journal of finance. 7 (1)*, 1952.

[24] P. MOHAJERIN ESFAHANI AND D. KUHN, *Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations*, Mathematical Programming, 171 (2018), pp. 115–166.

[25] A. NEMIROVSKI, *Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems*, SIAM Journal on Optimization, 15 (2004), pp. 229–251.

[26] A. NEMIROVSKIJ AND D. YUDIN, *Problem complexity and method efficiency in optimization*, New York, (1983).

[27] Y. NESTEROV, *Introductory lectures on convex optimization: A basic course*, vol. 87, Springer Science & Business Media, 2003.

[28] Y. NESTEROV, *Smooth minimization of non-smooth functions*, Mathematical programming, 103 (2005), pp. 127–152.

[29] D. H. NGUYEN AND T. SAKURAI, *A particle-based algorithm for distributional optimization on\textit {Constrained Domains} via variational transport and mirror descent*, arXiv preprint arXiv:2208.00587, (2022).

[30] V. A. NGUYEN, S. S. ABADEH, D. FILIPOVIĆ, AND D. KUHN, *Mean-covariance robust risk measurement*, preprint available at arXiv:2112.09959, (2021).

[31] G. PFLUG AND D. WOZABAL, *Ambiguity in portfolio selection*, Quantitative Finance, 7 (2007), pp. 435–442.

[32] K. POSTEK, D. DEN HERTOG, AND B. MELENBERG, *Computationally tractable counterparts of distributionally robust constraints on risk measures*, SIAM Review, 58 (2016), pp. 603–650.

[33] H. RAHIMIAN AND S. MEHROTRA, *Distributionally robust optimization: A review*, preprint available at arXiv:1908.05659, (2019).

[34] N. RUJEERAPAIBOON, D. KUHN, AND W. WIESEMANN, *Robust growth-optimal portfolios*, Management Science, 62 (2016), pp. 2090–2109.

[35] H. SCARF, *A min max solution of an inventory problem*, Studies in the mathematical theory of inventory and production, (1958).

[36] S. Shafieezadeh-Abadeh, L. Aolaritei, F. Dörfler, and D. Kuhn, *New perspectives on regularization and computation in optimal transport-based distributionally robust optimization*, preprint available at arXiv:2303.03900, (2023).

[37] S. Shafieezadeh-Abadeh, P. Mohajerin Esfahani, and D. Kuhn, *Distributionally robust logistic regression*, Advances in Neural Information Processing Systems, 28 (2015).

[38] J. E. Smith and R. L. Winkler, *The optimizer's curse: Skepticism and postdecision surprise in decision analysis*, Management Science, 52 (2006), pp. 311–322.

[39] H. Song, X. Zeng, Y. Chen, and Y. Hu, *Multivariate shortfall and divergence risk statistics*, Entropy, 21 (2019), p. 1031.

[40] F. Uhlig, *A recurring theorem about pairs of quadratic forms and extensions: A survey*, Linear algebra and its applications, 25 (1979), pp. 219–237.

[41] J. Wang, R. Gao, and Y. Xie, *Sinkhorn distributionally robust optimization*, preprint available at arXiv:2109.11926, (2021).

[42] W. Wiesemann, D. Kuhn, and M. Sim, *Distributionally robust convex optimization*, Operations Research, 62 (2014), pp. 1358–1376.

[43] J. Zhen, D. Kuhn, and W. Wiesemann, *Mathematical foundations of robust and distributionally robust optimization*, preprint available at arXiv:2105.00760, (2021).

[44] S. Zymler, D. Kuhn, and B. Rustem, *Worst-case value at risk of nonlinear portfolios*, Management Science, 59 (2013), pp. 172–188.