First-Order Methods for Distributionally Robust Mixed-Integer Optimization

Hubert Villuendas
Université Grenoble Alpes, Inria, CNRS, LIG, LJK
Grenoble, France
hubert.villuendas@univ-grenoble-alpes.fr

Mathieu Besançon
Université Grenoble Alpes, Inria, CNRS, LIG
Grenoble, France
mathieu.besancon@inria.fr

Jérôme Malick
Université Grenoble Alpes, CNRS, LJK
Grenoble, France
jerome.malick@univ-grenoble-alpes.fr

Abstract

We consider combinatorial optimization problems in which input data are only partly observed or subject to errors. When the probability distribution of uncertain parameters is unknown, one typically relies on a finite historical dataset to infer plausible distributions. In this context, Wasserstein Distributionally Robust Optimization (WDRO) has become a powerful framework, with its nice modeling and generalization properties.

In this paper, we consider WDRO models involving combinatorial or discrete decision structures, for which we provide a new approach. We propose a tractable approach, based on the entropic regularization of the value function, which enables the computation of stochastic gradient estimators. To tackle the problem, we propose to use a stochastic Frank–Wolfe algorithm to optimize the WDRO objective while preserving the combinatorial nature of the constraints. We illustrate our method on classical optimization problems, such as the minimum quadratic spanning tree.

In such uncertain settings, the standard approach is to rely on Empirical Risk Minimization, an efficient approach but notoriously sensitive to data quality, *i.e.* when the samples are limited or unrepresentative

1 Introduction

Context: decision, uncertainty, and examples. In combinatorial optimization, handling uncertain input parameters poses a significant challenge. Data-driven approaches typically rely on empirical risk minimization which consists of minimizing the expectation of the loss of a model on the empirical distribution of training data. In a context where the training data may be lacking, this empirical risk minimization optimization problem is solved with respect to an estimated distribution that may significantly differ from the true one. As a consequence, solutions obtained through (ERM) often suffer from *optimizer's curse*, which occurs when the solution to a data-driven optimization problem tends to leverage statistical noise in the data (Mohajerin Esfahani and Kuhn [2018]). Moreover, the optimization procedure itself has the potential to amplify estimation errors, due to the combinatorial structure of the problem (Elmachtoub and Grigas [2022]). Meanwhile, decision problems are increasingly characterized by high-dimensional data and large-scale models. The resulting optimization problems frequently encounter the *curse of dimensionality*, which limits the applicability of conventional stochastic programming methods. These limitations have motivated the development of Distributionally Robust Optimization (DRO), which aims to optimize the worst case among a set of plausible distributions centered around the observed data.

However, applying a robust distributional approach to combinatorial optimization problems faces the challenge of dealing with mixed-integer sets (a challenge that already exists in the deterministic case, as the problems of interest are generally $\mathcal{NP}$-hard). Thus, existing DRO approaches for combinatorial problems require problem-specific

formulations, *e.g.* for the robust vehicle routing problem (Ghosal and Wiesemann [2020]), for a facility location problem framework (Basciftci et al. [2021]), for fair transit resource allocation (Sun et al. [2023]) or scheduling for surgery allocation (Chow et al. [2022]). This is particularly true for DRO approaches using Wasserstein ambiguity sets, which are very popular in operations research and machine learning (Kuhn et al. [2019]), which is already a difficult problem to solve in the continuous case (Mohajerin Esfahani and Kuhn [2018]), and for which applications to discrete cases remain rare. The goal of our work is precisely to propose a general approach for solving WDRO models of mixed-integer optimization. We will illustrate our approach on two running examples: the Quadratic Minimum Spanning Tree Problem and Traffic Assignment Problem.

Example.

(i) *Uncertain Quadratic Minimum Spanning Tree Problem* (see *e.g.* (Assad and Xu [1992], de Meijer et al. [2025])). Let $G = (V, E)$ denote an undirected graph on n vertices and m edges. The Quadratic Minimum Spanning Tree Problem is a generalization of the classical Minimum Spanning Tree Problem where the objective function takes into account not only the cost of the selected edges, but also interaction costs $\mathcal{C}_{e,e'}$ between pairs of edges $e, e' \in E$. In our setting, the cost matrix is subject to uncertainty and is not known in advance; we only have access to training data taken from the set of possible cost matrices $\Xi \subset \mathbb{R}_+^{m\times m}$.

(ii) *Uncertain Traffic Assignment Problem* (see *e.g.* (Patriksson [2015])). Let $G = (V, A)$ be a directed graph representing a transportation network, subject to a set of origin-destination travel demands. Individual network users aim to minimize their own travel times and choose their paths accordingly. However, as link occupancy increases, congestion accumulates, increasing the travel time on that specific link. The objective of the classic Traffic Assignment Problem is to determine the flow to impose on each arc $a \in A$ such that no user is tempted to alter their path unilaterally. In this setting, uncertainty arises from two modeling parameters, denoted α and β, which are dependent on exogenous factors such as vehicle fleet composition or weather conditions. The full formulation of Traffic Assignment Problem is described in (3). In our setting, we thus have access to a finite amount of training data $\xi = (\alpha, \beta) \in \mathbb{R}_+^2$.

Contributions and outline. In this work, we propose a general framework to approach data-driven stochastic combinatorial problems, based on Wasserstein Distributionally Robust Optimization, motivated by its attractive guarantees (Mohajerin Esfahani and Kuhn [2018]). Our method proposes a differentiable surrogate of the WDRO objective by integrating entropic smoothing in the formulation of optimal transport-based ambiguity sets, following (Azizian et al. [2023], Vincent et al. [2024]); and we provide an algorithmic scheme to handle a full class of mixed-integer problems. We will illustrate our method on two running examples: the Uncertain Quadratic Minimum Spanning Tree Problem, and the Uncertain Traffic Assignment Problem.

The paper is organized as follows: Section 2 presents the mathematical framework of our work. Section 2.1 discusses the Wasserstein Distributionally Robust Optimization setup and its entropic regularization, while we derive the properties of our distributionally robust objective in Section 2.2 and establish our gradient estimators. Section 3 details our algorithmic framework: we propose to use a momentum-based stochastic Frank-Wolfe algorithm, and discuss the practical calibration of the bounds for the dual variable parameter. Finally, we illustrate our proposed framework on the Uncertain Quadratic Minimum Spanning Tree Problem and Traffic Assignment Problem in Section 4, showing in particular the robustness of our approach against distributional shift.

2 Smooth distributionally robust modeling

In this article, we consider a general class of data-driven combinatorial optimization problems

$$\min_{\mathbf{z}\in\mathcal{Z}} \quad \mathbb{E}_\xi\left[f(\mathbf{z},\xi)\right] \qquad \mathcal{Z} \subseteq \mathbb{R}^p \times \mathbb{Z}^{n-p}. \tag{1}$$

where ξ is an uncertain scenario lying in $\Xi \subseteq \mathbb{R}^d$. The function f denotes the value of the combinatorial problem with parameter $\mathbf{z}$ and scenario ξ. We aim to take a decision based on available training data $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$. Defining the empirical distribution as $\widehat{\mathbb{P}}_N \overset{\text{def}}{=} (\delta_{\widehat{\xi}_1} + \cdots + \delta_{\widehat{\xi}_N})/N$, the classical stochastic optimization approach called empirical risk minimization, or sample-average approximation, aims to minimize the average loss on the training dataset ([Shapiro et al., 2021]), namely:

$$\min_{\mathbf{z}\in\mathcal{Z}} \quad \frac{1}{N}\sum_{k=1}^{N} f(\mathbf{z},\widehat{\xi}_k). \tag{ERM}$$

While Empirical Risk Minimization (ERM) remains a classical framework in data-driven stochastic optimization, its performance can be sensitive to distributional shifts ([Duchi and Namkoong, 2019]), particularly when we only

have access to low-quality or unrepresentative training samples. This has motivated research for more robust learning paradigms. In this section, we present our regularized Wasserstein Distributionally Robust Optimization framework, following (Azizian et al. [2023]).

2.1 Mathematical setup

In this section, we present our mathematical setup and the set of assumptions that will be made throughout the paper. Since our approach relies on gradient-based methods, we assume that the loss function f is sufficiently smooth[1].

Assumption 1.

(i) The loss function f is defined on $\mathscr{O} \times \Xi$, where $\mathscr{O}$ is a convex open set containing $\mathcal{Z}$. The function $f(\cdot, \xi)$ is a $\mathscr{C}^2(\mathscr{O})$ function for all scenario $\xi \in \Xi$.

(ii) The function $\xi \mapsto \nabla_z f(\mathbf{z}, \xi)$ is continuous on Ξ for any $\mathbf{z} \in \mathscr{O}$.

This assumption is relatively general and satisfied by a wide range of applications in operations research. In particular, it is verified by our two running examples.

Example.

(i) *Uncertain Quadratic Minimum Spanning Tree Problem.* Using binary variables $z \in \{0,1\}^m$ modeling the membership of each edge in the spanning tree $\mathcal{T}$, the Uncertain Quadratic Minimum Spanning Tree Problem under cost $\xi \in \Xi$ writes

$$\left[\begin{array}{lll} \underset{\mathbf{z}\in\{0,1\}^m}{\text{minimize}} & f(\mathbf{z},\xi) = \mathbf{z}^\top \xi \mathbf{z} & \\ \text{subject to} & \sum_{i=1}^{m} \mathbf{z}_i = n-1 & \\ & \sum_{i\in S} \mathbf{z}_i \leq |S|-1 \quad \forall S \subseteq [\![m]\!]. & \end{array}\right. \tag{2}$$

Moreover, for a fixed scenario $\xi \in \Xi$, the function $f(\cdot, \xi)$ is a quadratic form, thus $f(\cdot, \xi) \in \mathscr{C}^2$. Its derivative with respect to $\mathbf{z}$ is given by $\mathbf{z} \mapsto (\xi + \xi^\top)\mathbf{z}$, and thus for any fixed spanning tree embedding $\mathbf{z} \in \mathcal{Z}$, the function $\xi \mapsto \nabla_z f(\mathbf{z}, \xi)$ is a continuous map, and f verifies the conditions of Assumption 1.

(ii) *Uncertain Traffic Assignment Problem*: on a network $G = (V, A)$, $q_{i\to j} \geq 0$ denotes the traffic demand between i and j, and $\mathscr{P}_{i\to j}$ stands for the set of paths connecting the origin i to the destination j, for each $(i,j) \in V^2$. For a deterministic scenario $\xi = (\alpha, \beta)$, the Traffic Assignment Problem writes:

$$\left[\begin{array}{lll} \underset{\mathbf{x}\in\mathbb{R}_+^{|A|}}{\text{minimize}} & f(\mathbf{x},\xi) = \sum_{a\in A} \int_0^{\mathbf{x}_a} t_a^{(0)} \left(1 + \alpha \left(\frac{s}{c_a}\right)^\beta\right) \mathrm{d}s & \\ \text{subject to} & \sum_{p\in\mathscr{P}_{i\to j}} \sum_{a\in p} f_p \mathbb{1}_p(a) = q_{i\to j} & \forall (i,j) \in V^2 \\ & \sum_{(i,j)\in V^2} \sum_{p\in\mathscr{P}_{i\to j}} f_p = \mathbf{x}_a & \forall a \in A \\ & f_p \geq 0 & \forall p \in \mathscr{P}_{i\to j}, \forall (i,j) \in V^2 \end{array}\right. \tag{3}$$

where $t_a^{(0)} > 0$ denotes the free-flow travel time, $c_a > 0$ represents the practical capacity of arc a, and $\alpha, \beta > 0$ are dimensionless tuning parameters.
The objective function is of $\mathscr{C}^2$, and its derivative with with respect to $\mathbf{x}$ is

$$\nabla_{\mathbf{x}} f(\mathbf{x}, \xi) = \sum_{a\in A} t_a^{(0)} \left(1 + \alpha \left(\mathbf{x}_a / c_a\right)^\beta\right)$$

which is continuous with respect to $\xi = (\alpha, \beta)$. Thus f verifies Assumption 1.

Since Ξ represents the set of all plausible scenarios for a combinatorial optimization problem, we can assume that $\Xi \subseteq \mathbb{R}^d$ is closed and bounded, and therefore compact. Moreover, we suppose that the decision space itself

[1] By requiring f to be $\mathscr{C}^2$, we ensure that the local curvature of the problem can be exploited by our stochastic gradient-based algorithm (see Section 3).

is compact, as is the case for our two running problems, as well as a wide variety of combinatorial problems, *e.g.* knapsack problems, TSP, shortest path, scheduling problems, *etc.*

Assumption 2.

(i) The set $\Xi \subset \mathbb{R}^d$ is compact and non-empty.

(ii) The set $\mathcal{Z}$ is compact, and its open cover $\mathcal{O}$ is bounded.

The paradigm of distributionally robust optimization aims to minimize the expected value of $f(z,\cdot)$ under the worst-case distribution within a neighborhood of the empirical measure, thereby accounting for data uncertainty. A particularly appealing way to define this neighborhood is through the optimal transport distance, called the Wasserstein distance (see *e.g.* (Peyré and Cuturi [2019])), defined as:

$$W_c(\mathbb{Q}_1,\mathbb{Q}_2) \stackrel{\text{def}}{=} \inf_{\substack{\pi \in \text{Prob}(\Xi\times\Xi) \\ [\pi]_1=\mathbb{Q}_1\,[\pi]_2=\mathbb{Q}_2}} \int_{\Xi\times\Xi} c\left(\xi,\xi'\right) \mathrm{d}\pi\left(\xi,\xi'\right).$$

with $c : \Xi \times \Xi \to \mathbb{R}_+$ a ground cost function. This leads to the Wasserstein Distributionally Robust Optimization (WDRO) problem:

$$\min_{\mathbf{z}\in\mathcal{Z}} \sup_{\substack{\mathbb{Q}\in\text{Prob}(\Xi) \\ W_c(\mathbb{Q},\widehat{\mathbb{P}}_N)\le\varrho}} \mathbb{E}_{\zeta\sim\mathbb{Q}}\left[f(\mathbf{z},\zeta)\right]. \tag{WDRO}$$

This formulation combines expressive modeling flexibility (Mohajerin Esfahani and Kuhn [2018]) and statistical guarantees (Le and Malick [2024]).

The inner supremum in (WDRO) is an optimization problem over an infinite-dimensional space, and is generally intractable. Then duality helps: under mild assumptions, the reformulation of (WDRO) yields a single-level minimization problem (Mohajerin Esfahani and Kuhn [2018]) of the form:

$$\min_{\mathbf{z}\in\mathcal{Z}} \inf_{\lambda\ge 0} \lambda\varrho + \mathbb{E}_{\widehat{\xi}\sim\widehat{\mathbb{P}}_N}\left[\sup_{\zeta\in\Xi}\left\{f(\mathbf{z},\zeta) - \lambda c\left(\widehat{\xi},\zeta\right)\right\}\right] \tag{4}$$

where λ stands for the dual variable associated to the constraint $W_c\left(\mathbb{Q},\widehat{\mathbb{P}}_N\right) \le \varrho$. However, the resulting dual function is non-smooth, which prevents the direct use of first-order methods to solve the problem. Following (Azizian et al. [2023]), we therefore consider a log-sum-exp smoothing of the inner supremum:

$$\min_{\mathbf{z}\in\mathcal{Z}} \inf_{\lambda\ge 0} \quad F(\mathbf{z},\lambda) \stackrel{\text{def}}{=} \lambda\varrho + \varepsilon\mathbb{E}_{\widehat{\xi}\sim\widehat{\mathbb{P}}_N}\left[\log\left(\mathbb{E}_{\zeta\sim\mathcal{N}(\widehat{\xi},\sigma^2\mathbf{I})}\left[\exp\left(\frac{f(\mathbf{z},\zeta) - \lambda c\left(\widehat{\xi},\zeta\right)}{\varepsilon}\right)\right]\right)\right] \tag{5}$$

where $\varepsilon > 0$ controls the smoothness, and σ^2 determines the sampling variance. The entropic approximation of the WDRO formulation of problem (1) becomes then (5), the minimization of F over $\mathcal{Z}\times\mathbb{R}_+$.

Remark 1 (Smoothing/Regularization). The above smoothing of WDRO has a nice interpretation by (double) regularization of WDRO. Under some mild assumptions (Azizian et al. [2023], Gao and Kleywegt [2023]), it can be shown that the smoothing is equivalent to introduice a Kullback-Leiber divergence term $\varepsilon\mathbf{KL}(\mathbb{Q}\,||\,\widehat{\mathbb{P}}_N)$ in the radius constraint in (WDRO). Interestingly, the nice statistical properties of WDRO extends to smoothed/regularized WDRO (Le and Malick [2024]).

2.2 The robust objective and its properties

The robust objective function ((5)) verifies useful properties that allow the employment of first-order methods. From a standard argument, we easily get the convexity of the robust objective, which we formalize in the next proposition for the sake of completeness.

Proposition 1 (Convexity of the regularized WDRO objective)**.** *Under Assumption 2, the function F defined in* (5) *is convex on $\mathcal{O}\times\mathbb{R}_+$.*

Proof. Let set for all $k \in [\![N]\!]$ the function $h_k : \mathcal{O}\times\mathbb{R}_+\times\Xi \to \mathbb{R}$ defined by $h_k(\mathbf{z},\lambda\,;\,\xi) \stackrel{\text{def}}{=} \left(f(\mathbf{z},\xi) - \lambda c(\widehat{\xi}_k,\xi)\right)/\varepsilon$. Since the function $f(\cdot,\xi)$ is convex, it comes that the functions $h_k(\cdot,\cdot\,;\,\xi)$ are convex on $\mathcal{O}\times\mathbb{R}_+$. Assumptions 2 and 4 and the continuity of f ensure that the functions $\zeta \mapsto h_k(\mathbf{z},\lambda;\zeta)$ are $\mathcal{N}\left(\widehat{\xi}_k,\sigma^2\mathbf{I}\right)$-integrable for any $(\mathbf{z},\lambda)\in\mathcal{O}\times\mathbb{R}_+$.

Let $k \in [\![N]\!]$, and $G_k : \mathscr{O} \times \mathbb{R}_+ \to \mathbb{R}$ defined by $G_k(\mathbf{z}, \lambda) \stackrel{\text{def}}{=} \log\left(\int_\Xi \exp\left(h_k(\mathbf{z}, \lambda\,;\,\zeta)\right) \mathrm{d}\mu(\zeta)\right)$, with $\mu = \mathscr{N}\left(\widehat{\xi}_k, \sigma^2\mathbf{I}\right)$, and show that G_k is convex on $\mathscr{O} \times \mathbb{R}_+$. Let $(\mathbf{z}_0, \lambda_0), (\mathbf{z}_1, \lambda_1) \in \mathscr{O} \times \mathbb{R}_+$ and $\tau \in [0,1]$. We write $y_\tau \stackrel{\text{def}}{=} \tau(\mathbf{z}_1, \lambda_1) + (1-\tau)(\mathbf{z}_0, \lambda_0)$. If $\tau \in \{0,1\}$, we immediately have $G_k(y_\tau) = \tau G_k(y_1) + (1-\tau)G_k(y_0)$, and otherwise, we suppose $\tau \in (0,1)$. For all $\zeta \in \Xi$, the convexity of $h_k(\cdot,\cdot\,;\,\zeta)$ gives

$$\begin{aligned} h_k(y_\tau\,;\,\zeta) &\leq \tau\, h_k(y_1\,;\,\zeta) + (1-\tau) h_k(y_0\,;\,\zeta) \\ \exp(h_k(y_\tau\,;\,\zeta)) &\leq \exp(h_k(y_1\,;\,\zeta))^\tau \cdot \exp(h_k(y_0\,;\,\zeta))^{(1-\tau)} \\ \int_\Xi \exp(h_k(y_\tau\,;\,\zeta))\mathrm{d}\mu(\zeta) &\leq \int_\Xi \exp(h_k(y_1\,;\,\zeta))^\tau \cdot \exp(h_k(y_0\,;\,\zeta))^{(1-\tau)} \mathrm{d}\mu(\zeta). \end{aligned} \tag{6}$$

Since $\tau \notin \{0,1\}$, we set $p \stackrel{\text{def}}{=} \frac{1}{\tau}$ and $q \stackrel{\text{def}}{=} \frac{1}{1-\tau}$ which are two real numbers in $(1,+\infty)$ such that $\frac{1}{p} + \frac{1}{q} = \tau + (1-\tau) = 1$. The functions $\exp(h_k(y_1\,;\,\cdot))^\tau$, and $\exp(h_k(y_0\,;\,\cdot))^{1-\tau}$ both are μ-integrable, thus the Hölder inequality on the right-hand side of (6) gives

$$\begin{aligned} \int_\Xi \exp(h_k(y_\tau\,;\,\zeta))\mathrm{d}\mu(\zeta) &\leq \left(\int_\Xi \left[\exp(h_k(y_1\,;\,\zeta))^\tau\right]^p \mathrm{d}\mu(\zeta)\right)^{\frac{1}{p}} \cdot \left(\int_\Xi \left[\exp(h_k(y_0\,;\,\zeta))^{1-\tau}\right]^q \mathrm{d}\mu(\zeta)\right)^{\frac{1}{q}} \\ &\leq \left(\int_\Xi \exp(h_k(y_1\,;\,\zeta))\mathrm{d}\mu(\zeta)\right)^\tau \cdot \left(\int_\Xi \exp(h_k(y_0\,;\,\zeta))\mathrm{d}\mu(\zeta)\right)^{1-\tau} \end{aligned}$$

and taking the logarithm on both sides yields:

$$\begin{aligned} G_k(y_\tau) &= \log\left(\int_\Xi \exp(h_k(y_\tau\,;\,\zeta))\mathrm{d}\mu(\zeta)\right) \\ &\leq \tau \log\left(\int_\Xi \exp(h_k(y_1\,;\,\zeta))\mathrm{d}\mu(\zeta)\right) + (1-\tau)\log\left(\int_\Xi \exp(h_k(y_0\,;\,\zeta))\mathrm{d}\mu(\zeta)\right) = \tau G_k(y_1) + (1-\tau)G_k(y_0). \end{aligned}$$

Since this inequality is verified for every $y_0, y_1 \in \mathscr{O} \times \mathbb{R}_+$ and $\tau \in [0,1]$, the function G_k is indeed convex. Therefore, each term $\log\left(\int_\Xi \exp\left(h_k(\mathbf{z}, \lambda\,;\,\xi)\right) \mathrm{d}\mathscr{N}\left(\widehat{\xi}_k, \sigma^2\mathbf{I}\right)(\zeta)\right)$ of the expectation on $\{\widehat{\xi}_1, \ldots, \widehat{\xi}_N\}$ in (5) is convex, and thus F is a convex function. □

As an application defined via a parameterized integral, the differentiability of F will follow from the properties of the loss function f.

Proposition 2 (Differentiability of the regularized WDRO objective)**.** *Under Assumptions 1 and 2, the function F defined in* (5) *is twice-differentiable on $\mathscr{O} \times \mathbb{R}_+$.*
Moreover, we have

$$\nabla_z F(\mathbf{z}, \lambda) = \mathbb{E}_{\widehat{\xi} \sim \widehat{\mathbb{P}}_N}\left[\mathbb{E}_{\zeta \sim \boldsymbol{\pi}_{\mathbf{z},\lambda}(\cdot|\widehat{\xi})}\left[\nabla_{\mathbf{z}} f(\mathbf{z}, \zeta)\right]\right] \tag{7}$$

$$\frac{\partial}{\partial \lambda} F(\mathbf{z}, \lambda) = \varrho - \mathbb{E}_{\widehat{\xi} \sim \widehat{\mathbb{P}}_N}\left[\mathbb{E}_{\zeta \sim \boldsymbol{\pi}_{\mathbf{z},\lambda}(\cdot|\widehat{\xi})}\left[c(\widehat{\xi}, \zeta)\right]\right]. \tag{8}$$

where $\boldsymbol{\pi}_{\mathbf{z},\lambda}(\cdot|\xi)$ is the probability distribution given by

$$\mathrm{d}\boldsymbol{\pi}_{\mathbf{z},\lambda}(\cdot|\xi) \propto \exp\left(\left(f(\mathbf{z}, \zeta) - \lambda c(\xi, \zeta)\right)/\varepsilon\right) \exp\left(-\|\xi - \zeta\|_2^2/2\sigma^2\right) \mathrm{d}\zeta. \tag{9}$$

Proof. Assumptions 1 and 2 gives all the conditions to apply *differentiation under the integral sign theorem* to the functions $(\mathbf{z}, \lambda) \mapsto \exp\left(\left(f(\mathbf{z}, \zeta) - \lambda c(\widehat{\xi}_k, \zeta)\right)/\varepsilon\right)$ for any $\zeta \in \Xi$, giving the differentiability of the functions

$$(\mathbf{z}, \lambda) \mapsto \mathbb{E}_{\zeta \sim \mathscr{N}(\widehat{\xi}_k, \sigma^2\mathbf{I})}\left[\exp\left(\left(f(\mathbf{z}, \zeta) - \lambda c(\widehat{\xi}_k, \zeta)\right)/\varepsilon\right)\right] \quad \text{with gradient} \quad \mathbb{E}_{\zeta \sim \mathscr{N}(\widehat{\xi}_k, \sigma^2\mathbf{I})}\left[\nabla \exp\left(\left(f(\mathbf{z}, \zeta) - \lambda c(\widehat{\xi}_k, \zeta)\right)/\varepsilon\right)\right].$$

Since this is true for all terms of the expectation on $\{\widehat{\xi}_1, \ldots, \widehat{\xi}_N\}$ in (5), and the log is $\mathscr{C}^\infty$, the function F is $\mathscr{C}^1$ on $\mathscr{O} \times \mathbb{R}_+$. A simple calculation gives an explicit expression for the gradient of F:

$$\nabla_z F(\mathbf{z}, \lambda) = \frac{1}{N}\sum_{k=1}^N \frac{\mathbb{E}_{\zeta \sim \mathscr{N}(\widehat{\xi}_k, \sigma^2\mathbf{I})}\left[\nabla_{\mathbf{z}} f(\mathbf{z}, \zeta) \exp\left(\left(f(\mathbf{z}, \zeta) - \lambda c(\widehat{\xi}_k, \zeta)\right)/\varepsilon\right)\right]}{\mathbb{E}_{\zeta \sim \mathscr{N}(\widehat{\xi}_k, \sigma^2\mathbf{I})}\left[\exp\left(\left(f(\mathbf{z}, \zeta) - \lambda c(\widehat{\xi}_k, \zeta)\right)/\varepsilon\right)\right]} = \mathbb{E}_{\widehat{\xi} \sim \widehat{\mathbb{P}}_N}\left[\mathbb{E}_{\zeta \sim \boldsymbol{\pi}_{\mathbf{z},\lambda}(\cdot|\widehat{\xi})}\left[\nabla_{\mathbf{z}} f(\mathbf{z}, \zeta)\right]\right]$$

$$\frac{\partial}{\partial \lambda} F(\mathbf{z}, \lambda) = \varrho - \frac{1}{N}\sum_{k=1}^N \frac{\mathbb{E}_{\zeta \sim \mathscr{N}(\widehat{\xi}_k, \sigma^2\mathbf{I})}\left[c(\widehat{\xi}_k, \zeta) \exp\left(\left(f(\mathbf{z}, \zeta) - \lambda c(\widehat{\xi}_k, \zeta)\right)/\varepsilon\right)\right]}{\mathbb{E}_{\zeta \sim \mathscr{N}(\widehat{\xi}_k, \sigma^2\mathbf{I})}\left[\exp\left(\left(f(\mathbf{z}, \zeta) - \lambda c(\widehat{\xi}_k, \zeta)\right)/\varepsilon\right)\right]} = \varrho - \mathbb{E}_{\widehat{\xi} \sim \widehat{\mathbb{P}}_N}\left[\mathbb{E}_{\zeta \sim \boldsymbol{\pi}_{\mathbf{z},\lambda}(\cdot|\widehat{\xi})}\left[c(\widehat{\xi}, \zeta)\right]\right].$$

Using the regularity of $\nabla_{\mathbf{z}} f(\mathbf{z},\cdot)$ given by Assumption 1, a similar argument applied to (7) and (8) gives that our robust objective function $F \in \mathscr{C}^2(\mathscr{O} \times \mathbb{R}_+)$. □

While equations (7) and (8) provide a closed-form expression for the gradient of F, an exact computation remains numerically difficult. Indeed, these expressions involve multi-dimensional integrals on Ξ. To overcome this issue, we use a Monte Carlo sampling approach to approximate the integrals. In practice, for a sampling budget $S \in \mathbb{N}$, and for samples $\zeta_1^{(k)}, \dots, \zeta_S^{(k)}$ sampled from the Gaussian measure $\mathcal{N}(\widehat{\xi}_k, \sigma^2 \mathbf{I})$, for all $k \in [\![N]\!]$, one define the Monte-Carlo estimate of the gradient as:

$$\mathbb{E}_{\zeta \sim \boldsymbol{\pi}_{\mathbf{z},\lambda}(\cdot|\widehat{\xi}_k)}\left[\varphi_k(\mathbf{z},\zeta)\right] \approx \sum_{s=1}^{S} \varphi_k(\mathbf{z},\zeta_s^{(k)}) \frac{w_s^{(k)}}{\|w^{(k)}\|_1} \tag{10}$$

where $\varphi_k(\mathbf{z},\zeta_s^{(k)})$ is to be replaced by $\nabla_{\mathbf{z}} f(\mathbf{z},\zeta_s^{(k)})$ (resp. $c(\widehat{\xi}_k,\zeta_s^{(k)})$)to match the terms of (7) (resp. (8)), and where $w_s^{(k)} \stackrel{\text{def}}{=} \exp\left(\left(f(\mathbf{z},\zeta_s^{(k)}) - \lambda c(\widehat{\xi}_k,\zeta_s^{(k)})\right)/\varepsilon\right)$ for $k \in [\![N]\!]$ and $s \in [\![S]\!]$.

This approximation can be intractable for problems where the dimension of the scenario is high, as it is the case for the Quadratic Minimum Spanning Tree Problem.

> *Example.* Computing the gradient estimates in (11) and (12) for the Quadratic Minimum Spanning Tree Problem with a sampling budget $S \in \mathbb{N}$ requires drawing $N \times S$ samples of $m \times m$ real matrices. In a graph with n vertices, the number of edges m is typically $m = \boldsymbol{\Theta}(n^2)$. Even for a relatively small graph where $n = 50$ and $m = 350$, evaluating the full-batch gradient with $S = 10$ and a training set of size $N = 100$ would involve generating and storing approximately 122 million scalars at each iteration, leading to a significant computational and memory burden.

To overcome these limitations, we adopt a mini-batch gradient scheme by restricting the gradient computation to a small subset of indexes $\mathcal{J} \subseteq [\![N]\!]$, following the method proposed in (Vincent et al. [2024]). Given a mini-batch $\mathcal{J} \subseteq [\![N]\!]$, and samples $\zeta_1^{(k)}, \dots, \zeta_S^{(k)} \sim \mathcal{N}(\widehat{\xi}_k, \sigma^2 \mathbf{I})$ for all $k \in \mathcal{J}$, one define our gradient estimator, in $\mathbf{z}$ and λ respectively, by taking the extern expectation on $\mathcal{J}$ of (7) and (8), and approximate each inner expectation as (10) proposes, thus giving the following estimates:

$$\mathscr{G}_{\mathbf{z}}^{\mathcal{J}}(\mathbf{z},\lambda) \stackrel{\text{def}}{=} \frac{1}{|\mathcal{J}|} \sum_{k \in \mathcal{J}} \sum_{s=1}^{S} \nabla_{\mathbf{z}} f(\mathbf{z},\zeta_s^{(k)}) \frac{w_s^{(k)}}{\|w^{(k)}\|_1} \tag{11}$$

$$\mathscr{G}_{\lambda}^{\mathcal{J}}(\mathbf{z},\lambda) \stackrel{\text{def}}{=} \varrho - \frac{1}{|\mathcal{J}|} \sum_{k \in \mathcal{J}} \sum_{s=1}^{S} c(\widehat{\xi}_k,\zeta_s^{(k)}) \frac{w_s^{(k)}}{\|w^{(k)}\|_1} \tag{12}$$

where $w^{(k)} \in \mathbb{R}^S$ is the vector defined by $w_s^{(k)} \stackrel{\text{def}}{=} \exp\left(\left(f(\mathbf{z},\zeta_s^{(k)}) - \lambda c(\widehat{\xi}_k,\zeta_s^{(k)})\right)/\varepsilon\right)$ for $s \in [\![S]\!]$.

The robust objective function (5) verifies useful properties that allow the use of first-order methods.

2.3 Properties of the gradient estimator

In this section, we investigate the basic useful properties verified by our gradient estimator that will be required to allow the use of our first-order-based algorithm (see Section 3 for more details).

Lemma 1 (Mini-batch unbiasedness)**.** *Let $x_1, \dots, x_n \in \mathbb{R}^d$. Let $\mathcal{J} \subseteq [\![N]\!]$ be a random subset of fixed size $b \in [\![N]\!]$, chosen uniformly and independently among all subsets of $[\![N]\!]$ with size b. Then, the empirical average over $\mathcal{J}$ is an unbiased estimator of the average over $[\![N]\!]$:*

$$\mathbb{E}\left[\frac{1}{|\mathcal{J}|} \sum_{k \in \mathcal{J}} x_k\right] = \frac{1}{N} \sum_{k=1}^{N} x_k.$$

Proof. For any $\mathcal{J} \subseteq [\![N]\!]$, since $\mathcal{J}$ has a known size b, we can write $\frac{1}{|\mathcal{J}|}\sum_{k\in\mathcal{J}} x_k$ as $\frac{1}{b}\sum_{k=1}^{N} x_k \mathbb{1}_{k\in\mathcal{J}}$, thus giving, by linearity:

$$\mathbb{E}\left[\frac{1}{b}\sum_{k=1}^{N} x_k \mathbb{1}_{k\in\mathcal{J}}\right] = \frac{1}{b}\sum_{k=1}^{N} x_k \mathbb{E}\left[\mathbb{1}_{k\in\mathcal{J}}\right] = \frac{1}{b}\sum_{k=1}^{N} x_k \mathbb{P}(k\in\mathcal{J}) = \frac{1}{b}\sum_{k=1}^{N} x_k \frac{\binom{N-1}{b-1}}{\binom{N}{b}} = \frac{1}{b}\sum_{k=1}^{N} x_k \frac{b}{N} = \frac{1}{N}\sum_{k=1}^{N} x_k.$$

□

To ensure that the expectations are finite and well-defined, we make the following further assumption:

Assumption 3. For any $\mathcal{J} \subseteq [\![N]\!]$, the family $\left(\mathscr{G}^{\mathcal{J}}\right)_{S\geq 1}$ is uniformly integrable.

Under this assumption, one can show that our gradient estimator doesn't depend on the choice of the mini-batch, and that we can find the appropriate sampling budget S to get arbitrarily close to ∇F:

Proposition 3. *Under Assumption 3, The gradient estimator $\mathscr{G}^{\mathcal{J}}$ is asymptotically unbiased:*

$$\mathbb{E}\left[\mathscr{G}^{\mathcal{J}}(\mathbf{z},\lambda)\right] \xrightarrow[S\to\infty]{} \nabla F(\mathbf{z},\lambda).$$

Proof. We only write the proof for the gradient in $\mathbf{z}$, since the proof for the derivative in λ is similar. The randomness comes from both the choice of the mini-batch $\mathcal{J}$ and the Monte Carlo sampling $\zeta_1^{(k)},\dots,\zeta_S^{(k)} \sim \mathcal{N}\left(\widehat{\xi}_k, \sigma^2\mathbf{I}\right)$. Since $\mathcal{J}$ is independently sampled out of $[\![N]\!]$, Lemma 1 ensures that the average over the mini-batch $\mathcal{J}$ is an unbiased estimator of the total empirical estimator. Applying the *law of total expectation* (Le Gall [2022]) – **Theorem 11.3** – over $\mathscr{F}$ the σ-algebra generated by all Monte-Carlo samples, we have

$$\begin{aligned}\mathbb{E}\left[\mathscr{G}^{\mathcal{J}}_{\mathbf{z}}(\mathbf{z},\lambda)\right] &= \mathbb{E}_{(\zeta_s^\bullet)_{s\geq 1}\sim\mathcal{N}(\widehat{\xi}_\bullet,\sigma^2\mathbf{I})}\left[\mathbb{E}_{\mathcal{J}}\left(\frac{1}{|\mathcal{J}|}\sum_{k\in\mathcal{J}}\sum_{s=1}^{S}\nabla_{\mathbf{z}} f(\mathbf{z},\zeta_s^{(k)})\frac{w_s^{(k)}}{\|w^{(k)}\|_1}\,\middle|\,\mathscr{F}\right)\right] \\ &= \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}_{(\zeta_s^{(k)})_{s\geq 1}\sim\mathcal{N}(\widehat{\xi}_k,\sigma^2\mathbf{I})}\left[\sum_{s=1}^{S}\nabla_{\mathbf{z}} f(\mathbf{z},\zeta_s^{(k)})\frac{w_s^{(k)}}{\|w^{(k)}\|_1}\right].\end{aligned}$$

Now, for every $k\in[\![N]\!]$, the *law of large numbers* (classical, see e.g. (Le Gall [2022]) – **Theorem 10.8**) ensures that

$$\frac{1}{S}\sum_{s=1}^{S}\nabla_{\mathbf{z}} f(\mathbf{z},\zeta_s^{(k)})\,w_s^{(k)} \xrightarrow[S\to\infty]{\text{a.s.}} \mathbb{E}\left[\nabla_{\mathbf{z}} f(\mathbf{z},\cdot)\,w^{(k)}\right] \qquad (13)$$

$$\frac{1}{S}\|w^{(k)}\|_1 = \frac{1}{S}\sum_{s=1}^{S} w_s^{(k)} \xrightarrow[S\to\infty]{\text{a.s.}} \mathbb{E}\left[w^{(k)}\right] \qquad (14)$$

Since both maps $\nabla_{\mathbf{z}} f(\mathbf{z},\cdot)$ and $w^{(k)}$ are continuous, Assumption 2 – (i) ensures that both integrals of (13) and (14) are finite. Since the function $f(\mathbf{z},\cdot) - \lambda c(\widehat{\xi}_k,\cdot)$ is positive and Ξ is a non-empty subset (Assumption 2 – (i)), the limit in (14) is positive, therefore the *continuous mapping theorem* (classical, see e.g. (Van der Vaart [2000]) – **Theorem 2.3**) gives

$$\sum_{s=1}^{S}\nabla_{\mathbf{z}} f(\mathbf{z},\zeta_s^{(k)})\frac{w_s^{(k)}}{\|w^{(k)}\|_1} \xrightarrow[S\to\infty]{\text{a.s.}} \frac{\mathbb{E}\left[\nabla_{\mathbf{z}} f(\mathbf{z},\cdot)\,w^{(k)}\right]}{\mathbb{E}\left[w^{(k)}\right]}. \qquad (15)$$

Assumption 3 and the almost sure convergence of (15) allows to apply *Vitali's convergence theorem* ((Bogachev and Ruas [2007]), Theorem 4.5.4) – thus giving

$$\mathbb{E}\left[\mathscr{G}^{\mathcal{J}}_{\mathbf{z}}(\mathbf{z},\lambda)\right] = \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}_{(\zeta_s^{(k)})_{s\geq 1}\sim\mathcal{N}(\widehat{\xi}_k,\sigma^2\mathbf{I})}\left[\sum_{s=1}^{S}\nabla_{\mathbf{z}} f(\mathbf{z},\zeta)\frac{w_s^{(k)}}{\|w^{(k)}\|_1}\right] \xrightarrow[s\to\infty]{} \frac{1}{N}\sum_{k=1}^{N}\frac{\mathbb{E}\left[\nabla_{\mathbf{z}} f(\mathbf{z},\cdot)\,w^{(k)}\right]}{\mathbb{E}\left[w^{(k)}\right]} = \nabla_{\mathbf{z}} F(\mathbf{z},\lambda).$$

□

In the next section, we apply the stochastic Frank-Wolfe algorithm to tackle our robust problem.

3 Stochastic Frank-Wolfe

Optimizing our distributionally robust objective (5) over combinatorial domains involves managing the discrete geometry of the feasible set. To tackle the constraints, we use a Frank-Wolfe algorithm that imposes the constraints thanks to Linear Minimization Oracle (Frank and Wolfe [1956]). This approach has been applied to mixed-integer and robust optimization problems (Braun et al. [2025], Besançon and Kurtz [2024]).

In this section, we first describe the proposed stochastic Frank-Wolfe algorithm, which uses momentum and mini-batch sampling to handle the gradient of the regularized objective function. We then analyze its convergence properties.

3.1 Frank-Wolfe at work

To address the regularized WDRO problem defined in (5), we adopt a *Frank-Wolfe* framework. This choice is motivated by the combinatorial structure of the feasible set $\mathcal{Z}$, whose convex-hull $\text{conv}(\mathcal{Z})$ is a polytope with a number of facets that increases exponentially with the size of the problem (Schrijver [2002]), making standard projected gradient methods intractable in a general framework, as the projection step requires solving a constrained quadratic program, which is intractable in the general case. Conversely, *Frank-Wolfe*-type algorithms are projection-free (Jaggi [2013]), relying only on a *Linear Minimization Oracle* (LMO):

$$\text{LMO}_{\mathcal{Z}}(\mathbf{g}) \in \underset{\mathbf{s}\in\text{conv}(\mathcal{Z})}{\text{argmin}} \langle \mathbf{g}, \mathbf{s} \rangle. \tag{16}$$

For a wide variety of combinatorial problems, this linear sub-problem is highly tractable and can be solved efficiently via specialized methods ([Besançon et al., 2022, 2025]).

Example.

(i) *Quadratic Minimum Spanning Tree Problem.* Over the convex hull of the spanning tree polytope, minimizing a linear objective function decouples from the quadratic interactions. Consequently, the LMO can be solved exactly and in polynomial time *Kruskal's* algorithm

(ii) *Traffic Assignment Problem.* Optimizing over the flow polytope described in (3), the LMO corresponds to a linear network flow subproblem. By decomposing the total demand across pairs of origin-destination, the linear oracle reduces to finding the shortest paths on the network given the current gradient weights (Mitradjieva and Lindberg [2013]). This shortest path subproblem is efficiently solved using *Dijkstra's* algorithm.

Algorithm 1 *(Informal) Stochastic Frank-Wolfe algorithm*

Input: step sizes $0 \le \alpha_t \le 1$
for $t = 0$ to $\ldots$ **do**
  $\widetilde{\nabla} F(y_t) \leftarrow$ gradient estimator
  $v_t \leftarrow \text{LMO}_{\mathcal{Z}}\left(\widetilde{\nabla} F(y_t)\right)$ ▷ v_t is an optimal solution of the linear subproblem (16)
  $y_{t+1} \leftarrow y_t + \alpha_t \left(v_t - y_t\right)$
end for

The convergence of such methods is very sensitive to the properties of the estimator $\widetilde{\nabla} F$. Specifically, our framework requires the estimator to be unbiased to avoid systematic drift in the descent direction (Bottou et al. [2018]). The following lemma establishes that mini-batch doesn't introduce any bias, and will allow us to show that our estimator is unbiased.

While the asymptotic unbiasedness proven in the previous section ensures that our estimator $\mathcal{G}^{\partial}$ is consistent, in the absence of variance reduction, stochastic Frank-Wolfe methods (Algorithm 1) often suffer from the constant gradient noise, which can prevent convergence. While classical solutions typically rely on increasing the sampling budget S at each iteration to control the gradient noise, such strategies lead to an increasing computational cost. Instead, we integrate a momentum term that provides a stable estimation of the descent direction by taking into account the previous gradient direction (Braun et al. [2025]), thus ensuring convergence without expanding the sample size. We thus propose a Stochastic Frank-Wolfe variant with momentum and mini-batching to handle our WDRO objective function:

Algorithm 2 *Momentum Stochastic Frank-Wolfe algorithm with mini-batch*

Input: $\mathbf{z}_0 \in \mathrm{conv}(\mathcal{Z})$, $\lambda_{\max}$ an upper bound for the dual variable λ, $\lambda_0 \in [0, \lambda_{\max})$, step sizes $0 \leq \alpha_t \leq 1$ and momentum terms $\beta_t \in [0,1]$ for $t \geq 0$ with $\beta_0 = 1$.
A budget $b \in [\![N]\!]$ for the mini-batch size, a sampling budget $S \in \mathbb{N}$ to compute integrals
for $t = 0$ to $\ldots$ **do**
  $\mathcal{J}_t \leftarrow$ random subset of $[\![N]\!]$ with $|\mathcal{J}_t| = b$
  for $k \in \mathcal{J}_t$ **do**
    sample $\zeta_1^{(k)}, \ldots, \zeta_S^{(k)} \sim \mathcal{N}\big(\widehat{\xi}_k, \sigma^2 \mathbf{I}\big)$ i.i.d.
    $w_s^{(k)} \leftarrow \exp\Big(\Big(f(\mathbf{z}_t, \zeta_s^{(k)}) - \lambda_t c(\widehat{\xi}_k, \zeta_s^{(k)})\Big)\big/_{\varepsilon}\Big)$
  end for
  $\mathscr{G}^{\mathcal{J}_t}(\mathbf{z}_t, \lambda_t) \leftarrow$ Estimate given by (11) and (12)
  $\widehat{\nabla} F(\mathbf{z}_t, \lambda_t) \leftarrow \beta_t \mathscr{G}_S^{\mathcal{J}_t}(\mathbf{z}_t, \lambda_t) + (1-\beta_t)\widehat{\nabla} F(\mathbf{z}_{t-1}, \lambda_{t-1})$ ▷ with $\widehat{\nabla} F(\mathbf{z}_0, \lambda_0) \stackrel{\text{def}}{=} \mathscr{G}_S^{\mathcal{J}_0}(\mathbf{z}_0, \lambda_0)$ for first iteration
  $\upsilon_t \leftarrow \mathrm{argmin}_{\upsilon} \mathrm{LMO}\big(\widehat{\nabla} F(\mathbf{z}_t, \lambda_t)\big)$
  $(\mathbf{z}_{t+1}, \lambda_{t+1}) \leftarrow (\mathbf{z}_t, \lambda_t) + \alpha_t (\upsilon_t - (\mathbf{z}_t, \lambda_t))$
end for

3.2 Convergence analysis

The standard convergence theory for *Frank-Wolfe* algorithms requires the objective function to be minimized over a convex and compact domain (Braun et al. [2025]). While the convex hull $\mathrm{conv}(\mathcal{Z})$ is necessarily compact from Assumption 2 – (ii), the dual variable λ is defined over the interval $[0, +\infty)$. To ensure the good convergence of the algorithm, we must identify a sufficiently large upper bound $\lambda_{\max}$ such that the solution space for λ can be restricted to the compact interval $[0, \lambda_{\max}]$ without losing the optimal solution. This is the purpose of Proposition 4 that relies on the following lemma:

Lemma 2. *Suppose Assumption 2 – (i) hold, let* $\mathbb{Q} \in \mathrm{Proba}(\Xi)$ *and* $g : \Xi \to \mathbb{R}$ *be a bounded* $\mathbb{Q}$*-measurable function. Then*

$$\mathbb{E}_{\zeta \sim \mathbb{Q}^g}\big[g(\zeta)\big] \geq \log\big(\mathbb{E}_{\zeta \sim \mathbb{Q}}\big[\exp\big(g(\zeta)\big)\big]\big)$$

where $\mathbb{Q}^g$ *is the distribution defined by* $\mathrm{d}\mathbb{Q}^g(\zeta) \propto \exp\big(g(\zeta)\big)\,\mathrm{d}\mathbb{Q}(\zeta)$.

Since Assumption 2 ensures that both $\mathcal{Z}$ and Ξ are compact, and f is continuous (Assumption 1), the quantity

$$\|f\|_\infty \stackrel{\text{def}}{=} \sup_{\mathbf{z} \in \mathcal{Z}, \xi \in \Xi} |f(\mathbf{z}, \xi)|$$

is well-defined and finite. We can now show that there exists $\lambda_{\max}$ large enough such that any optimal dual parameter $\lambda^\star$ of (5) can be found within $[0, \lambda_{\max}]$.

Proposition 4 (Upper bound on λ)**.** *Let* $m_c \stackrel{def}{=} \max_{k \in [\![N]\!]} \mathbb{E}_{\zeta \sim \mathcal{N}(\widehat{\xi}_k, \sigma^2 \mathbf{I})}\big[c(\widehat{\xi}_k, \zeta)\big]$ *and suppose that* $\varrho > m_c$.
Let $\lambda_{\max} \stackrel{def}{=} 2\|f\|_\infty \big/ (\varrho - m_c)$. *Then for all* $\mathbf{z} \in \mathcal{Z}$*:*

$$\inf_{\lambda \in \mathbb{R}_+} \lambda\varrho + \mathbb{E}_{\widehat{\xi} \sim \widehat{\mathbb{P}}_N}\big[\varphi_\varepsilon(\mathbf{z}, \lambda, \widehat{\xi})\big] = \inf_{\lambda \in [0, \lambda_{\max}]} \lambda\varrho + \mathbb{E}_{\widehat{\xi} \sim \widehat{\mathbb{P}}_N}\big[\varphi_\varepsilon(\mathbf{z}, \lambda, \widehat{\xi})\big]$$

where φ_ε *is the function on* $\mathcal{Z} \times \mathbb{R}_+ \times \Xi$ *defined by*

$$\varphi_\varepsilon(\mathbf{z}, \lambda, \xi) \stackrel{def}{=} \varepsilon \log\left(\mathbb{E}_{\zeta \sim \mathcal{N}(\xi, \sigma^2 \mathbf{I})}\left[\exp\left(\frac{f(\mathbf{z}, \zeta) - \lambda c(\xi, \zeta)}{\varepsilon}\right)\right]\right).$$

Proof. Let $\mathbf{z} \in \mathcal{Z}$ and $k \in [\![N]\!]$. An explicit calculation gives

$$\partial_\lambda \varphi_\varepsilon(\mathbf{z}, \lambda, \widehat{\xi}_k) = -\mathbb{E}_{\zeta \sim \boldsymbol{\pi}_{\mathbf{z}, \lambda}(\cdot | \widehat{\xi}_k)}\big[c(\widehat{\xi}_k, \zeta)\big]$$

where $\boldsymbol{\pi}_{\mathbf{z},\lambda}(\cdot|\widehat{\xi}_k)$ is the distribution defined in (9). We bound $-\partial_\lambda\varphi_\varepsilon$ uniformly in $z\in\mathcal{Z}$ and $\{\widehat{\xi}_k\}_{k\in[\![N]\!]}$. We have:

$$\begin{aligned}\mathbb{E}_{\zeta\sim\boldsymbol{\pi}_{\mathbf{z},\lambda}(\cdot|\widehat{\xi}_k)}\left[c(\widehat{\xi}_k,\zeta)\right] &= \mathbb{E}_{\zeta\sim\boldsymbol{\pi}_{\mathbf{z},\lambda}(\cdot|\widehat{\xi}_k)}\left[\frac{f(\mathbf{z},\zeta)-f(\mathbf{z},\zeta)+\lambda c(\widehat{\xi}_k,\zeta)}{\lambda}\right]\\ &= \frac{1}{\lambda}\mathbb{E}_{\zeta\sim\boldsymbol{\pi}_{\mathbf{z},\lambda}(\cdot|\widehat{\xi}_k)}\left[f(\mathbf{z},\zeta)\right]-\frac{\varepsilon}{\lambda}\mathbb{E}_{\zeta\sim\boldsymbol{\pi}_{\mathbf{z},\lambda}(\cdot|\widehat{\xi}_k)}\left[\frac{f(\mathbf{z},\zeta)-\lambda c(\widehat{\xi}_k,\zeta)}{\varepsilon}\right] \qquad (17)\end{aligned}$$

The first term in (17) is uniformly bounded by $\mathbb{E}_{\zeta\sim\boldsymbol{\pi}_{\mathbf{z},\lambda}(\cdot|\widehat{\xi}_k)}\left[\|f\|_\infty\right]/\lambda = \|f\|_\infty/\lambda$. To bound the second term, we apply Lemma 2 with function $g\overset{\text{def}}{=}(f(\mathbf{z},\cdot)-\lambda c(\widehat{\xi}_k,\cdot))/\varepsilon$ and the distribution $\mathbb{Q}\overset{\text{def}}{=}\mathcal{N}(\widehat{\xi}_k,\sigma^2\mathbf{I})\in\text{Proba}(\Xi)$. It gives:

$$\begin{aligned}-\frac{\varepsilon}{\lambda}\mathbb{E}_{\zeta\sim\boldsymbol{\pi}_{\mathbf{z},\lambda}(\cdot|\widehat{\xi}_k)}\left[\frac{f(\mathbf{z},\zeta)-\lambda c(\widehat{\xi}_k,\zeta)}{\varepsilon}\right] &\le -\frac{\varepsilon}{\lambda}\log\left(\mathbb{E}_{\zeta\sim\mathcal{N}(\widehat{\xi}_k,\sigma^2\mathbf{I})}\left[\exp\left(\frac{f(\mathbf{z},\zeta)-\lambda c(\widehat{\xi}_k,\zeta)}{\varepsilon}\right)\right]\right)\\ &\le -\frac{1}{\lambda}\mathbb{E}_{\zeta\sim\mathcal{N}(\widehat{\xi}_k,\sigma^2\mathbf{I})}\left[f(\mathbf{z},\zeta)-\lambda c(\widehat{\xi}_k,\zeta)\right] \qquad (18)\\ &= -\frac{1}{\lambda}\mathbb{E}_{\zeta\sim\mathcal{N}(\widehat{\xi}_k,\sigma^2\mathbf{I})}\left[f(\mathbf{z},\zeta)\right]+\mathbb{E}_{\zeta\sim\mathcal{N}(\xi,\sigma^2\mathbf{I})}\left[c(\widehat{\xi}_k,\zeta)\right]\\ &\le \frac{1}{\lambda}\|f\|_\infty+m_c\end{aligned}$$

where (18) uses Jensen's inequality on the convex function $-\log$. Finally, we have

$$-\partial_\lambda\varphi_\varepsilon(\mathbf{z},\lambda,\widehat{\xi}_k)\le\frac{2\|f\|_\infty}{\lambda}+m_c.$$

Assuming $\varrho>m_c$ and setting $\lambda_{\max}\overset{\text{def}}{=}2\|f\|_\infty/(\varrho-m_c)$, we have $0\le\varrho+\partial_\lambda\varphi_\varepsilon(\mathbf{z},\lambda_{\max},\widehat{\xi}_k)$ for all $\mathbf{z}\in\mathscr{O}$ and $k\in[\![N]\!]$. In particular, for all $\mathbf{z}\in\mathscr{O}$:

$$0\le\varrho+\mathbb{E}_{\widehat{\xi}\sim\widehat{\mathbb{P}}_N}\left[\partial_\lambda\varphi_\varepsilon(\mathbf{z},\lambda_{\max},\widehat{\xi}_k)\right]=\partial_\lambda\left(\lambda\varrho+\mathbb{E}_{\widehat{\xi}\sim\widehat{\mathbb{P}}_N}\left[\varphi_\varepsilon(\mathbf{z},\lambda_{\max},\widehat{\xi}_k)\right]\right)=\partial_\lambda F(\mathbf{z},\lambda_{\max}). \qquad (19)$$

Thanks to Proposition 1, we have that $F(\mathbf{z},\cdot)$ is convex for any fixed $\mathbf{z}\in\mathscr{O}$, thus (19) ensures that the dual minimizer $\lambda^\star$ on $\mathbb{R}_+$ may be found on $[0,\lambda_{\max}]$. □

Proposition 4 proposes an upper bound on the optimal dual variable $\lambda^\star\le\lambda_{\max}$. To make this upper bound even more explicit, we make an assumption on the cost ground-function c, to get rid of the quantity m_c defined in the proposition. Until now, c appearing in (WDRO) was considered generic. We add a mild Lipschitzness assumption on c (satisfied by many popular cost-function, including all p-norms):

Assumption 4. The cost function $c:\Xi\times\Xi\to\mathbb{R}_+$ is Lipschitz continuous with respect to the squared Euclidean norm, *i.e.* $c(\xi,\zeta)\le\mathbf{L}_c\|\xi-\zeta\|_2^2$ for all $\xi,\zeta\in\Xi$. In particular, $c(\xi,\xi)=0$ for all $\xi\in\Xi$.

Corollary 1. *Under Assumption 4, if $\varrho>\mathbf{L}_c\sigma^2 d$, then any optimal dual parameter $\lambda^\star$ of* (5) *can be found within* $[0,\lambda_{\max}]$ *with* $\lambda_{\max}\overset{def}{=}2\|f\|_\infty/(\varrho-\mathbf{L}_c\sigma^2 d)$.

Proof. Let $k\in[\![N]\!]$. Then $\mathbb{E}_{\zeta\sim\mathcal{N}(\widehat{\xi}_k,\sigma^2\mathbf{I})}\left[c(\widehat{\xi}_k,\zeta)\right]\le\mathbf{L}_c\mathbb{E}_{\zeta\sim\mathcal{N}(\widehat{\xi}_k,\sigma^2\mathbf{I})}\left[\|\widehat{\xi}_k-\zeta\|_2^2\right]$ by Assumption 4. We can change variable for $\zeta=\widehat{\xi}_k+\sigma\mathbf{Z}$ and $\|\widehat{\xi}_k-\zeta\|_2^2=\sigma^2\|\mathbf{Z}\|_2^2=\sigma^2(\mathbf{Z}_1^2+\cdots+\mathbf{Z}_1^d)$, with $\mathbf{Z}_i\sim\mathcal{N}(0,1)$ for all $i\in[\![d]\!]$. In particular $\mathbb{E}_{\zeta\sim\mathcal{N}(\widehat{\xi}_k,\sigma^2\mathbf{I})}\left[c(\widehat{\xi}_k,\zeta)\right]\le\mathbf{L}_c\sigma^2\left(\mathbb{E}_{\mathcal{N}(0,1)}\left[\mathbf{Z}_1^2\right]+\cdots+\mathbb{E}_{\mathcal{N}(0,1)}\left[\mathbf{Z}_d^2\right]\right)$. Integration by part gives $\mathbb{E}_{\mathcal{N}(0,1)}\left[\mathbf{Z}_i^2\right]=1$, such that $\mathbb{E}_{\zeta\sim\mathcal{N}(\widehat{\xi}_k,\sigma^2\mathbf{I})}\left[c(\widehat{\xi}_k,\zeta)\right]\le\mathbf{L}_c\sigma^2 d$. Since this is true for all $k\in[\![N]\!]$, defining m_c as in Proposition 4, we have $2\|f\|_\infty/(\varrho-m_c)\le\lambda_{\max}\overset{\text{def}}{=}2\|f\|_\infty/(\varrho-\mathbf{L}_c\sigma^2 d)$, and Proposition 4 gives the wanted result. □

Proposition 4 and corollary 1 allows us to can choose a compact set on which to optimize our function, allowing us to utilize the convergence properties of the FW algorithm. A first result is the Lipschitzness properties of the function's gradients, which we formalize in the following lemma:

Lemma 3. *Let Assumptions 1 and 2 hold. Then there exists $\mathbf{L}_F>0$ such that ∇F is $\mathbf{L}_F$-Lipschitz on $\mathcal{Z}\times[0,\lambda_{\max}]$.*

Proof. Using Assumptions 1 and 2, Proposition 2 shows that our robust objective F is twice differentiable on the compact set $\mathcal{Z} \times [0, \lambda_{\max}]$. Thus, there exists $\mathbf{L}_F > 0$ such that

$$\sup_{(\mathbf{z},\lambda)\in \mathrm{conv}(\mathcal{Z}\times[0,\lambda_{\max}])} \|\nabla^2 F(\mathbf{z},\lambda)\| \leq \mathbf{L}_F.$$

The *mean value theorem* allows us to state that $\|\nabla F(\mathbf{z}_2,\lambda_2) - \nabla F(\mathbf{z}_1,\lambda_1)\| \leq \mathbf{L}_F \cdot \|(\mathbf{z}_2,\lambda_2) - (\mathbf{z}_1,\lambda_1)\|$ for any $(\mathbf{z}_1,\lambda_1),(\mathbf{z}_2,\lambda_2) \in \mathrm{conv}(\mathcal{Z}\times[0,\lambda_{\max}])$, and F is thus $\mathbf{L}_F$-smooth on $\mathrm{conv}(\mathcal{Z}\times[0,\lambda_{\max}])$. □

The $\mathbf{L}_F$-smoothness of F provides the theoretical foundation for the convergence of first-order methods ([Nesterov et al., 2018, Braun et al., 2025]), and is in particular needed for the convergence guarantees of stochastic *Frank-Wolfe* Algorithm 1.

Under this setting, we can establish the rate of convergence of the stochastic Frank-Wolfe algorithm.

Theorem 1. *Let Assumptions 1 to 4 hold, and let $B, C > 0$ be such that $\|f(\mathbf{z},\cdot)\| \leq B$ and $\|\nabla_{\mathbf{z}} f(\mathbf{z},\cdot)\| \leq C$ for all $(\mathbf{z},\lambda) \in \mathrm{conv}(\mathcal{Z}\times[0,\lambda_{\max}])$ and $\mathbf{L}_F$ a Lipschitz constant for F, whose existence is given by Assumption 1 and lemma 3. Let $D \overset{def}{=} \mathrm{diam}(\mathcal{Z}\times[0,\lambda_{\max}]) < +\infty$. With batch size b and sampling budget S, step-sizes $\alpha_t = 2/(t+7)$ and momentum terms $\beta_t = 4/(t+8)^{2/3}$, Algorithm 2 gives iterates $(\mathbf{z}_t,\lambda_t)_{t\geq 0} \in \mathrm{conv}(\mathcal{Z}\times[0,\lambda_{\max}])$ that verify, for all $t \geq 0$:*

$$\mathbb{E}\left[F(\mathbf{z}_t,\lambda_t)\right] - F(\mathbf{z}^\star,\lambda^\star) \leq \frac{Q}{\sqrt[3]{t+9}}$$

with $(\mathbf{z}^\star,\lambda^\star) \in \mathrm{conv}(\mathcal{Z}\times[0,\lambda_{\max}])$ a pair of minimizer of F and

$$Q \overset{def}{=} \max\left\{ \sqrt[3]{9}\left(F(\mathbf{z}_0,\lambda_0) - F(\mathbf{z}^\star,\lambda^\star)\right), \right.$$
$$\left. \frac{\mathbf{L}_F D}{2} + 2D\max\left\{ 3\|\nabla F(\mathbf{z}_0,\lambda_0)\|, \sqrt{\frac{64C^2}{bS}\cdot\exp\left(\frac{4B}{\varepsilon}\right)\left(1+\frac{8\|f\|_\infty \mathbf{L}_c\sigma^2}{\varepsilon(\varrho-\mathbf{L}_c\sigma^2 d)}\right)^{d/2} + 2\mathbf{L}_F^2 D^2} \right\} \right\}.$$

The proof relies on the fact that under the given hypothesis, we can give explicit bounds on the variance of the gradient estimator:

Lemma 4 (Variance of the gradient estimator)**.** *Let Assumption 4 hold, and let $B, C > 0$ be such that $\|f(\mathbf{z},\cdot)\| \leq B$ and $\|\nabla_{\mathbf{z}} f(\mathbf{z},\cdot)\| \leq C$ for all $(\mathbf{z},\lambda) \in \mathrm{conv}(\mathcal{Z}\times[0,\lambda_{\max}])$, whose existence is given by Assumption 1. With batch size b and sampling budget S, the gradient estimator given in* (11) *and* (12) *has variance at most:*

$$\mathrm{Var}\left[\mathcal{G}^{\mathcal{J}}(\mathbf{z},\lambda)\right] \leq \frac{4C^2}{bS}\cdot\exp\left(\frac{4B}{\varepsilon}\right)\left(1+\frac{4\lambda\mathbf{L}_c\sigma^2}{\varepsilon}\right)^{d/2}. \tag{20}$$

Proof. Let $\mathcal{J} \subseteq [\![N]\!]$ with size $b = |\mathcal{J}|$ and let $\mathbf{G}_k(\mathbf{z},\lambda)$ be the random variable defined as a function of the Monte-Carlo samples $\zeta_1^{(k)},\ldots,\zeta_S^{(k)} \sim \mathcal{N}\left(\widehat{\xi}_k,\sigma^2\right)$ i.i.d.:

$$\mathbf{G}_k(\mathbf{z},\lambda) \overset{\text{def}}{=} \sum_{s=1}^{S} \nabla_{\mathbf{z}} f(\mathbf{z},\zeta_s^{(k)}) \frac{w_s^{(k)}}{\|w^{(k)}\|_1} \quad \text{with} \quad w_s^{(k)} \overset{\text{def}}{=} \exp\left(\left(f(\mathbf{z},\zeta_s^{(k)}) - \lambda c(\zeta_s^{(k)},\widehat{\xi}_k)\right)\Big/_{\varepsilon}\right). \tag{21}$$

Then all the $\mathbf{G}_k(\mathbf{z},\lambda)$'s are independent random variables with identical variance, and thus

$$\mathrm{Var}\left[\mathcal{G}^{\mathcal{J}}(\mathbf{z},\lambda)\right] = \mathrm{Var}\left[\frac{1}{b}\sum_{k\in\mathcal{J}}\mathbf{G}_k(\mathbf{z},\lambda)\right] = \sum_{k\in\mathcal{J}}\frac{1}{b^2}\mathrm{Var}\left[\mathbf{G}_k(\mathbf{z},\lambda)\right] = \frac{1}{b}\mathrm{Var}\left[\mathbf{G}_0(\mathbf{z},\lambda)\right]$$

where $\mathbf{G}_0(\mathbf{z},\lambda)$ is the random variable defined as in (21) with $w_s^{(0)} \overset{\text{def}}{=} \exp\left(\left(f(\mathbf{z},\boldsymbol{\zeta}_s) - \lambda c(\boldsymbol{\zeta}_s,0)\right)\Big/_{\varepsilon}\right)$, and $\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_S \sim \mathcal{N}\left(0,\sigma^2\right)$. Since $\mathbf{G}_0(z,\lambda)$ can be written as a ratio of empirical averages, the delta method approximates its variance by identifying the its moments in the linearization of the mapping $(a,b) \mapsto a/b$ around around the expectations

$a = \mathbb{E}[\nabla_z f(z,\cdot)w]$ and $b = \mathbb{E}[w]$ through the first-order Taylor expansion, giving us

$$\begin{aligned}\mathbf{G}_0(\mathbf{z},\lambda) &\approx \mu + \frac{1}{\mathbb{E}[w]}\left((\nabla_{\mathbf{z}} f(\mathbf{z},\cdot)w - \mathbb{E}\left[\nabla_{\mathbf{z}} f(\mathbf{z},\cdot)\right]) - \mu(\mathbb{E}[w] - w)\right)\\ &\approx \mu + \frac{1}{\mathbb{E}[w]}\cdot\frac{1}{S}\sum_{s=1}^{S}\left(H(\boldsymbol{\zeta}_s) - \mathbb{E}[H]\right).\end{aligned}$$

with $\mu \overset{\text{def}}{=} \mathbb{E}[\nabla_{\mathbf{z}} f(\mathbf{z},\cdot)w]/\mathbb{E}[w]$ and H the random variable $H \overset{\text{def}}{=} w\left(\nabla_{\mathbf{z}} f(\mathbf{z},\cdot) - \mu\right)$. Note that $\mathbb{E}[H] = 0$ by construction. Since the $\boldsymbol{\zeta}_s$ are i.i.d., we have

$$\mathrm{Var}\left[\mathbf{G}_0(\mathbf{z},\lambda)\right] = \mathrm{Var}\left[\frac{1}{\mathbb{E}[w]}\cdot\frac{1}{S}\sum_{s=1}^{S}H(\boldsymbol{\zeta}_s)\right] = \frac{1}{\mathbb{E}[w]^2}\sum_{s=1}^{S}\frac{1}{S^2}\mathrm{Var}[H(\boldsymbol{\zeta}_s)] = \frac{1}{\mathbb{E}[w]^2}\cdot\frac{1}{S}\mathrm{Var}[H]$$

Now, we notice that $\mathrm{Var}[H] = \mathbb{E}[\|H - \mathbb{E}[H]\|^2] = \mathbb{E}\left[\|\nabla_{\mathbf{z}} f(\mathbf{z},\cdot) - \mu\|^2 w^2\right]$ which yields

$$\mathrm{Var}\left[\mathbf{G}_0(\mathbf{z},\lambda)\right] = \frac{1}{S}\cdot\frac{\mathbb{E}\left[\|\nabla_{\mathbf{z}} f(\mathbf{z},\cdot) - \mu\|^2 w^2\right]}{\mathbb{E}[w]^2} \quad \text{with} \quad \mu \overset{\text{def}}{=} \frac{\mathbb{E}\left[\nabla_{\mathbf{z}} f(\mathbf{z},\cdot)w\right]}{\mathbb{E}[w]} \quad \text{and} \quad w(\zeta) \overset{\text{def}}{=} \exp\left(\left(f(\mathbf{z},\zeta) - \lambda c(\zeta,0)\right)/\varepsilon\right).$$

Assumption 1 allows to find two constants $B, C > 0$ such that $\|f(\mathbf{z},\cdot)\| \le B$ and $\|\nabla_{\mathbf{z}} f(\mathbf{z},\cdot)\| \le C$ on Ξ. Thus we have $\|\mu\| \le C$, and the triangle inequality gives $\|\nabla_{\mathbf{z}} f(\mathbf{z},\cdot) - \mu\|^2 \le 4C^2$. It thus gives a first bound for $\mathrm{Var}\left[\mathbf{G}_0(\mathbf{z},\lambda)\right]$:

$$\mathrm{Var}\left[\mathbf{G}_0(\mathbf{z},\lambda)\right] \le \frac{4C^2}{S}\cdot\frac{\mathbb{E}\left[w^2\right]}{\mathbb{E}[w]^2}. \tag{22}$$

Thanks to Assumption 4, we can bound w^2 by:

$$w^2(\zeta) \le \exp\left(\frac{2B}{\varepsilon}\right)\exp\left(-\frac{2\lambda}{\varepsilon}c(\zeta,0)\right) \le \exp\left(\frac{2B}{\varepsilon}\right)\exp\left(-\frac{2\lambda\mathbf{L}_c}{\varepsilon}\|\zeta\|_2^2\right) \qquad \forall \zeta \in \Xi.$$

Now, using $\|\zeta\|^2 = \zeta_1^2 + \cdots + \zeta_d^2$ and the fact that all ζ_i are independent and identically distributed, we have

$$\begin{aligned}\mathbb{E}\left[w^2\right] &\le \exp\left(\frac{2B}{\varepsilon}\right)\cdot\mathbb{E}\left[\exp\left(-\frac{2\lambda\mathbf{L}_c}{\varepsilon}\|\zeta\|_2^2\right)\right] = \exp\left(\frac{2B}{\varepsilon}\right)\cdot\mathbb{E}\left[\prod_{i=1}^{d}\exp\left(-\frac{2\lambda\mathbf{L}_c}{\varepsilon}\zeta_i^2\right)\right]\\ &= \exp\left(\frac{2B}{\varepsilon}\right)\prod_{i=1}^{d}\mathbb{E}\left[\exp\left(-\frac{2\lambda\mathbf{L}_c}{\varepsilon}\zeta_i^2\right)\right] = \exp\left(\frac{2B}{\varepsilon}\right)\cdot\mathbb{E}\left[\exp\left(-\frac{2\lambda\mathbf{L}_c}{\varepsilon}\zeta_1^2\right)\right]^d\\ &\le \exp\left(\frac{2B}{\varepsilon}\right)\cdot\left(\int_{-\infty}^{+\infty}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{2\lambda\mathbf{L}_c}{\varepsilon}x^2\right)\exp\left(-\frac{x^2}{2\sigma^2}\right)\mathrm{d}x\right)^d\\ &= \exp\left(\frac{2B}{\varepsilon}\right)\left(1 + \frac{4\lambda\mathbf{L}_c\sigma^2}{\varepsilon}\right)^{-d/2}\end{aligned} \tag{23}$$

Similarly, using $w(\zeta) \ge \exp\left(-B/\varepsilon\right)\cdot\exp\left(-\lambda\mathbf{L}_c\|\zeta\|_2^2/\varepsilon\right)$, we have a lower bound for $\mathbb{E}[w]^2$:

$$\mathbb{E}[w]^2 \ge \exp\left(\frac{-2B}{\varepsilon}\right)\cdot\left(1 + \frac{2\lambda\mathbf{L}_c\sigma^2}{\varepsilon}\right)^{-d}. \tag{24}$$

Combining (22) to (24), we get to

$$\mathrm{Var}\left[\mathbf{G}_0(\mathbf{z},\lambda)\right] \le \frac{4C^2}{S}\cdot\exp\left(\frac{2B}{\varepsilon}\right)\cdot\left(1 + \frac{4\lambda\mathbf{L}_c\sigma^2}{\varepsilon}\right)^{-d/2}\cdot\exp\left(\frac{2B}{\varepsilon}\right)\cdot\left(1 + \frac{2\lambda\mathbf{L}_c\sigma^2}{\varepsilon}\right)^{d} \le \frac{4C^2}{S}\cdot\exp\left(\frac{4B}{\varepsilon}\right)\left(1 + \frac{4\lambda\mathbf{L}_c\sigma^2}{\varepsilon}\right)^{d/2}$$

and finally,

$$\mathrm{Var}\left[\mathcal{G}^{\mathcal{J}}(\mathbf{z},\lambda)\right] \le \frac{4C^2}{bS}\cdot\exp\left(\frac{4B}{\varepsilon}\right)\left(1 + \frac{4\lambda\mathbf{L}_c\sigma^2}{\varepsilon}\right)^{d/2}. \tag{25}$$

□

Proof of Theorem 1. This result is a direct application of Theorem 4.12 in (Braun et al. [2025]), involving the the fact that, F is convex and L-smooth, $\mathcal{Z}\times[0,\lambda_{\max}]$ has diameter D, and Lemma 4 ensures that stochastic gradient are generated with variance at most $4C^2/bS\cdot\exp\left(4B/\varepsilon\right)\left(1 + 4\lambda_t\mathbf{L}_c\sigma^2/\varepsilon\right)^{d/2}$ at each iteration.

As Corollary 1 states, we can bound each λ_t's by $\lambda_{\max} \le 2\|f\|_\infty/(\varrho - \mathbf{L}_c\sigma^2 d)$. Setting this as an upper bound for λ_t's, the wanted result holds. □

4 Numerical illustrations

We illustrate our tests on our two running examples, the Quadratic Minimum Spanning Tree Problem and Traffic Assignment Problem: for a training data set $\widehat{\xi}_1,\dots,\widehat{\xi}_{N_{\text{train}}}$, we want to compare the robustness of solutions obtained via our WDRO framework with those obtained via standard empirical risk minimization. We thus set

$$\mathbf{z}_{\text{WDRO}} \overset{\text{def}}{=} \operatorname*{argmin}_{\mathbf{z}\in\mathcal{Z}} \min_{\lambda\in[0,\lambda_{\max}]} F(\mathbf{z},\lambda) \tag{26}$$

$$\mathbf{z}_{\text{ERM}} \overset{\text{def}}{=} \operatorname*{argmin}_{\mathbf{z}\in\mathcal{Z}} \frac{1}{N_{\text{train}}} \sum_{k=1}^{N_{\text{train}}} f(\mathbf{z},\xi). \tag{27}$$

Solutions of problems (26) and (27) are obtained by running both models:

	ERM	WDRO
Objective	$\frac{1}{N_{\text{train}}} \sum_{k=1}^{N_{\text{train}}} f(\mathbf{z},\widehat{\xi}_k)$	$F(\mathbf{z},\lambda)$ as in (5)
Gradient	$\frac{1}{N_{\text{train}}} \sum_{k=1}^{N_{\text{train}}} \nabla_{\mathbf{z}} f(\mathbf{z},\widehat{\xi}_k)$	$\mathscr{G}^{\partial}(\mathbf{z},\lambda)$ as in (11) and (12)
LMO	Same linear oracle	
Algorithm	*Classical Frank-Wolfe*	Algorithm 2

Once both solutions are obtained, we want to compare how the robust model performs in comparison with ERM, facing unknown test data: we compare the how the parametrized losses $f(\mathbf{z}_{\text{WDRO}},\cdot)$ and $f(\mathbf{z}_{\text{ERM}},\cdot)$ on a shifted dataset $\widetilde{\boldsymbol{\xi}}_1,\dots,\widetilde{\boldsymbol{\xi}}_{N_{\text{test}}}$.

The empirical study of both approaches on toy-example of our data-driven combinatorial problems highlights that the out-of-sample performance of standard ERM is related to the complexity of the solution set $\mathcal{Z}$. Specifically, when the problem of interest has few solutions, ERM is structurally guided toward reliable solutions and offers good generalization guarantees, even under moderate distributional shifts. This empirical finding is formalized and demonstrated in Appendix A, Theorems 2 and 3. We therefore created a generator to build our instances for the spanning tree problem, and derived variant of existing instances for the traffic assignment, which allows us to control our custom instances of our running examples. This allows us to deliberately push the ERM to its limits, and to evaluate our WDRO framework in complex and flexible scenarios.

In Section 4.1, we discuss the experimental setup used to run our model, and to provide a comprehensive analysis, we evaluate the out-of-sample performance of the ERM and our robust framework on our two running examples individually. We will highlight how the distributionally robust approach adapts to different structural environments, from the discrete combinatorial structure of the Quadratic Minimum Spanning Tree Problem in Section 4.2 to the continuous case of the Traffic Assignment Problem in Section 4.3.

4.1 Experimental setting

For our setting, we choose $c(\xi,\zeta) \overset{\text{def}}{=} \|\xi-\zeta\|_2^2$ as ground-cost function for the Wasserstein distance. As mentioned before, to run Algorithm 2, one needs to give an explicit upper bound $\lambda_{\max}$ for the dual parameter λ. To ensure numerical instabilities, we want to avoid that $\lambda\|\widehat{\xi}_k-\zeta\|_2^2 \gg f(\mathbf{z},\zeta)$. The heuristic we propose is to set $\Delta_f > 0$ as the average dispersion of the function $f(\mathbf{z},\cdot)$ over Ξ, namely

$$\Delta_f \overset{\text{def}}{=} \mathbb{E}_{\mathbf{z}}\left[\operatorname{diam}(f(\mathbf{z},\Xi))\right] = \mathbb{E}_{\mathbf{z}}\left[\max_{\xi\in\Xi} f(\mathbf{z},\xi) - \min_{\zeta\in\Xi} f(\mathbf{z},\zeta)\right]$$

and $\overline{c} \overset{\text{def}}{=} \mathbb{E}_{\widehat{\xi}\sim\widehat{\mathbb{P}}_N}\left[\mathbb{E}_{\zeta}\left[\|\widehat{\xi}-\zeta\|_2^2\right]\right]$ the average cost. We then choose $\lambda_{\max}$ such that $\lambda_{\max}\overline{c}$ is of the same order of magnitude as Δ_f. Since Δ_f and $\overline{c}$ are hard to compute exactly, and given that we only seek a practical computation to get the upper bound $\lambda_{\max}$, we propose the following approximation-scheme heuristic to provide an upper bound for λ:

Algorithm 3 *Calibration of* $\lambda_{\max}$

Input: training samples $\{\widehat{\xi}_1,\dots,\widehat{\xi}_N\}$ of the learning problem

for $k \in [\![N]\!]$ **do**
  Sample $\zeta_1^{(k)}, \ldots, \zeta_S^{(k)} \sim \mathcal{N}(\widehat{\xi}_k, \sigma^2 \mathbf{I})$
  $\mathbf{z}_k \leftarrow \mathrm{LMO}(\zeta_k)$ with $\zeta_k \sim \mathcal{N}(\widehat{\xi}_k, \sigma^2 \mathbf{I})$
end for
$\widetilde{c} \leftarrow \dfrac{1}{NS} \displaystyle\sum_{k=1}^{N} \sum_{s=1}^{S} \|\widehat{\xi}_k - \zeta_s^{(k)}\|_2^2$ ▷ Approximation of the average cost
$\widetilde{\Delta_f} \leftarrow \dfrac{1}{N} \displaystyle\sum_{k=1}^{N} \left(\max_{s \in [\![S]\!]} \left\{ f(\mathbf{z}_k, \zeta_s^{(k)}) \right\} - \min_{s' \in [\![S]\!]} \left\{ f(\mathbf{z}_k, \zeta_s^{(k)}) \right\} \right)$ ▷ Approximation of the average dispersion
return $\lambda_{\max} \overset{\text{def}}{=} \widetilde{\Delta_f} / 2\widetilde{c}$

Remark 2. In Algorithm 3, the empirical average is computed by sampling $\{\zeta_s^{(k)}\}_s$, which can be done by calling the same method as for the gradient estimation in Algorithm 2.

Both algorithm, to compute $\mathbf{z}_{\text{WDRO}}$ and $\mathbf{z}_{\text{ERM}}$, were given a maximum iteration budget of 5000 iterations. The parameters for our robust setting *i.e.* the radius $\rho > 0$, the sampling variance σ^2 and the smoothing temperature $\varepsilon > 0$ has been individually and empirically fine-tuned by experimentation.

4.2 Uncertain Quadratic Minimum Spanning Tree Problem

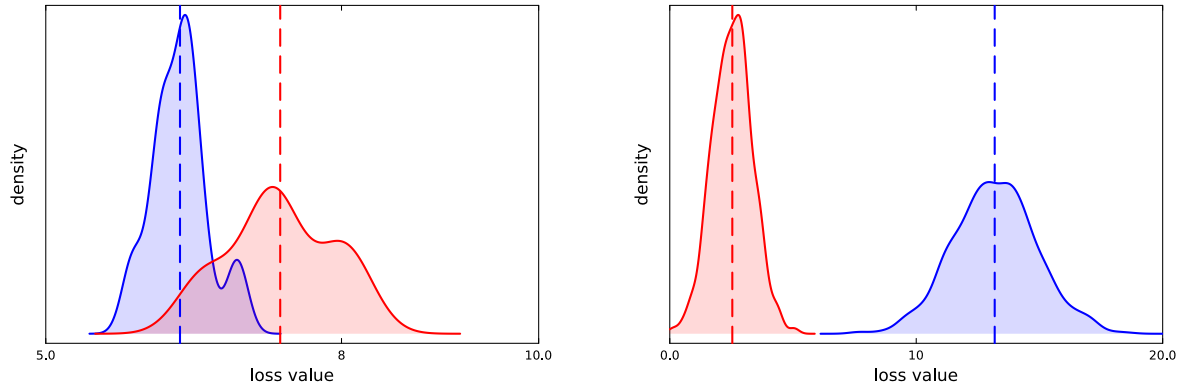
We generate instances of the problem with different properties: from very to less constrained settings. We refer to Appendix B.1 to discuss how the instances were generated (see Algorithm 4). The linear minimization oracle used is the Kruskal's algorithm (Schrijver [2002]) that runs $\mathcal{O}(m \log(m))$. We then performed the following tests:

SHIFT: we build a shifted distribution $\widetilde{\mathbf{P}_G}$ defined in the same way as in Algorithm 4, only with a different ground cost $\widetilde{\mu_G}$, a denser mask $\widetilde{M}$ and an increased noise. This distribution aims to simulate a regime shift toward greater instability. We then sample our test data according to this shifted distribution: $\Xi_{\text{shift}} \overset{\text{def}}{=} \{\widetilde{\boldsymbol{\xi}}_1, \ldots, \widetilde{\boldsymbol{\xi}}_{N_{\text{test}}}\}$ where $\widetilde{\boldsymbol{\xi}}_1, \ldots, \widetilde{\boldsymbol{\xi}}_{N_{\text{test}}} \sim \widetilde{\mathbf{P}_G}$.

Once we have our test data, we evaluate how both solutions perform when facing test data, thus building the following four sets:

$$\mathcal{L}_{\text{test}}(\mathbf{z}) \overset{\text{def}}{=} \left\{ f(\mathbf{z}, \xi) = \mathbf{z}^\top \boldsymbol{\xi} \mathbf{z} \,\middle|\, \boldsymbol{\xi} \in \Xi_{\text{test}} \right\} \qquad \mathbf{z} \in \{\mathbf{z}_{\text{ERM}}, \mathbf{z}_{\text{WDRO}}\} \quad \text{and} \quad \Xi_{\text{test}} \in \{\Xi_{\text{train}}, \Xi_{\text{shift}}\}.$$

To illustrate how both solutions perform facing a test dataset Ξ_{test}, we plot the histogram of the losses of $\mathcal{L}_{\text{test}}(\mathbf{z}_{\text{ERM}})$ and $\mathcal{L}_{\text{test}}(\mathbf{z}_{\text{WDRO}})$[2]



(a) training data (b) shift test

Figure 1: Quadratic Minimum Spanning Tree Problem example. Red smoothed histograms represent the losses obtained with the robust solution, blue histograms represent the losses obtained with the ERM. The ERM solution is overconfident and underperforms on novel data.

[2]For better readability, we illustrated the Kernel density estimation of the histograms.

Figure 1a illustrates the optimal behavior of the solution obtained by the empirical risk minimization facing the test data; on average, the losses of the robust solution are almost the worst every time. However, Figure 1b indicates that the robust tree offers smaller losses than the ERM tree when facing shifted data, indicating a tendency to be robust against distributional shift: the robust method has the opposite behavior compared to ERM: it promises with care and outperforms.

> *Remark* 3. As detailed in Appendix A, for combinatorial problems with a finite number of solutions, in highly constrained settings, the feasible set is restricted to a limited number of solutions. Thus, ERM is naturally guided toward reliable solutions, achieving satisfying performance even facing a distributional shift. The true advantages of our robust paradigm emerge in flexible, loosely constrained environments. More precisely, as shown by Theorem 2, the more elements $\mathcal{Z}$ has, the less efficient ERM should be, and our goal is to determine whether the robust WDRO objective can compensate for this effect.

As a consequence of the previous remark, we will measure how well WDRO performs compared to ERM as the problem gets less constrained. We control the complexity of the problem through two parameters: the size of the graph n, and its density $d(m,n) \overset{\text{def}}{=} 2m/n(n-1)$.

DENSITY TEST: for a fixed graph size $n = 50$, we increase the density of the graph by considering graphs on m_k edges, where $d(m_k, n) \approx k/10$, $k \in [\![10]\!]$, thus spreading the density evenly from the lowest to the maximal density. For each size and density, we generate a graph and multiple instances regarding only the generation of the training data: each instance has its own base cost μ_G, its own mask, and its own generation parameters (see Appendix B.1 for more details about the generation of instances). Each instance has thus different training scenarios and different dynamics on a shared graph structure, with common connectivity properties. We interpret the performance of the robust solution at a fixed size and density by aggregating the obtained differences between the ERM and the robust losses across the n_{inst} solved at each fixed dimension

$$\mathcal{D}^{(m)}_{\text{shift}} \overset{\text{def}}{=} \bigcup_{i=1}^{n_{\text{inst}}} \left\{ f(\mathbf{z}^{(m,i)}_{\text{ERM}}, \boldsymbol{\xi}) - f(\mathbf{z}^{(m,i)}_{\text{WDRO}}, \boldsymbol{\xi}) \,\middle|\, \boldsymbol{\xi} \in \Xi_{\text{shift}} \right\}$$

where $\mathbf{z}^{(m,i)}_{\text{ERM}}$ (respectively $\mathbf{z}^{(m,i)}_{\text{WDRO}}$) is the solution obtained by ERM (respectively WDRO) at the i^{th} instance of size m, for $i \in [\![n_{\text{inst}}]\!]$. For this experiment, we solved $n_{\text{inst}} = 10$ instances at each fixed size.

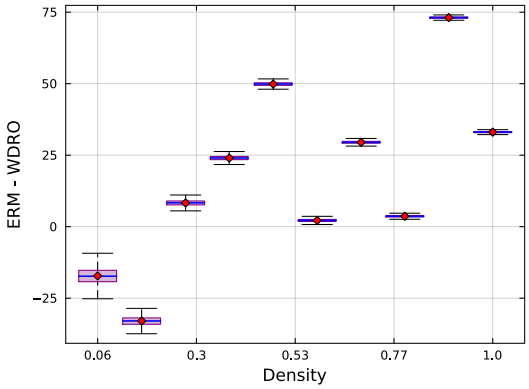


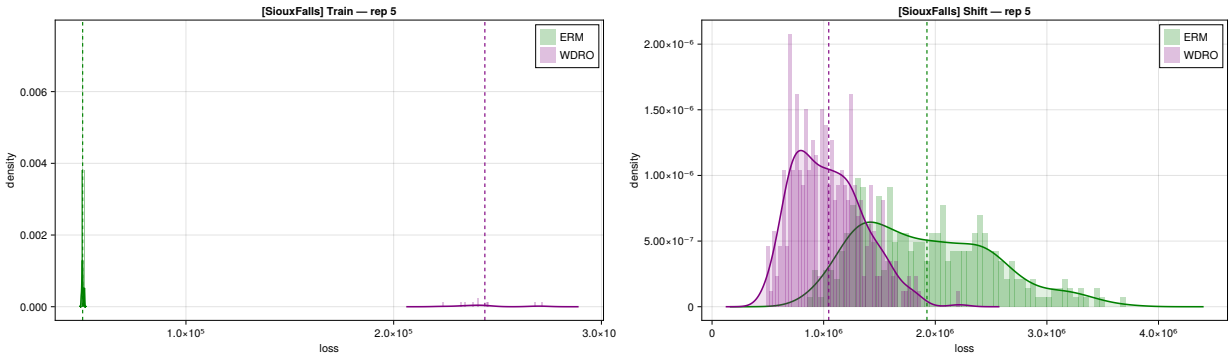
Figure 2: aggregated losses for instances of the Quadratic Minimum Spanning Tree Problem on a graph G on $n = 50$ vertices and $m \in \{75, 203, \ldots, 1225\}$ edges.

Section 4.2 shows a shrinking of the dispersion effect. Specifically, as the graph density increases, the ERM solution degrades while the WDRO framework yields lower regret. Moreover, the variance gets narrower as the density increases, showing that the robust method becomes superior to ERM under higher density. On the other hand, the averages of $\mathcal{D}^{(m)}_{\text{shift}}$ don't follow a particular trend, indicating that the evaluation of the robustness is very instance dependent, and also very dependent of the shifted distribution $\widetilde{\mathbf{P}_G}$. Despite this randomness effect, as the density of the graph increases, the averages of $\mathcal{D}^{(m)}_{\text{shift}}$ are above 0, meaning that we have a gain of robustness against distributional shift in average for the robust method compared to ERM.

4.3 Uncertain Traffic Assignment Problem

We construct our instances based on the "SiouxFalls" network, with 24 nodes and 76 links, from the *TransportationNetwork* dataset ([Xu et al., 2024]) [3]. For this problem, our linear minimisation oracle is solving a shortest paths problem on the network, with weights given by the current gradient (Mitradjieva and Lindberg [2013]). In our experiments, these shortest path subproblems are solved using *Dijkstra's* algorithm, wich runs in $\mathcal{O}\big(|A| + n\log(n)\big)$, where n is the number of nodes of the network.

[3] https://github.com/bstabler/TransportationNetworks



(a) training data

(b) shift test

Figure 3: example of Traffic Assignment Problem on "SiouxFalls". Purple histograms represent the losses obtained with the robust solution, green histograms represent the losses obtained with the ERM.

Figures 3a and 3b illustrate that on the training data, the ERM approach outperforms the robust method, achieving an overly optimistic low travel time. Conversely, the WDRO objective function leads to much higher total travel times on the training set, of the order of 2.44×10^5. On the shifted test data, the ERM severely degrades and underperforms, with an average loss of 1.92×10^6, while the robust solution remains stable and achieves a lower out-of-sample travel time, with an average objective of 1.04×10^6. In the next figure, we visualize the routing profiles yielded by the ERM and the robust solutions under two specific instances: a training scenario and an unseen out-of-sample test scenario:

(a) shift ERM flow

(b) shift robust flow

Figure 4: example on "SiouxFalls". The two solutions are illustrated on a scenario generated with a distributional shift. On each arc, a red shade indicates a high congestion over capacity ratio, while a green shade indicates that the considered link is uncongested.

Figures 4a and 4b illustrate the behavior of the ERM solution (Figure 4a) and the robust solution (Figure 4b) on a common instance generated using a shifted distribution. It can be seen that the robust solution alleviates congestion on many arcs in the city center, providing a solution with a reduced objective value.

Discussion on the experiments. One of the empirical findings of our experiments is that the effects of robustness become apparent when the shifted distribution is far enough from the empirical distribution $\widehat{\mathbb{P}}_N$, for example, where the training set is small or exhibits limited variability. In practice, robustness emerges when the variance of the sampling scenarios is large enough to induce different behaviors when evaluating our gradient estimators. When evaluating the losses, the proposed WDRO framework is particularly relevant in settings where distributional shifts may significantly alter the underlying decision problem.

5 Conclusion

This paper introduced a unified framework for data-driven stochastic combinatorial optimization with unknown distribution. By employing an entropic regularization scheme, we developed a momentum-based stochastic Frank-Wolfe algorithm to handle large-scale combinatorial constraints through a Linear Minimization Oracle. Our numerical experiments on the Quadratic Minimum Spanning Tree Problem and Traffic Assignment Problem illustrate that the proposed approach is computationally tractable but also provides robustness against distribution shifts. Our results also highlight the benefits of the WDRO framework for optimization problems that are loosely constrained, while ERM and WDRO are likely to give similar solutions to very constrained problems, suggesting that the greater the degrees of freedom in the decision space, the more useful the robust approach is. In future work, we aim to adapt this framework to two-stage distributionally robust problems, *i.e.* when the loss function itself is a value of a sub-problem, *e.g.* facility location.

6 Acknowledgments

This work was supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025) and by MIAI @ Grenoble Alpes (ANR-19-P3IA-0003), whose support is gratefully acknowledged.

References

Arjang Assad and Weixuan Xu. The quadratic minimum spanning tree problem. *Naval Research Logistics*, 1992.

Waïss Azizian, Franck Iutzeler, and Jérôme Malick. Regularization for Wasserstein distributionally robust optimization. *ESAIM: Control, Optimisation and Calculus of Variations*, 2023.

Beste Basciftci, Shabbir Ahmed, and Siqian Shen. Distributionally robust facility location problem under decision-dependent stochastic demand. *European Journal of Operational Research*, 2021.

Besançon and Kurtz. A Frank-Wolfe algorithm for oracle-based robust optimization. *arXiv preprint arXiv:2411.19848*, 2024.

Mathieu Besançon, Alejandro Carderera, and Sebastian Pokutta. FrankWolfe.jl: A high-performance and flexible toolbox for Frank-Wolfe algorithms and conditional gradients. *INFORMS Journal on Computing*, 34(5):2611–2620, 2022.

Mathieu Besançon, Sébastien Designolle, Jannis Halbey, Deborah Hendrych, Dominik Kuzinowicz, Sebastian Pokutta, Hannah Troppens, Daniel Viladrich Herrmannsdoerfer, and Elias Wirth. Improved algorithms and novel applications of the FrankWolfe.jl library. *ACM Transactions on Mathematical Software*, 51(4):1–33, 2025.

Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

Vladimir Igorevich Bogachev and Maria Aparecida Soares Ruas. *Measure theory*, volume 1. Springer, 2007.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

Gábor Braun, Alejandro Carderera, Cyrille W. Combettes, Hamed Hassani, Amin Karbasi, Aryan Mokhtari, and Sebastian Pokutta. Conditional gradient methods, 2025. URL `https://arxiv.org/abs/2211.14103`.

Vincent Tsz Fai Chow, Zheng Cui, and Daniel Zhuoyu Long. Target-oriented distributionally robust optimization and its applications to surgery allocation. *INFORMS Journal on Computing*, 2022.

Frank de Meijer, Melanie Siebenhofer, Renata Sotirov, and Angelika Wiegele. Spanning and splitting: Integer semidefinite programming for the quadratic minimum spanning tree problem. *European Journal of Operational Research*, 2025.

John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55, 2019.

Adam N Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1):9–26, 2022.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3 (1-2):95–110, 1956.

Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

Shubhechyya Ghosal and Wolfram Wiesemann. The distributionally robust chance-constrained vehicle routing problem. *Operations Research*, 68(3):716–732, 2020.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *International conference on machine learning*, pages 427–435. PMLR, 2013.

Anton J Kleywegt, Alexander Shapiro, and Tito Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on optimization*, 12(2):479–502, 2002.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pages 130–166. Informs, 2019.

Tam Le and Jérôme Malick. Universal generalization guarantees for Wasserstein distributionally robust models. *arXiv preprint*, 2024.

Jean-François Le Gall. *Measure theory, probability, and stochastic processes*. Springer, 2022.

Maria Mitradjieva and Per Olov Lindberg. The stiff is moving—conjugate direction Frank-Wolfe methods with applications to traffic assignment. *Transportation Science*, 47(2):280–293, 2013.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 2018.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Michael Patriksson. *The traffic assignment problem: models and methods*. Courier Dover Publications, 2015.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11, 2019.

Alexander Schrijver. *Combinatorial optimization*. Springer, 2002.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

Luying Sun, Weijun Xie, and Tim Witten. Distributionally robust fair transit resource allocation during a pandemic. *Transportation science*, 2023.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Vincent, Azizian, Iutzeler, and Malick. `skwdro`: a library for Wasserstein distributionally robust machine learning. *arXiv preprint*, 2024.

Xiaotong Xu, Zhenjie Zheng, Zijian Hu, Kairui Feng, and Wei Ma. A unified dataset for the city-scale traffic assignment model in 20 us cities. *Scientific data*, 11(1):325, 2024.

A Properties of the empirical risk minimization paradigm

In this section, we aim to show some of the properties of (ERM) under the assumption that the considered combinatorial problem is a pure integer problem with a finite number of solutions, *i.e.* the $\mathcal{Z}$ defined in (1) is a finite subset of $\mathbb{Z}^n$.

Theorem 2 formalizes the observation that, unlike the unconstrained setting, (ERM) for constrained discrete stochastic optimization can perform well in generalization in the constrained setting with a finite $\mathcal{Z}$, in a spirit similar to (Kleywegt et al. [2002]). We then extend this to the performance of ERM under a moderate distributional shift in Theorem 3.

Theorem 2 (ERM generalizes in finite sets)**.** *Let $\mathcal{Z}$ be a finite set of points. Let Assumption 1 hold, thus there exists $B > 0$ such that*

$$|f(\mathbf{z},\xi)| \le B \qquad \forall \mathbf{z} \in \mathcal{Z}, \xi \in \Xi.$$

Let $\mathbf{P}$ be an unknown probability distribution over Ξ and let $\widehat{\xi}_1,\dots,\widehat{\xi}_N$ be drawn i.i.d. from $\mathbf{P}$. Define the true risk and empirical risk, respectively, as

$$\mathbf{F}(\mathbf{z}) \stackrel{def}{=} \mathbb{E}_{\xi\sim\mathbf{P}}[f(\mathbf{z},\xi)], \qquad \textit{and} \qquad \widehat{\mathbf{F}}_N(\mathbf{z}) \stackrel{def}{=} \frac{1}{N}\sum_{k=1}^{N} f(\mathbf{z},\widehat{\xi}_k).$$

Define the ERM value $\mathbf{F}_{\text{ERM}}$ and the true optimal value $\mathbf{F}^\star$ as the minimal value of $\widehat{\mathbf{F}}_N$, and $\mathbf{F}$ respectively. Then, for a threshold $\delta \in (0,1)$, with probability at least $1-\delta$:

$$\mathbf{F}_{\text{ERM}} - \mathbf{F}^\star \le 2B\sqrt{\frac{\log\left(2|\mathcal{Z}|/\delta\right)}{2N}}. \tag{28}$$

Proof. Fix any $\mathbf{z} \in \mathcal{Z}$, the random variables $\{f(\mathbf{z},\widehat{\xi}_i)\}_{i\in[\![N]\!]}$ are i.i.d. with common expectation $\mathbf{F}(z)$ and take values in the interval $[-B,B]$. For any $t>0$, we can apply Hoeffding's inequality (Hoeffding [1963]):

$$\mathbb{P}\left(|\widehat{\mathbf{F}}_N(\mathbf{z}) - \mathbf{F}(\mathbf{z})| \ge t\right) \le 2\exp\left(-\frac{Nt^2}{2B^2}\right).$$

By applying the union bound to the event of a deviation t between the ERM and true values over all $\mathbf{z} \in \mathcal{Z}$,

$$\begin{aligned}\mathbb{P}\left(\sum_{\mathbf{z}\in\mathcal{Z}} |\widehat{\mathbf{F}}_N(\mathbf{z}) - F(z)| \ge t\right) &\le \sum_{\mathbf{z}\in\mathcal{Z}} \mathbb{P}\left(|\widehat{\mathbf{F}}_N(\mathbf{z}) - \mathbf{F}(\mathbf{z})| \ge t\right) \\ &\le \sum_{\mathbf{z}\in\mathcal{Z}} 2\exp\left(-\frac{Nt^2}{2B^2}\right) = 2|\mathcal{Z}|\exp\left(-\frac{Nt^2}{2B^2}\right).\end{aligned} \tag{29}$$

To upper-bound the right-hand side by δ, we obtain:

$$2|\mathcal{Z}|\exp\left(-\frac{Nt^2}{2B^2}\right) \le \delta \quad \Rightarrow \quad t \ge B\sqrt{\frac{2\log\left(2|\mathcal{Z}|/\delta\right)}{N}}.$$

Define the uniform convergence event:

$$\mathscr{E} = \left\{\sup_{\mathbf{z}\in\mathcal{Z}} |\widehat{\mathbf{F}}_N(\mathbf{z}) - \mathbf{F}(\mathbf{z})| \le B\sqrt{\frac{2\log\left(2|\mathcal{Z}|/\delta\right)}{N}}\right\} \tag{30}$$

From complementing the event in the probability in (29), we obtain $\mathbb{P}(\mathscr{E}) \ge 1-\delta$.

Denote with $\mathbf{z}_{\text{ERM}}$ an optimizer of $\widehat{\mathbf{F}}_N$ over $\mathcal{Z}$: $\widehat{F}_N(\mathbf{z}_{\text{ERM}}) = \mathbf{F}_{\text{ERM}}$, and with $\mathbf{z}^\star$ an optimizer of $\mathbf{F}$ over $\mathcal{Z}$. We have with probability $1-\delta$:

$$\mathbf{F}(\mathbf{z}_{\text{ERM}}) \le \widehat{\mathbf{F}}_N(\mathbf{z}_{\text{ERM}}) + B\sqrt{\frac{2\log\left(2|\mathcal{Z}|/\delta\right)}{N}} \tag{31}$$

$$\le \widehat{\mathbf{F}}_N(\mathbf{z}^\star) + B\sqrt{\frac{2\log\left(2|\mathcal{Z}|/\delta\right)}{N}} \tag{32}$$

$$\le \mathbf{F}(\mathbf{z}^\star) + 2B\sqrt{\frac{2\log\left(2|\mathcal{Z}|/\delta\right)}{N}}. \tag{33}$$

Inequalities (31) and (33) are obtained from the uniform convergence event valid at $\mathbf{z}_{\text{ERM}}$ and $\mathbf{z}^\star$ respectively, and (32) comes from $\mathbf{z}_{\text{ERM}}$ being a minimizer of $\widehat{\mathbf{F}}_N$. Rearranging provides (28) with probability $1-\delta$. □

We now state a technical lemma akin to the Kantorovich–Rubinstein theorem but without the requirement for the ground cost c to be a metric; this is closely related, *e.g.*, to the weak duality result of (Blanchet and Murthy [2019]).

Lemma 5 (Lipschitz bound on expectations)**.** *Under Assumption 4, suppose $h : \Xi \to \mathbb{R}$ satisfies $|h(\xi) - h(\xi')| \le \mathbf{L}_h \cdot c(\xi, \xi')$ for all $\xi, \xi' \in \Xi$. Then for any two distributions $\mathbb{Q}_1, \mathbb{Q}_2$ over Ξ:*

$$\left|\mathbb{E}_{\xi\sim\mathbb{Q}_1}[h(\xi)] - \mathbb{E}_{\xi'\sim\mathbb{Q}_2}[h(\xi')]\right| \le \mathbf{L}_h \cdot W_c(\mathbb{Q}_1, \mathbb{Q}_2).$$

Proof. Let π be a coupling, *i.e.* a joint distribution over $\Xi \times \Xi$ with marginals $\mathbb{Q}_1$ and $\mathbb{Q}_2$, then

$$\mathbb{E}_{\xi\sim\mathbb{Q}_1}[h(\xi)] - \mathbb{E}_{\xi'\sim\mathbb{Q}_2}[h(\xi')] = \mathbb{E}_{(\xi,\xi')\sim\pi}[h(\xi) - h(\xi')] \le \mathbb{E}_{(\xi,\xi')\sim\pi}[|h(\xi) - h(\xi')|] \le \mathbf{L}_h \mathbb{E}_{(\xi,\xi')\sim\pi}[c(\xi,\xi')].$$

Since this holds for every coupling π, taking the infimum over all couplings gives

$$\mathbb{E}_{\xi\sim\mathbb{Q}_1}[h(\xi)] - \mathbb{E}_{\xi'\sim\mathbb{Q}_2}[h(\xi')] \le \mathbf{L}_h \inf_{\pi} \mathbb{E}_{(\xi,\xi')\sim\pi}[c(\xi,\xi')] = \mathbf{L}_h \cdot W_c(\mathbb{Q}_1, \mathbb{Q}_2).$$

Applying the same argument to $-h$ gives the two-sided bound. □

We can now extend the generalization bound to a performance guarantee of ERM under distributional shift.

Theorem 3 (ERM robustness to distribution shift)**.** *Retaining the notation of Theorem 2, suppose Assumption 4 holds and assume that the loss function $f(\mathbf{z}, \cdot)$ is $\mathbf{L}_f$-Lipschitz continuous with respect to the cost function c at any $\mathbf{z} \in \mathcal{Z}$:*

$$|f(\mathbf{z}, \xi) - f(\mathbf{z}, \xi')| \le \mathbf{L}_f c(\xi, \xi') \quad \forall \mathbf{z} \in \mathcal{Z}.$$

Let $\mathbb{Q}$ be a distribution over Ξ such that for some $\varrho > 0$, $W_c(\mathbb{P}, \mathbb{Q}) \le \varrho$. Define $\mathbf{F}_{\mathbb{Q}}(\mathbf{z})$, $\mathbf{F}(\mathbf{z})$ as the risk of under distribution $\mathbb{Q}$, $\mathbb{P}$ respectively and $\mathbf{z}_{\mathbb{Q}} \in \operatorname{argmin}_{\mathbf{z}\in\mathcal{Z}} \mathbf{F}_{\mathbb{Q}}(\mathbf{z})$. For any δ, with probability $1-\delta$, we have:

$$\mathbf{F}_{\mathbb{Q}}(\mathbf{z}_{\text{ERM}}) - \mathbf{F}_{\mathbb{Q}}(\mathbf{z}_{\mathbb{Q}}) \le 2B\sqrt{\frac{2\log\left(2|\mathcal{Z}|/\delta\right)}{N}} + 2\mathbf{L}_f\varrho. \tag{34}$$

Proof. For any fixed $\mathbf{z} \in \mathcal{Z}$, applying Lemma 5 with distributions $\mathbb{Q}, \mathbb{P}$ with $W_c(\mathbb{P}, \mathbb{Q}) \le \varrho$ and $h(\xi) = f(\mathbf{z}, \xi)$,

$$|\mathbf{F}_{\mathbb{Q}}(\mathbf{z}) - \mathbf{F}(\mathbf{z})| = |\mathbb{E}_{\xi\sim\mathbb{Q}}[f(\mathbf{z},\xi)] - \mathbb{E}_{\xi'\sim\mathbb{P}}[f(\mathbf{z},\xi')]| \le \mathbf{L}_f \cdot W_c(\mathbb{P}, \mathbb{Q}) \le \mathbf{L}_f\varrho \quad \forall \mathbf{z} \in \mathcal{Z}. \tag{35}$$

Using the uniform convergence event $\mathscr{E}$ from Theorem 2:

$$\mathscr{E} = \left\{ \sup_{\mathbf{z}\in\mathcal{Z}} |\widehat{\mathbf{F}}_N(\mathbf{z}) - \mathbf{F}(\mathbf{z})| \le B\sqrt{\frac{2\log\left(2|\mathcal{Z}|/\delta\right)}{N}} \right\} \tag{36}$$

holding with probability $1-\delta$, we can derive:

$$\mathbf{F}_{\mathbb{Q}}(\mathbf{z}_{\mathrm{ERM}}) \leq \mathbf{F}(\mathbf{z}_{\mathrm{ERM}}) + \mathbf{L}_f \varrho \tag{37a}$$

$$\leq \widehat{\mathbf{F}}_N(\mathbf{z}_{\mathrm{ERM}}) + B\sqrt{\frac{2\log\left(2|\mathcal{Z}|/\delta\right)}{N}} + \mathbf{L}_f \varrho \tag{37b}$$

$$\leq \widehat{\mathbf{F}}_N(\mathbf{z}_{\mathbb{Q}}) + B\sqrt{\frac{2\log\left(2|\mathcal{Z}|/\delta\right)}{N}} + \mathbf{L}_f \varrho \tag{37c}$$

$$\leq \mathbf{F}(\mathbf{z}_{\mathbb{Q}}) + 2B\sqrt{\frac{2\log\left(2|\mathcal{Z}|/\delta\right)}{N}} + \mathbf{L}_f \varrho \tag{37d}$$

$$\leq \mathbf{F}_{\mathbb{Q}}(\mathbf{z}_{\mathbb{Q}}) + 2B\sqrt{\frac{2\log\left(2|\mathcal{Z}|/\delta\right)}{N}} + 2\mathbf{L}_f \varrho \tag{37e}$$

where (37a) comes from applying (35) at $\mathbf{z}_{\mathrm{ERM}}$, (37b) uses the uniform convergence event bound holding at $\mathbf{z}_{\mathrm{ERM}}$ for $\mathbb{P}$, (37c) uses optimality of $\mathbf{z}_{\mathrm{ERM}}$ for $\widehat{\mathbf{F}}_N$, (37c) applies the uniform convergence event bound, this time at $\mathbf{z}_{\mathbb{Q}}$, and (37e) uses (35) a second time at $\mathbf{z}_{\mathbb{Q}}$. □

The key take-away of Theorem 2 and Theorem 3 is that if the number of vertices of the feasible region remains polynomial in the dimension n with maximum degree k, we obtain a generalization bounded by a quantity $\mathcal{O}\left(\sqrt{k\log(n)}\right)$. This analysis heavily relies on the finiteness of $\mathcal{Z}$ and on the union bound, and it could be that mixed-integer sets do not provide any satisfactory bound.

B Numerical illustration – generation of instances

B.1 Instances for the Quadratic Minimum Spanning Tree Problem

For the *Uncertain Quadratic Minimum Spanning Tree Problem*, we generated our instances following the following protocol:

Algorithm 4 *Generation of Uncertain Quadratic Minimum Spanning Tree Problem instances*

Input: n the number of nodes, m the number of edges
  $G \leftarrow$ random graph on n vertices and m edges, generated by Erdős–Rényi's algorithm
  while G is not connected **do**
    $G \leftarrow$ Erdős–Rényi(n, m)
  end while
  Sample $\mu_G \in \mathbb{R}_+^{m\times m}$ random ▷ Base cost
  Sample $M \in \{0,1\}^{m\times m}$, with $M_{ij} \sim \mathcal{B}(0,7), \forall i,j \in [\![m]\!]$ ▷ 30% chance of missing data
  Build distribution $\mathbf{P}_G$ defined as
    function $\mathbf{P}_G$:
      Sample $\mathbf{C}$ with $\mathbf{C}_{ij} \sim \mathcal{N}(0,1)$ ▷ pertubation
      $N \leftarrow M \otimes \left(\mu_G + 0,1 \times \mathbf{C}\right)$ ▷ Base cost + noise + missing data
      return $N/\|N\|_2$ ▷ normalized to facilitate comparison across instances
    end
  return $(G, \mathbf{P}_G)$

Once an instance is generated, we are given a graph G on n vertices and m edges, along with a "true" probability $\mathbf{P}_G$ from which we generate training data

$$\widehat{\xi}_1, \ldots, \widehat{\xi}_{N_{\mathrm{train}}} \sim \mathbf{P}_G.$$

B.2 Instances for the Traffic Assignment Problem

The instances used for the Traffic Assignment Problem experiments comes from the *TransportationNetwork* dataset[4]. Each instance provides a network $G = (V, A)$ with nominal capacities $c_a > 0$, free flow times $t_a^{(0)} > 0$

[4] https://github.com/bstabler/TransportationNetworks

for all $a \in A$, along with a ground multiplier α and a power parameter $\beta > 0$. The uncertain instances are then generated thanks to the following method:

Algorithm 5 *Generation of Uncertain Traffic Assignment Problem training instances*

Input: **Inst** $= \left(G = (V, A), (c_a)_{a\in A}, (t_a^{(0)})_{a\in A}, \alpha, \beta\right)$ the nominal instance
  Uncertainty parameters:
    $\sigma_\alpha^2 > 0$ ▷ controls the variance of the uncertainty on the multiplier
    B_β>0 ▷ length of the sampling interval for the power parameter
    μ_α, μ_β: shifting the training parameters ▷ $\mu_\alpha, \mu_\beta > 0$ shifts towards an optimistic setting
  Build distribution $\mathbf{P_{Inst}}$ defined as
    function $\mathbf{P_{Inst}}$:
      for $a \in A$ **do**
        Sample $m_a \sim \mathscr{U}([0.75, 1])$ ▷ change the free flow time of arc a
      end for
      Sample $\widetilde{\alpha} > 0$ with $\widetilde{\alpha} \sim \mathscr{N}\left(\alpha - \mu_\alpha, \sigma_\alpha^2\right)$ ▷ normal perturbation
      Sample $\widetilde{\beta} > 0$ with $\widetilde{\beta} \sim \mathscr{U}\left(\left[\left(\beta - \mu_\beta\right) \pm B_\beta/2\right]\right)$ ▷ uniform distribution for the power
      return $\mathbf{Inst}_{\text{sample}} = \left(G = (V, A), (c_a)_{a\in A}, (m_\alpha \times t_a^{(0)})_{a\in A}, \widetilde{\alpha}, \widetilde{\beta}\right)$
    end
  return $\mathbf{P_{Inst}}$

This algorithm is designed to provide us with uncertain instances of the Traffic Assignment Problem, based on real-life scenarios, where the parameters of the objective function have a bias towards an optimistic paradigm. The capacities and the network stays however, unchanged. We optimize our robust objective and the ERM objective with training scenarios

$$\widehat{\xi}_1, \ldots, \widehat{\xi}_{N_{\text{train}}} \sim \mathbf{P_{Inst}}.$$

In contrast, the shifted scenarios are generated from a shifted distribution using the same method but with significantly more pessimistic parameters. Capacities are also slightly reduced. Typically, the results illustrated in Figures 3a and 3b and the illustration of Figures 4a and 4b are obtained with generating parameters $\sigma_\alpha^2 = 0.3 \times \alpha$, $B_\beta = 0.9$, $\mu_\alpha = 0.15 \times \alpha$, $\mu_\beta = 1$.

C Usefull proof

For sake of completness, we provide a proof of a useful lemma in convex analysis. **# to do**

Lemma 1. *Under Assumption 2 – (i), let* $\mathbb{Q} \in \text{Proba}(\Xi)$ *and* $g : \Xi \to \mathbb{R}$ *be a bounded* $\mathbb{Q}$*-mesurable function. Then*

$$\mathbb{E}_{\zeta\sim\mathbb{Q}^g}\left[g(\zeta)\right] \geq \log\left(\mathbb{E}_{\zeta\sim\mathbb{Q}}\left[\exp\left(g(\zeta)\right)\right]\right)$$

where $\mathbb{Q}^g$ *is the distribution defined by* $\mathrm{d}\mathbb{Q}^g(\zeta) \propto \exp\left(g(\zeta)\right)\mathrm{d}\mathbb{Q}(\zeta)$.

Proof of Lemma 2. Let us consider the function $t \mapsto \log\left(\mathbb{E}_{\zeta\sim\mathbb{Q}}\left[\exp\left(tg(\zeta)\right)\right]\right)$. This function is convex and differentiable with derivative

$$\partial_t \log\left(\mathbb{E}_{\zeta\sim\mathbb{Q}}\left[\exp\left(tg(\zeta)\right)\right]\right) = \frac{\mathbb{E}_{\zeta\sim\mathbb{Q}}\left[g(\zeta)\exp\left(tg(\zeta)\right)\right]}{\mathbb{E}_{\zeta\sim\mathbb{Q}}\left[\exp\left(tg(\zeta)\right)\right]} = \mathbb{E}_{\zeta\sim\mathbb{Q}^{tg}}\left[g(\zeta)\right]$$

with $\mathrm{d}\mathbb{Q}^{tg}(\zeta) \propto \exp\left(tg(\zeta)\right)\mathrm{d}\mathbb{Q}(\zeta)$. The function is thus above all its tangents, and in particular above the tangent at 1:

$$\log\left(\mathbb{E}_{\zeta\sim\mathbb{Q}}\left[\exp\left(tg(\zeta)\right)\right]\right) \geq \mathbb{E}_{\zeta\sim\mathbb{Q}^g}\left[g(\zeta)\right](t-1) + \log\left(\mathbb{E}_{\zeta\sim\mathbb{Q}}\left[\exp\left(g(\zeta)\right)\right]\right) \qquad \forall t \in \mathbb{R}.$$

This identity at $t = 0$ gives us

$$0 \geq \mathbb{E}_{\zeta\sim\mathbb{Q}^g}\left[g(\zeta)\right](0-1) + \log\left(\mathbb{E}_{\zeta\sim\mathbb{Q}}\left[\exp\left(g(\zeta)\right)\right]\right) \qquad \textit{i.e.} \qquad \mathbb{E}_{\zeta\sim\mathbb{Q}^g}\left[g(\zeta)\right] \geq \log\left(\mathbb{E}_{\zeta\sim\mathbb{Q}}\left[\exp\left(g(\zeta)\right)\right]\right).$$

□