First-Order Methods for Wasserstein Distributionally Robust Constrained Optimization

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

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

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

Abstract

We consider constrained problems in which input data are affected by errors. In such settings, Wasserstein distributionally robust optimization provides a principled framework to mitigate model risk by optimizing against worst-case data distributions within Wasserstein ambiguity sets. However, the numerical resolution of the resulting problems remains challenging, especially in constrained settings. In this paper, we provide a general, practical way to solve Wasserstein distributionally robust formulations in the presence of constraints. Our approach only requires a linear minimization oracle for the feasible set, and combines two key ingredients: (i) an entropic regularization of the distributionally robust value function, which makes it possible to compute stochastic gradient estimators, and (ii) a stochastic Frank–Wolfe algorithm, which minimizes the regularized robust objective while naturally handling constraints. We illustrate the method, its tractability, and its interests against empirical risk minimization, on two problems: the traffic assignment and the minimum quadratic spanning tree.

Keywords: Distributionally robust optimization, data-driven optimization, conditional gradients, constrained optimization

1 Introduction: decision, uncertainty, and examples

1.1 Context

We consider a general class of data-driven constrained optimization problems of the form

$$\min_{\mathbf{x}\in\mathcal{X}} \quad \mathbb{E}_{\xi}\left[f(\mathbf{x},\xi)\right] \tag{1}$$

where $\mathbf{x}$ is the decision variable lying in the feasible convex set $\mathcal{X}\subseteq\mathbb{R}^n$ and ξ is an uncertain scenario lying in a set of possible scenarios $\Xi\subseteq\mathbb{R}^d$. We assume that the distribution of ξ is unknown, and that we only have access to training data $\widehat{\xi}_1,\dots,\widehat{\xi}_N\in\Xi$. The standard data-driven approach consists in replacing the unknown distribution by the empirical distribution $\widehat{\mathbb{P}}_N \stackrel{\text{def}}{=} (\delta_{\widehat{\xi}_1}+\cdots+\delta_{\widehat{\xi}_N})/N$. This leads to Empirical Risk Minimization (ERM) which minimizes the average loss on the training dataset [Shapiro et al., 2021, Kleywegt et al., 2002]:

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

While simple and widely used, ERM may produce over-optimistic decisions that overfit the observed scenarios and perform poorly on new data, especially when the training samples are limited, noisy, or unrepresentative (see

e.g. [Duchi and Namkoong, 2019, Mohajerin Esfahani and Kuhn, 2018]). Distributionally robust optimization addresses this limitation by replacing the single empirical distribution with a worst-case distribution chosen in an ambiguity set around the data. A particularly appealing way to define such an ambiguity set is through optimal transport. Given a ground cost function $c : \Xi \times \Xi \to \mathbb{R}_+$, the associated Wasserstein distance between two probability distributions $\mathbb{Q}_1$ and $\mathbb{Q}_2$ on Ξ is defined by

$$W_c(\mathbb{Q}_1, \mathbb{Q}_2) \stackrel{\text{def}}{=} \inf_{\substack{\pi \in \text{Prob}(\Xi \times \Xi) \\ [\pi]_1 = \mathbb{Q}_1,\, [\pi]_2 = \mathbb{Q}_2}} \int_{\Xi \times \Xi} c(\xi, \xi') \, \mathrm{d}\pi(\xi, \xi'), \tag{2}$$

where $[\pi]_1$ and $[\pi]_2$ denote the two marginals of the transport plan π; see *e.g.* [Peyré and Cuturi, 2019]. This leads to Wasserstein distributionally robust optimization (WDRO)

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

Here, $\rho \ge 0$ is the radius of the ambiguity set. This formulation combines expressive modeling flexibility (see *e.g.* [Mohajerin Esfahani and Kuhn, 2018]) and statistical guarantees (see *e.g.* [Le and Malick, 2025]). The inner supremum in (WDRO) is an optimization problem over an infinite-dimensional space of probability distributions, and is generally intractable in its primal form. Under mild assumptions, however, duality yields the reformulation [Mohajerin Esfahani and Kuhn, 2018, Blanchet and Murthy, 2019, Gao and Kleywegt, 2023]

$$\min_{\mathbf{x} \in \mathcal{X}} \inf_{\lambda \ge 0} \lambda \rho + \mathbb{E}_{\widehat{\xi} \sim \widehat{\mathbb{P}}_N} \left[\sup_{\zeta \in \Xi} \left\{ f(\mathbf{x}, \zeta) - \lambda c\left(\widehat{\xi}, \zeta\right) \right\} \right], \tag{3}$$

where λ is the dual variable associated with the constraint $W_c\left(\mathbb{Q}, \widehat{\mathbb{P}}_N\right) \le \rho$. This dual reformulation nicely removes the optimization over probability distributions, but the supremum over ζ typically induces the non-smoothness of the objective function. Thus, in general, solving (3) is challenging, and existing solution approaches typically rely on additional structural assumptions on both f and Ξ: indeed, in some specific situations, the supremum leads to convex optimization [Mohajerin Esfahani and Kuhn, 2018] or sometimes admits a closed-form reformulation (see an example in Appendix B), but in general the inner sup cannot be easily tackled.

When the constraint set $\mathcal{X}$ is difficult to handle (*e.g.* no compact algebraic representation or no easy-to-compute projection), the tractability of (3) is even more complicated. Existing distributionally robust approaches for such constrained problems are therefore mostly problem-specific: see *e.g.* applications for vehicle routing [Ghosal and Wiesemann, 2020], facility location [Basciftci et al., 2021], transit resource allocation [Sun et al., 2023], or surgery scheduling [Chow et al., 2022]. Among these, Ghosal and Wiesemann [2020] uses a Benders decomposition, where the supremum is computed over a set of critical scenarios iteratively updated by a separation oracle that identifies worst-case situations. Another example, Zhen et al. [2025] uses a standard Lagrangian reformulation applied to the inner maximization problem to yield a single minimization program.

The goal of our work is to propose a *general* method for Wasserstein distributionally robust constrained optimization, applicable to any problem with only a gradient oracle for the loss function and a linear minimization oracle for the feasible set. This requirement is satisfied for many important feasible sets (*e.g.* flow polytopes, matching polytopes, matroid polytopes, the Birkhoff polytope, spectrahedra), opening many applications in operations research (network routing, matching, matrix completion, or submodular optimization); see the survey of Braun et al. [2025]. In this paper, we illustrate our approach with the following two running examples.

Example 1 (Traffic Assignment– see *e.g.* Patriksson [2015]). On a transportation network with origin-destination travel demands, the objective is to determine the flow on each arc while accounting for congestion. The mathematical formulation is described in (5), and its linear minimization oracle is described in Example 5. ■

Example 2 (Quadratic Minimum Spanning Tree– see *e.g.* Assad and Xu [1992]). For a given graph, this 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 induced by selecting pairs of edges. The problem is further formulated in (7), and its linear minimization oracle is described in Example 6. ■

1.2 Contributions and outline

This paper presents a practical approach to data-driven stochastic constrained optimization based on Wasserstein distributional robustness. As discussed above, two main computational difficulties arise in this setting: handling

the worst-case scenario problem induced by the Wasserstein ambiguity set, and optimizing the resulting objective function over a hard-to-handle feasible region. We address these challenges simultaneously by combining two existing ideas (namely (i) entropic smoothing of the distributionally robust objective and (ii) conditional gradient methods for constrained optimization) into a unified and flexible framework. This algorithmic framework does not require tailoring to specific loss functions, uncertainty models, or algebraic representations of feasible sets.

Starting from the dual WDRO formulation in (3), we first replace the inner sample-wise supremum by an entropic log-sum-exp approximation, following Azizian et al. [2023], Gao and Kleywegt [2023], Florian et al. [2026]. This yields the smoothed WDRO problem

$$\min_{\mathbf{x}\in\mathcal{X}}\inf_{\lambda\geq 0} \quad F(\mathbf{x},\lambda) \stackrel{\text{def}}{=} \lambda\varrho + \varepsilon\mathbb{E}_{\widehat{\xi}\sim\widehat{\mathbb{P}}_N}\left[\log\left(\mathbb{E}_{\zeta\sim\mathcal{N}(\widehat{\xi},\sigma^2\mathbf{I})}\left[\exp\left(\frac{f(\mathbf{x},\zeta)-\lambda c\left(\widehat{\xi},\zeta\right)}{\varepsilon}\right)\right]\right)\right], \tag{4}$$

where $\varepsilon > 0$ controls the smoothing level and σ^2 determines the sampling variance. This smoothing has both an algorithmic and a modeling role. Algorithmically, it turns the non-smooth dual objective into a differentiable surrogate amenable to stochastic first-order schemes. From a modeling viewpoint, it can be interpreted as an entropic regularization of WDRO, while preserving the statistical guarantees of the Wasserstein formulation (see [Azizian et al., 2023, Le and Malick, 2025]).

We then propose to solve (4) with a momentum stochastic Frank-Wolfe scheme (see *e.g.* [Mokhtari et al., 2020, Braun et al., 2025]) which only requires access to a linear minimization oracle over the feasible set. This makes the method applicable to a broad class of constrained problems without deriving problem-specific robust reformulations. We then provide arguments to make the smoothed WDRO formulation algorithmically analyzable. Although the dual formulation involves an unbounded multiplier $\lambda \in \mathbb{R}_+$, we prove that this apparent non-compactness is harmless, as the dual search can be explicitly restricted to a compact interval $[0,\bar{\lambda}]$. Then, we show that the stochastic gradients of the robust objective F have uniformly bounded variance. These two estimates provide the technical arguments of the standard convergence analysis: compactness controls the size of the feasible set, while the variance bound controls the stochastic noise.

We provide numerical illustrations on two problems where existing methods do not apply, namely Traffic Assignment and Quadratic Minimum Spanning Tree. We recover in this constrained setting the behavior of the unconstrained setting [Mohajerin Esfahani and Kuhn, 2018]: ERM tends to over-promise and under-perform, whereas Wasserstein distributionally robust modeling under-promises and over-performs on new test data.

This paper is organized as follows. Section 2 introduces the smoothed WDRO model, its assumptions, and the associated gradient estimators. We establish the main regularity properties of this surrogate and derive Monte Carlo gradient estimators adapted to its structure. Section 3 presents the stochastic Frank-Wolfe algorithm and its convergence guarantees. Section 4 illustrates the robustness obtained by our model on the two running problems. Finally, four Appendices provide complementary informations (generation of instances, special cases, auxiliary results on ERM, and two technical results).

2 Smooth distributionally robust framework

In this section, we present the smooth WDRO framework, adapted from Azizian et al. [2023], Gao and Kleywegt [2023], Florian et al. [2026] to treat the general constrained case. Section 2.1 presents our assumptions on the objective function, the constraints and scenario sets, and on the ground cost. Section 2.2 discusses the properties of the smoothed robust function and its gradient, key for the first-order method considered in Section 3.

2.1 Assumptions and examples

This section presents the assumptions that will be made throughout the paper.

Assumption 1 (On the original objective function f)**.**

(i) For any scenario $\xi \in \Xi$, the function $\mathbf{x} \mapsto f(\mathbf{x},\xi)$ is a $\mathscr{C}^2$ convex function.

(ii) For any decision $\mathbf{x} \in \mathcal{X}$, the function $\xi \mapsto \nabla_{\mathbf{x}} f(\mathbf{x},\xi)$ is continuous on Ξ.

Since our approach will rely on gradient-based methods, we thus assume that the loss function f is sufficiently smooth with respect to the decision variable. This assumption is relatively general and satisfied by a wide range of applications. In particular, it is verified by our two running examples.

Example 3 (Traffic Assignment). For a network $G = (V, A)$, with origin-destination travel demands, the objective of the Traffic Assignment is to determine the flow on each arc while accounting for congestion. In our case, the travel time is dependent on uncertain exogenous factors (such as vehicle fleet composition or weather conditions), and we make our decision based on a finite amount of observations. For each $(i, j) \in V^2$, denote $q_{i\to j} \geq 0$ the traffic demand between i and j, and $\mathscr{P}_{i\to j}$ the set of paths connecting the origin i to the destination j. For a determinist scenario $\xi = ((t_a^{(0)})_{a\in A}, \alpha, \beta)$, the Traffic Assignment writes:

$$\underset{\mathbf{x}\in\mathcal{X}_{\text{flow}}}{\text{minimize}} \quad f(\mathbf{x},\xi) = \sum_{a\in A} \int_0^{\mathbf{x}_a} t_a^{(0)} \left(1 + \alpha\left(\frac{s}{c_a}\right)^{\beta}\right) \mathrm{d}s \tag{5}$$

where $t_a^{(0)} > 0$ denotes the free-flow travel time, $c_a > 0$ is the capacity of arc a, and $\alpha, \beta > 0$ are dimensionless tuning parameters, and $\mathcal{X}_{\text{flow}}$ is the flow-polytope defined by:

$$\mathcal{X}_{\text{flow}} \overset{\text{def}}{=} \left\{ \mathbf{x} \in \mathbb{R}_+^{|A|} \;\middle|\; \begin{array}{ll} \sum\limits_{p\in\mathscr{P}_{i\to j}} \sum\limits_{a\in p} f_p \mathbb{1}_p(a) = q_{i\to j} & \forall (i,j) \in V^2, \\ \sum\limits_{(i,j)\in V^2} \sum\limits_{p\in\mathscr{P}_{i\to j}} f_p = \mathbf{x}_a & \forall a \in A, \\ f_p \geq 0 & \forall p \in \mathscr{P}_{i\to j}, \forall (i,j) \in V^2 \end{array} \right\}. \tag{6}$$

The objective function is $\mathscr{C}^2$, and its gradient with respect to $\mathbf{x}_a$ is

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

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

Example 4. *Quadratic Minimum Spanning Tree.* We consider the quadratic spanning tree problem over a graph $G = (V, E)$, which includes bilinear cost terms for the selection of some pairs of edges. In our case, the costs are not known in advance, and we instead have access to a finite number of training samples. Here, the relaxed problem for a given cost scenario $\xi \in \Xi \subset \mathbb{R}_+^{m\times m}$ is written as

$$\underset{\mathbf{x}\in\mathcal{X}_{\text{tree}}}{\text{minimize}} \quad f(\mathbf{x},\xi) = \mathbf{x}^\top \xi \mathbf{x} \tag{7}$$

where $\mathcal{X}_{\text{tree}}$ is the tree-polytope, defined by:

$$\mathcal{X}_{\text{tree}} \overset{\text{def}}{=} \left\{ \mathbf{x} \in [0,1]^m \;\middle|\; \sum_{i=1}^m \mathbf{x}_i = n-1, \sum_{i\in E(S)} \mathbf{x}_i \leq |S| - 1, \forall S \subseteq V \right\}, \tag{8}$$

with $E(S)$ the set of edges adjacent to vertices in S. The function $f(\cdot,\xi)$ is quadratic, thus $f(\cdot,\xi) \in \mathscr{C}^2$. Its derivative with respect to $\mathbf{x}$ is given by $\mathbf{x} \mapsto (\xi + \xi^\top)\mathbf{x}$, and thus for any fixed spanning tree embedding $\mathbf{x} \in \mathcal{X}$, the function $\xi \mapsto \nabla_{\mathbf{x}} f(\mathbf{x},\xi)$ is a continuous map. Thus f verifies the conditions of Assumption 1. ■

On the uncertain parameter space, we further assume that the set of all plausible scenarios Ξ is closed and bounded. Moreover, we suppose that the decision space itself is compact, as is the case for our two running problems, as well as a wide variety of problems; see *e.g.* Bertsimas et al. [2011] for a more detailed survey.

Assumption 2 (On the constraints and scenario sets)**.**

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

(ii) The set $\mathcal{X}$ is compact.

Note that the following quantity (which will appear in our analysis) is well-defined and finite, under Assumptions 1 and 2:

$$\|f\|_\infty \overset{\text{def}}{=} \sup_{\mathbf{x}\in\mathcal{X}, \xi\in\Xi} |f(\mathbf{x},\xi)|.$$

We also assume that the ground cost is comparable to the squared Euclidean norm. This controls the cost terms that appear in the dual variable and in the gradient of the regularized WDRO objective. This assumption is not used to apply the developments of this paper, but it is convenient to establish the mathematical properties and the convergence analysis. It is satisfied by many ground costs, including all the squared p-norm costs.

Assumption 3 (On the ground cost)**.** There exist constants $\mu_c, \mathbf{L}_c > 0$ such that the ground cost $c : \Xi \times \Xi \to \mathbb{R}_+$ satisfies

$$\mu_c \|\xi - \zeta\|_2^2 \le c(\xi, \zeta) \le \mathbf{L}_c \|\xi - \zeta\|_2^2 \qquad \forall \xi, \zeta \in \Xi.$$

In particular, $c(\xi, \xi) = 0$ for all $\xi \in \Xi$.

2.2 Differentiability and gradient of robust objective

The distributionally robust objective function (4) has useful properties that allow the employment of first-order methods, which we detail in this section. These results are already present, more or less explicitly, in the literature, in particular Azizian et al. [2023], Florian et al. [2026]. We summarize the properties needed below and give a short proof for completeness.

Proposition 1 (Regularized WDRO objective)**.** *Under Assumptions 1 to 3, the function F defined in* (4) *is convex and twice-differentiable on $\mathcal{X} \times \mathbb{R}_+$. Moreover,*

$$\nabla_{\mathbf{x}} F(\mathbf{x}, \lambda) = \mathbb{E}_{\widehat{\xi} \sim \widehat{\mathbb{P}}_N} \left[\mathbb{E}_{\zeta \sim \boldsymbol{\pi}_{\mathbf{x},\lambda}(\cdot|\widehat{\xi})} \left[\nabla_{\mathbf{x}} f(\mathbf{x}, \zeta) \right] \right], \tag{9}$$

$$\frac{\partial}{\partial \lambda} F(\mathbf{x}, \lambda) = \varrho - \mathbb{E}_{\widehat{\xi} \sim \widehat{\mathbb{P}}_N} \left[\mathbb{E}_{\zeta \sim \boldsymbol{\pi}_{\mathbf{x},\lambda}(\cdot|\widehat{\xi})} \left[c(\widehat{\xi}, \zeta) \right] \right]. \tag{10}$$

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

$$\mathrm{d}\boldsymbol{\pi}_{\mathbf{x},\lambda}(\cdot|\xi) \propto \exp\left(\left(f(\mathbf{x}, \zeta) - \lambda c(\xi, \zeta) \right) / \varepsilon \right) \exp\left(-\|\xi - \zeta\|_2^2 / 2\sigma^2 \right) \mathrm{d}\zeta. \tag{11}$$

Proof of Proposition 1. For $k \in [\![N]\!]$, set

$$h_k(\mathbf{x}, \lambda; \zeta) \overset{\text{def}}{=} f(\mathbf{x}, \zeta) - \lambda c(\widehat{\xi}_k, \zeta) / \varepsilon, \qquad G_k(\mathbf{x}, \lambda) \overset{\text{def}}{=} \log \mathbb{E}_{\zeta \sim \mathcal{N}(\widehat{\xi}_k, \sigma^2 \mathbf{I})} \left[\exp\left(h_k(\mathbf{x}, \lambda; \zeta) \right) \right].$$

For fixed ζ, the map $(\mathbf{x}, \lambda) \mapsto h_k(\mathbf{x}, \lambda; \zeta)$ is convex. The log-sum-exp operator preserves convexity; equivalently, this follows from Hölder's inequality applied to $\exp(h_k)$ along any segment in $\mathcal{X} \times \mathbb{R}_+$. Hence each G_k is convex, and so is F, as a nonnegative sum of the G_k plus the linear term $\lambda\varrho$. The compactness, smoothness, and cost-comparison assumptions allow differentiation under the integral sign for the exponential terms, and the normalizing denominator is strictly positive. Therefore,

$$\nabla_{\mathbf{x}} F(\mathbf{x}, \lambda) = \frac{1}{N} \sum_{k=1}^{N} \frac{\mathbb{E}_{\zeta \sim \mathcal{N}(\widehat{\xi}_k, \sigma^2 \mathbf{I})} \left[\nabla_{\mathbf{x}} f(\mathbf{x}, \zeta) \exp\left(h_k(\mathbf{x}, \lambda; \zeta) \right) \right]}{\mathbb{E}_{\zeta \sim \mathcal{N}(\widehat{\xi}_k, \sigma^2 \mathbf{I})} \left[\exp\left(h_k(\mathbf{x}, \lambda; \zeta) \right) \right]} = \mathbb{E}_{\widehat{\xi} \sim \widehat{\mathbb{P}}_N} \left[\mathbb{E}_{\zeta \sim \boldsymbol{\pi}_{\mathbf{x},\lambda}(\cdot|\widehat{\xi})} \left[\nabla_{\mathbf{x}} f(\mathbf{x}, \zeta) \right] \right],$$

and similarly,

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

Applying the same differentiation argument to these gradient expressions gives $F \in \mathscr{C}^2(\mathcal{X} \times \mathbb{R}_+)$. □

While equations (10) 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 $\zeta_1^{(k)}, \ldots, \zeta_S^{(k)}$ 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:

$$\mathbb{E}_{\zeta \sim \boldsymbol{\pi}_{\mathbf{x},\lambda}(\cdot|\widehat{\xi}_k)} \left[\varphi_k(\mathbf{x}, \zeta) \right] \approx \sum_{s=1}^{S} \varphi_k(\mathbf{x}, \zeta_s^{(k)}) \frac{w_s^{(k)}}{\|w^{(k)}\|_1} \tag{12}$$

where $\varphi_k(\mathbf{x}, \zeta_s^{(k)})$ is to be replaced by $\nabla_{\mathbf{x}} f(\mathbf{x}, \zeta_s^{(k)})$ (resp. $c(\widehat{\xi}_k, \zeta_s^{(k)})$) to match the terms of (9) and (10), and where $w_s^{(k)} \overset{\text{def}}{=} \exp\left(\left(f(\mathbf{x}, \zeta_s^{(k)}) - \lambda c(\widehat{\xi}_k, \zeta_s^{(k)}) \right) / \varepsilon \right)$ for $k \in [\![N]\!]$ and $s \in [\![S]\!]$. This approximation can be intractable for problems where the dimension of the scenario is high, as is the case for the Quadratic Minimum Spanning Tree[1]. Thus, we

[1] Computing the gradient estimates in (13) for the Quadratic Minimum Spanning Tree 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.

adopt a mini-batch gradient scheme by restricting the gradient computation to a small subset of indices $\mathcal{J} \subseteq [\![N]\!]$, following Florian et al. [2026]. Given a mini-batch $\mathcal{J} \subseteq [\![N]\!]$, and samples $\zeta_1^{(k)}, \ldots, \zeta_S^{(k)} \sim \mathcal{N}\left(\widehat{\xi}_k, \sigma^2 \mathbf{I}\right)$ for all $k \in \mathcal{J}$, we define our estimator of (9) and (10) by taking the external sum over $\mathcal{J}$ only, and approximate each inner expectation as (12); this gives the following estimates:

$$\mathscr{G}_{\mathbf{x}}^{\mathcal{J}}(\mathbf{x}, \lambda) \stackrel{\text{def}}{=} \frac{1}{|\mathcal{J}|} \sum_{k \in \mathcal{J}} \sum_{s=1}^{S} \nabla_{\mathbf{x}} f(\mathbf{x}, \zeta_s^{(k)}) \frac{w_s^{(k)}}{\|w^{(k)}\|_1} \qquad \mathscr{G}_{\lambda}^{\mathcal{J}}(\mathbf{x}, \lambda) \stackrel{\text{def}}{=} \rho - \frac{1}{|\mathcal{J}|} \sum_{k \in \mathcal{J}} \sum_{s=1}^{S} c(\widehat{\xi}_k, \zeta_s^{(k)}) \frac{w_s^{(k)}}{\|w^{(k)}\|_1}. \tag{13}$$

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

The resulting estimator is the practical object used by our first-order method. We record below the minimal consistency property needed for the stochastic Frank-Wolfe scheme introduced in Section 3.

Proposition 2. *Under Assumptions 1 and 2, the gradient estimator $\mathscr{G}^{\mathcal{J}}$ is asymptotically unbiased:*

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

Proof. We prove the result for the $\mathbf{x}$-component; the argument for the λ-component is identical. Here the expectation is taken over both the random mini-batch $\mathcal{J}$ and the Monte Carlo samples used to construct the estimator. Let $\mathscr{F}_S$ be the σ-algebra generated by all Monte Carlo samples $\left(\zeta_s^{(k)}\right)_{k \in [\![N]\!], s \in [\![S]\!]}$ and define

$$Y_{k,S} \stackrel{\text{def}}{=} \sum_{s=1}^{S} \nabla_{\mathbf{x}} f(\mathbf{x}, \zeta_s^{(k)}) \frac{w_s^{(k)}}{\|w^{(k)}\|_1}, \qquad k \in [\![N]\!],$$

which is deterministic conditionally on $\mathscr{F}_S$. Since the mini-batch $\mathcal{J}$ is sampled uniformly and independently of $\mathscr{F}_S$, the standard calculus gives (see Lemma 4 in Appendix D for a formal statements)

$$\mathbb{E}\left[\mathscr{G}_{\mathbf{x}}^{\mathcal{J}}(\mathbf{x}, \lambda)\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{1}{|\mathcal{J}|} \sum_{k \in \mathcal{J}} Y_{k,S} \,\middle|\, \mathscr{F}_S\right]\right] = \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\left[Y_{k,S}\right].$$

It remains to identify the limit of each local Monte Carlo estimator $Y_{k,S}$. Fix k and set

$$\omega_k(\zeta) \stackrel{\text{def}}{=} \exp\left(f(\mathbf{x}, \zeta) - \lambda c(\widehat{\xi}_k, \zeta)/\varepsilon\right).$$

The law of large numbers gives

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

The denominator limit is finite by the definition of F and strictly positive because $\omega_k > 0$ almost surely. Hence the continuous mapping theorem yields

$$Y_{k,S} \xrightarrow[S \to \infty]{\text{a.s.}} \frac{\mathbb{E}\left[\nabla_{\mathbf{x}} f(\mathbf{x}, \cdot) \omega_k(\cdot)\right]}{\mathbb{E}\left[\omega_k(\cdot)\right]} = \mathbb{E}_{\zeta \sim \boldsymbol{\pi}_{\mathbf{x},\lambda}(\cdot|\widehat{\xi}_k)}\left[\nabla_{\mathbf{x}} f(\mathbf{x}, \zeta)\right].$$

We now justify the passage from almost sure convergence to convergence of expectations. By Assumptions 1 and 2, the map $(\mathbf{x}, \zeta) \mapsto \nabla_{\mathbf{x}} f(\mathbf{x}, \zeta)$ is bounded on $\mathcal{X} \times \Xi$. Let

$$C_z \stackrel{\text{def}}{=} \sup_{\mathbf{x} \in \mathcal{X}, \zeta \in \Xi} \|\nabla_{\mathbf{x}} f(\mathbf{x}, \zeta)\| < +\infty.$$

Since $w_s^{(k)} > 0$ and $\sum_{s=1}^{S} w_s^{(k)}/\|w^{(k)}\|_1 = 1$, the estimator $Y_{k,S}$ is a convex combination of vectors whose norms are bounded by C_z. Hence, for every $S \geq 1$,

$$\|Y_{k,S}\| \leq \sum_{s=1}^{S} \|\nabla_{\mathbf{x}} f(\mathbf{x}, \zeta_s^{(k)})\| \frac{w_s^{(k)}}{\|w^{(k)}\|_1} \leq C_z.$$

Thus the family $(Y_{k,S})_{S\geq 1}$ is uniformly bounded in L^∞, and therefore uniformly integrable. Vitali's convergence theorem gives convergence of $\mathbb{E}[Y_{k,S}]$ for every k. Since N is finite,

$$\mathbb{E}\left[\mathscr{G}^{\mathcal{J}}_{\mathbf{x}}(\mathbf{x},\lambda)\right] \xrightarrow[S\to\infty]{} \frac{1}{N}\sum_{k=1}^{N} \mathbb{E}_{\zeta\sim\boldsymbol{\pi}_{\mathbf{x},\lambda}(\cdot|\widehat{\xi}_k)}\left[\nabla_{\mathbf{x}} f(\mathbf{x},\zeta)\right] = \nabla_{\mathbf{x}} F(\mathbf{x},\lambda).$$

The same argument applies to the λ-component. Since Ξ is compact and c is bounded on $\Xi\times\Xi$ under Assumption 3, the constant

$$C_c \overset{\text{def}}{=} \sup_{\xi,\zeta\in\Xi} c(\xi,\zeta)$$

is finite. The self-normalized cost average is bounded by C_c, and therefore $\mathscr{G}^{\mathcal{J}}_{\lambda}(\mathbf{x},\lambda)$ is bounded by $\varrho + C_c$, uniformly in S. This again gives uniform integrability and allows Vitali's theorem to identify the limit of the expectation. Consequently, both components converge to the corresponding components of $\nabla F(\mathbf{x},\lambda)$. □

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

3 First-order methods tackling WDRO

Optimizing our distributionally robust objective (4) over constrained sets involves managing the geometry of the feasible set. To do so, we propose to use a Frank-Wolfe algorithm, following recent success in both ML and OR communities [Jaggi, 2013, Besançon et al., 2022, Braun et al., 2025]. 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 Stochastic Frank-Wolfe at work

To address the constraints of the regularized WDRO problem (4), we adopt a *Frank-Wolfe* framework [Frank and Wolfe, 1956, Jaggi, 2013, Harchaoui et al., 2015]. This method is adapted to smooth functions over a constrained feasible set $\mathcal{X}$ which admits a *Linear Minimization Oracle* (LMO):

$$\mathrm{LMO}_{\mathcal{X}}(\mathbf{g}) \in \operatorname*{argmin}_{\mathbf{s}\in\mathcal{X}} \langle \mathbf{g}, \mathbf{s}\rangle. \tag{14}$$

For a wide variety of constrained problems, this linear subproblem is tractable and can be solved efficiently via specialized methods [Besançon et al., 2022, 2025]. It is the case for our two running examples.

Example 5 (Traffic Assignment). Optimizing over the flow polytope described in (5), 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. ■

Example 6 (Quadratic Minimum Spanning Tree). Over the spanning-tree polytope, the linear oracle associated with the quadratic objective reduces to minimizing its linearization at the current point. This is a linear optimization problem over the spanning-tree polytope and can therefore be solved exactly by computing a minimum spanning tree, which can be achieved in polynomial time via *Kruskal*'s algorithm. We underline that we have a highly tractable linear minimization oracle, even if the spanning-tree polytope is difficult to describe efficiently (indeed, Pokutta [2026] recently showed that every symmetric extended formulation for polytope (8) has $\mathcal{O}(n^3)$ inequalities). ■

The standard (stochastic) Frank-Wolfe algorithm is given in Algorithm 1: an iteration consists of a gradient estimation, a resolution of the LMO, and an update of the iterate in the obtained direction.

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

Input: step sizes $0 \leq \alpha_t \leq 1$
for $t = 0$ to $\ldots$ **do**
  $\widetilde{\nabla} F(y_t) \leftarrow$ gradient estimator
  $v_t \leftarrow \mathrm{LMO}_{\mathcal{X}}\left(\widetilde{\nabla} F(\mathbf{y}_t)\right)$ ▷ v_t is an optimal solution of the linear subproblem (14)
  $\mathbf{y}_{t+1} \leftarrow \mathbf{y}_t + \alpha_t\left(v_t - \mathbf{y}_t\right)$
end for

Since the iterates of Algorithm 2 are formed as convex combinations of the optimal vertices returned by the linear oracle, the algorithm computes a solution in $\mathcal{X}$. Depending on the application, this fractional solution can either be used directly (as in the continuous Traffic Assignment) or rounded via problem-specific heuristics[2].

Stochastic Frank-Wolfe methods (Algorithm 1) 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 propose here to use the version of Mokhtari et al. [2020], integrating a momentum term. This indeed provides a stable estimation of the descent direction by taking into account the previous gradient direction, and ensures 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: $\mathbf{x}_0 \in \mathcal{X}$, $\lambda_{\max}$ 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 $\dots$ **do**

  $\mathcal{J}_t \leftarrow$ random subset of $[\![N]\!]$ with $|\mathcal{J}_t| = b$

  for $k \in \mathcal{J}_t$ **do**

    sample $\zeta_1^{(k)}, \dots, \zeta_S^{(k)} \sim \mathcal{N}\left(\widehat{\xi}_k, \sigma^2 \mathbf{I}\right)$ i.i.d.

    $w_s^{(k)} \leftarrow \exp\left(\left(f(\mathbf{x}_t, \zeta_s^{(k)}) - \lambda_t c(\widehat{\xi}_k, \zeta_s^{(k)})\right)/\varepsilon\right)$

  end for

  $\mathcal{G}^{\mathcal{J}_t}(\mathbf{x}_t, \lambda_t) \leftarrow$ Estimate given by (13)

  $\widehat{\nabla} F(\mathbf{x}_t, \lambda_t) \leftarrow \beta_t \mathcal{G}_S^{\mathcal{J}_t}(\mathbf{x}_t, \lambda_t) + (1-\beta_t)\widehat{\nabla} F(\mathbf{x}_{t-1}, \lambda_{t-1})$

  $v_t \leftarrow \operatorname{argmin}_v \mathrm{LMO}\left(\widehat{\nabla} F(\mathbf{x}_t, \lambda_t)\right)$

  $(\mathbf{x}_{t+1}, \lambda_{t+1}) \leftarrow (\mathbf{x}_t, \lambda_t) + \alpha_t (v_t - (\mathbf{x}_t, \lambda_t))$

end for

This algorithm has been applied in other contexts to build convex relaxations to combinatorial and mixed-integer optimization problems [Hendrych et al., 2025].

Remark 1. A practical advantage of the sampling routine in Algorithm 2 is its ability to handle cases where the uncertainty support Ξ has known boundaries within $\mathbb{R}^d$. Indeed, a simple rejection sampling strategy can be integrated into the loop to avoid sampling infeasible instances. For instance, in the context of the Traffic Assignment, the components of $\zeta_s^{(k)}$ represent edge weights and must naturally satisfy non-negativity constraints (*i.e.*, $\Xi \subseteq \mathbb{R}_+^{m \times m}$): if a generated sample $\zeta_s^{(k)}$ has negative coefficients, it can simply be discarded and resampled.

3.2 Convergence analysis

This section analyzes the convergence properties of Algorithm 2 when applied to (4), or more precisely, to its convex relaxation. Specifically, Theorem 1 establishes the convergence of our method toward an optimal solution $(\mathbf{x}^\star, \lambda^\star)$ of the relaxed problem. Furthermore, the numerical experiments presented in Section 4 demonstrate that the solution $\mathbf{x}^\star$ offers robustness properties when evaluated on unseen out-of-sample data.

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 $\mathcal{X}$ is necessarily compact from Assumption 2 – (ii), the dual variable λ is defined over the interval $[0, +\infty)$. To ensure convergence of the algorithm, we must identify a sufficiently large upper bound $\lambda_{\max}$ such that the solution space can be restricted to the compact interval $[0, \lambda_{\max}]$ without losing the optimal solution. This is the purpose of Proposition 3; the auxiliary log-sum-exp inequality used in its proof is stated in Lemma 5 in Appendix D.

[2] Constructing an explicit discrete feasible solution from such a relaxation is a widely studied problem in operations research, for which numerous primal heuristics, rounding schemes, and relaxation-based reconstruction procedures are available. Our framework is complementary to these techniques: we focus on solving the constrained WDRO relaxation, while an application-specific primal heuristic can subsequently be employed whenever an integer solution is required.

Proposition 3 (Upper bound on λ)**.** *Let* $m_c \overset{def}{=} \max_{k\in[\![N]\!]} \mathbb{E}_{\zeta\sim\mathcal{N}(\widehat{\xi}_k,\sigma^2\mathbf{I})}\left[c(\widehat{\xi}_k,\zeta)\right]$ *and suppose that* $\varrho > m_c$. *Let* $\lambda_{\max} \overset{def}{=} 2\|f\|_\infty/(\varrho - m_c)$. *Then for all* $\mathbf{x}\in\mathcal{X}$*:*

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

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

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

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

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

where $\boldsymbol{\pi}_{\mathbf{x},\lambda}(\cdot|\widehat{\xi}_k)$ is the distribution defined in (11). We bound $-\partial_\lambda\varphi_\varepsilon$ uniformly in $\mathbf{x}\in\mathcal{X}$ and $\{\widehat{\xi}_k\}_{k\in[\![N]\!]}$. We have:

$$\begin{aligned}\mathbb{E}_{\zeta\sim\boldsymbol{\pi}_{\mathbf{x},\lambda}(\cdot|\widehat{\xi}_k)}\left[c(\widehat{\xi}_k,\zeta)\right] &= \mathbb{E}_{\zeta\sim\boldsymbol{\pi}_{\mathbf{x},\lambda}(\cdot|\widehat{\xi}_k)}\left[\frac{f(\mathbf{x},\zeta)-f(\mathbf{x},\zeta)+\lambda c(\widehat{\xi}_k,\zeta)}{\lambda}\right]\\ &= \frac{1}{\lambda}\mathbb{E}_{\zeta\sim\boldsymbol{\pi}_{\mathbf{x},\lambda}(\cdot|\widehat{\xi}_k)}\left[f(\mathbf{x},\zeta)\right] - \frac{\varepsilon}{\lambda}\mathbb{E}_{\zeta\sim\boldsymbol{\pi}_{\mathbf{x},\lambda}(\cdot|\widehat{\xi}_k)}\left[\frac{f(\mathbf{x},\zeta)-\lambda c(\widehat{\xi}_k,\zeta)}{\varepsilon}\right] \qquad (15)\end{aligned}$$

The first term in (15) is uniformly bounded by $\mathbb{E}_{\zeta\sim\boldsymbol{\pi}_{\mathbf{x},\lambda}(\cdot|\widehat{\xi}_k)}\left[\|f\|_\infty\right]/\lambda = \|f\|_\infty/\lambda$. To bound the second term, we apply Lemma 5 in Appendix D with function $g \overset{\text{def}}{=} (f(\mathbf{x},\cdot)-\lambda c(\widehat{\xi}_k,\cdot))/\varepsilon$ and the distribution $\mathbb{Q} \overset{\text{def}}{=} \mathcal{N}(\widehat{\xi}_k,\sigma^2\mathbf{I}) \in \text{Proba}(\Xi)$. In this setting, the lemma gives the explicit inequality

$$\mathbb{E}_{\zeta\sim\boldsymbol{\pi}_{\mathbf{x},\lambda}(\cdot|\widehat{\xi}_k)}\left[\frac{f(\mathbf{x},\zeta)-\lambda c(\widehat{\xi}_k,\zeta)}{\varepsilon}\right] \geq \log\left(\mathbb{E}_{\zeta\sim\mathcal{N}(\widehat{\xi}_k,\sigma^2\mathbf{I})}\left[\exp\left(\frac{f(\mathbf{x},\zeta)-\lambda c(\widehat{\xi}_k,\zeta)}{\varepsilon}\right)\right]\right). \qquad (16)$$

Therefore:

$$\begin{aligned}-\frac{\varepsilon}{\lambda}\mathbb{E}_{\zeta\sim\boldsymbol{\pi}_{\mathbf{x},\lambda}(\cdot|\widehat{\xi}_k)}\left[\frac{f(\mathbf{x},\zeta)-\lambda c(\widehat{\xi}_k,\zeta)}{\varepsilon}\right] &\leq -\frac{\varepsilon}{\lambda}\log\left(\mathbb{E}_{\zeta\sim\mathcal{N}(\widehat{\xi}_k,\sigma^2\mathbf{I})}\left[\exp\left(\frac{f(\mathbf{x},\zeta)-\lambda c(\widehat{\xi}_k,\zeta)}{\varepsilon}\right)\right]\right)\\ &\leq -\frac{1}{\lambda}\mathbb{E}_{\zeta\sim\mathcal{N}(\widehat{\xi}_k,\sigma^2\mathbf{I})}\left[f(\mathbf{x},\zeta)-\lambda c(\widehat{\xi}_k,\zeta)\right] \qquad (17)\\ &= -\frac{1}{\lambda}\mathbb{E}_{\zeta\sim\mathcal{N}(\widehat{\xi}_k,\sigma^2\mathbf{I})}\left[f(\mathbf{x},\zeta)\right] + \mathbb{E}_{\zeta\sim\mathcal{N}(\widehat{\xi}_k,\sigma^2\mathbf{I})}\left[c(\widehat{\xi}_k,\zeta)\right]\\ &\leq \frac{1}{\lambda}\|f\|_\infty + m_c\end{aligned}$$

where (17) uses Jensen's inequality on the convex function $-\log$. Finally, we have

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

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

$$0 \leq \varrho + \mathbb{E}_{\widehat{\xi}\sim\widehat{\mathbb{P}}_N}\left[\partial_\lambda\varphi_\varepsilon(\mathbf{x},\lambda_{\max},\widehat{\xi}_k)\right] = \partial_\lambda\left(\lambda\varrho + \mathbb{E}_{\widehat{\xi}\sim\widehat{\mathbb{P}}_N}\left[\varphi_\varepsilon(\mathbf{x},\lambda_{\max},\widehat{\xi}_k)\right]\right) = \partial_\lambda F(\mathbf{x},\lambda_{\max}). \qquad (18)$$

Thanks to Proposition 1, we have that $F(\mathbf{x},\cdot)$ is convex for any fixed $\mathbf{x}\in\mathcal{X}$, thus (18) ensures that the dual minimizer $\lambda^\star$ on $\mathbb{R}_+$ may be found on $[0,\lambda_{\max}]$. □

Proposition 3 proposes an upper bound on the optimal dual variable $\lambda^\star \leq \lambda_{\max}$. Under Assumption 3, this upper bound can be made explicit by replacing the quantity m_c defined in the proposition.

Corollary 1. *Under Assumptions 1 to 3, if* $\varrho > \mathbf{L}_c\sigma^2 d$*, then any optimal dual parameter* $\lambda^\star$ *of* (4) *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 $\mathbb{E}_{\zeta\sim\mathcal{N}(\widehat{\xi}_k,\sigma^2\mathbf{I})}\left[c(\widehat{\xi}_k,\zeta)\right] \leq \mathbf{L}_c\mathbb{E}_{\zeta\sim\mathcal{N}(\widehat{\xi}_k,\sigma^2\mathbf{I})}\left[\|\widehat{\xi}_k-\zeta\|_2^2\right]$ by Assumption 3. 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 $\mathbb{E}_{\zeta\sim\mathcal{N}(\widehat{\xi}_k,\sigma^2\mathbf{I})}\left[c(\widehat{\xi}_k,\zeta)\right] \leq \mathbf{L}_c\sigma^2\left(\mathbb{E}_{\mathcal{N}(0,1)}\left[\mathbf{Z}_1^2\right]+\cdots+\mathbb{E}_{\mathcal{N}(0,1)}\left[\mathbf{Z}_d^2\right]\right)$. Integration by part gives $\mathbb{E}_{\mathcal{N}(0,1)}\left[\mathbf{Z}_i^2\right]=1$, such that $\mathbb{E}_{\zeta\sim\mathcal{N}(\widehat{\xi}_k,\sigma^2\mathbf{I})}\left[c(\widehat{\xi}_k,\zeta)\right] \leq \mathbf{L}_c\sigma^2 d$. Since this is true for all $k \in [\![N]\!]$, defining m_c as in Proposition 3, we have $2\|f\|_\infty/(\varrho-m_c) \leq \lambda_{\max} \overset{\text{def}}{=} 2\|f\|_\infty/(\varrho-\mathbf{L}_c\sigma^2 d)$, and Proposition 3 gives the wanted result. □

A first consequence of Proposition 3 and corollary 1 is the Lipschitz continuity of the gradients, which we formalize in the following lemma.

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

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

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

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

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 Algorithm 1. Under this setting, we can establish the rate of convergence of the stochastic Frank-Wolfe algorithm.

Theorem 1. *Let Assumptions 1 to 3 hold, and suppose $\varrho > \mathbf{L}_c\sigma^2 d$. With batch size b, sampling budget S, step-sizes $\alpha_t = 2/(t+7)$ and momentum terms $\beta_t = 4/(t+8)^{2/3}$, Algorithm 2 gives iterates $(\mathbf{x}_t,\lambda_t)_{t\geq 0} \in \mathcal{X}\times[0,\lambda_{\max}]$ satisfying, for all $t \geq 0$,*

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

where $(\mathbf{x}^\star,\lambda^\star)\in\mathcal{X}\times[0,\lambda_{\max}]$ is a minimizer of F and the constant Q_{corr} is defined as

$$Q_{\text{corr}} \overset{def}{=} \max\left\{\sqrt[3]{9}\left(F(\mathbf{x}_0,\lambda_0)-F(\mathbf{x}^\star,\lambda^\star)\right), \frac{\mathbf{L}_F D}{2} + 2D\max\left\{3\|\nabla F(\mathbf{x}_0,\lambda_0)\|, \sqrt{16\mathsf{V}_{\max}+2\mathbf{L}_F^2 D^2}\right\}\right\}.$$

The constants appearing above are all well-defined: $\mathbf{L}_F$ is the Lipschitz constant for ∇F, D is the diameter of $\mathcal{X}\times[0,\lambda_{\max}]$, C is such that $\|\nabla_{\mathbf{x}} f(\mathbf{x},\cdot)\| \leq C$ on Ξ, μ_c is defined in the ground cost assumption, and $C_c \overset{def}{=} \sup_{\xi,\zeta\in\Xi} c(\xi,\zeta) < +\infty$. Finally

$$\mathsf{V}_{\max} \overset{def}{=} \frac{4(C^2+C_c^2)}{bS}\exp\left(\frac{4\|f\|_\infty}{\varepsilon}\right)\left(1+\frac{4\lambda_{\max}\mu_c\sigma^2}{\varepsilon}\right)^{-d/2}\left(1+\frac{2\lambda_{\max}\mathbf{L}_c\sigma^2}{\varepsilon}\right)^d + \frac{N-b}{b(N-1)}(C^2+C_c^2).$$

The proof relies on the fact that under the given hypothesis, we can give explicit bounds on the variance of the gradient estimator. This is developed in the next important technical lemma.

Lemma 2 (Corrected variance bound for the gradient estimator)**.** *Under the notation and the assumptions of Theorem 1. For a mini-batch $\mathcal{J}$ sampled uniformly without replacement among all subsets of $[\![N]\!]$ with size b, and for a Monte Carlo budget S, the $\mathbf{x}$-component of the gradient estimator satisfies the total variance bound*

$$\mathrm{Var}\left[\mathscr{G}_{\mathbf{x}}^{\mathcal{J}}(\mathbf{x},\lambda)\right] \lesssim \frac{4C^2}{bS}\exp\left(\frac{4\|f\|_\infty}{\varepsilon}\right)\left(1+\frac{4\lambda\mu_c\sigma^2}{\varepsilon}\right)^{-d/2}\left(1+\frac{2\lambda\mathbf{L}_c\sigma^2}{\varepsilon}\right)^d + \frac{N-b}{b(N-1)}C^2.$$

The same bound holds for the λ-component after replacing C by C_c. Consequently, for the full estimator,

$$\mathrm{Var}\left[\mathscr{G}^{\mathcal{J}}(\mathbf{x},\lambda)\right] \lesssim \frac{4(C^2+C_c^2)}{bS}\exp\left(\frac{4\|f\|_\infty}{\varepsilon}\right)\left(1+\frac{4\lambda\mu_c\sigma^2}{\varepsilon}\right)^{-d/2}\left(1+\frac{2\lambda\mathbf{L}_c\sigma^2}{\varepsilon}\right)^d + \frac{N-b}{b(N-1)}(C^2+C_c^2)$$

where $\lesssim$ means that the left-hand sides are bounded up to a multiplicative constant, independent of the parameters.

Proof. We give the proof for the $\mathbf{x}$-component. The proof for the λ-component is identical, replacing the bounded vector $\nabla_{\mathbf{x}} f(\mathbf{x},\cdot)$ by the bounded scalar $c(\widehat{\xi}_k,\cdot)$. For $k \in [\![N]\!]$, define the local self-normalized estimator

$$\mathbf{G}_{k,S}(\mathbf{x},\lambda) \overset{\text{def}}{=} \sum_{s=1}^{S} \nabla_{\mathbf{x}} f(\mathbf{x},\zeta_s^{(k)}) \frac{w_s^{(k)}}{\|w^{(k)}\|_1}, \qquad w_s^{(k)} \overset{\text{def}}{=} \exp\left(\frac{f(\mathbf{x},\zeta_s^{(k)}) - \lambda c(\widehat{\xi}_k,\zeta_s^{(k)})}{\varepsilon}\right),$$

with $\zeta_s^{(k)} \sim \mathcal{N}\big(\widehat{\xi}_k, \sigma^2 \mathbf{I}\big)$ independent across s and k. The global estimator is

$$\mathscr{G}_{\mathbf{x}}^{\mathcal{J}}(\mathbf{x},\lambda) = \frac{1}{b} \sum_{k\in\mathcal{J}} \mathbf{G}_{k,S}(\mathbf{x},\lambda).$$

Conditionally on the mini-batch $\mathcal{J}$, the variables $\mathbf{G}_{k,S}$ are independent across k. Therefore,

$$\mathrm{Var}_{\mathrm{MC}}\left[\mathscr{G}_{\mathbf{x}}^{\mathcal{J}}(\mathbf{x},\lambda)\middle|\mathcal{J}\right] = \frac{1}{b^2} \sum_{k\in\mathcal{J}} \mathrm{Var}_{\mathrm{MC}}\left[\mathbf{G}_{k,S}(\mathbf{x},\lambda)\right] \le \frac{1}{b} \sup_{k\in[\![N]\!]} \mathrm{Var}_{\mathrm{MC}}\left[\mathbf{G}_{k,S}(\mathbf{x},\lambda)\right].$$

It remains to bound the local variance uniformly in k.

To do so, we fix k and write

$$w_k(\zeta) \overset{\text{def}}{=} \exp\left(\frac{f(\mathbf{x},\zeta) - \lambda c(\widehat{\xi}_k,\zeta)}{\varepsilon}\right), \qquad \mu_k \overset{\text{def}}{=} \frac{\mathbb{E}\left[\nabla_{\mathbf{x}} f(\mathbf{x},\zeta) w_k(\zeta)\right]}{\mathbb{E}\left[w_k(\zeta)\right]},$$

where the expectations are taken with respect to $\zeta \sim \mathcal{N}\big(\widehat{\xi}_k, \sigma^2\mathbf{I}\big)$. The classical delta-method linearization of the self-normalized estimator (see *e.g.* Owen [2013]) gives

$$\mathrm{Var}_{\mathrm{MC}}\left[\mathbf{G}_{k,S}(\mathbf{x},\lambda)\right] \lesssim \frac{1}{S} \frac{\mathbb{E}\left[\|\nabla_{\mathbf{x}} f(\mathbf{x},\zeta) - \mu_k\|^2 w_k(\zeta)^2\right]}{\mathbb{E}\left[w_k(\zeta)\right]^2}.$$

Since $\|\nabla_{\mathbf{x}} f(\mathbf{x},\cdot)\| \le C$, we also have $\|\mu_k\| \le C$, and hence $\|\nabla_{\mathbf{x}} f(\mathbf{x},\zeta) - \mu_k\|^2 \le 4C^2$. Thus,

$$\mathrm{Var}_{\mathrm{MC}}\left[\mathbf{G}_{k,S}(\mathbf{x},\lambda)\right] \lesssim \frac{4C^2}{S} \frac{\mathbb{E}\left[w_k(\zeta)^2\right]}{\mathbb{E}\left[w_k(\zeta)\right]^2}.$$

The numerator is bounded using the lower comparison $c(\widehat{\xi}_k,\zeta) \ge \mu_c \|\widehat{\xi}_k - \zeta\|_2^2$ and $|f| \le \|f\|_\infty$:

$$\mathbb{E}\left[w_k(\zeta)^2\right] \le \exp\left(\frac{2\|f\|_\infty}{\varepsilon}\right) \mathbb{E}\left[\exp\left(-\frac{2\lambda\mu_c}{\varepsilon}\|\widehat{\xi}_k - \zeta\|_2^2\right)\right].$$

With $\zeta = \widehat{\xi}_k + \sigma Z$, $Z \sim \mathcal{N}(0,\mathbf{I})$, this gives the uniform bound

$$\mathbb{E}\left[w_k(\zeta)^2\right] \le \exp\left(\frac{2\|f\|_\infty}{\varepsilon}\right)\left(1 + \frac{4\lambda\mu_c\sigma^2}{\varepsilon}\right)^{-d/2}.$$

For the denominator, Assumption 3 also gives the upper comparison

$$c(\widehat{\xi}_k,\zeta) \le \mathbf{L}_c \|\widehat{\xi}_k - \zeta\|_2^2.$$

Since the cost appears with a negative sign in the exponential, we get

$$w_k(\zeta) \ge \exp\left(-\frac{\|f\|_\infty}{\varepsilon}\right) \exp\left(-\frac{\lambda\mathbf{L}_c}{\varepsilon}\|\widehat{\xi}_k - \zeta\|_2^2\right) \quad \text{and} \quad \mathbb{E}\left[w_k(\zeta)\right]^2 \ge \exp\left(-\frac{2\|f\|_\infty}{\varepsilon}\right)\left(1 + \frac{2\lambda\mathbf{L}_c\sigma^2}{\varepsilon}\right)^{-d}.$$

Combining the bounds above gives, uniformly in k, a bound on $\mathrm{Var}_{\mathrm{MC}}\left[\mathbf{G}_{k,S}(\mathbf{x},\lambda)\right]$ and directly

$$\mathbb{E}_{\mathcal{J}}\left[\mathrm{Var}_{\mathrm{MC}}\left[\mathscr{G}_{\mathbf{x}}^{\mathcal{J}}(\mathbf{x},\lambda)\middle|\mathcal{J}\right]\right] \lesssim \frac{4C^2}{bS} \exp\left(\frac{4\|f\|_\infty}{\varepsilon}\right)\left(1 + \frac{4\lambda\mu_c\sigma^2}{\varepsilon}\right)^{-d/2}\left(1 + \frac{2\lambda\mathbf{L}_c\sigma^2}{\varepsilon}\right)^{d}.$$

It remains to account for the randomness of the mini-batch. Let

$$m_{k,S}(\mathbf{x},\lambda) \overset{\text{def}}{=} \mathbb{E}_{\mathrm{MC}}\left[\mathbf{G}_{k,S}(\mathbf{x},\lambda)\right], \qquad \bar{m}_S(\mathbf{x},\lambda) \overset{\text{def}}{=} \frac{1}{N}\sum_{k=1}^{N} m_{k,S}(\mathbf{x},\lambda).$$

The finite-population variance formula for uniform sampling without replacement gives

$$\mathrm{Var}_{\mathcal{J}}\left[\frac{1}{b}\sum_{k\in\mathcal{J}} m_{k,S}(\mathbf{x},\lambda)\right] = \frac{N-b}{b(N-1)}\cdot\frac{1}{N}\sum_{k=1}^{N}\left\|m_{k,S}(\mathbf{x},\lambda)-\bar{m}_S(\mathbf{x},\lambda)\right\|^2.$$

Each $\mathbf{G}_{k,S}$ is a convex combination of gradients with norm at most C: Jensen's inequality gives $\|m_{k,S}(\mathbf{x},\lambda)\|\le C$. Therefore,

$$\frac{1}{N}\sum_{k=1}^{N}\left\|m_{k,S}(\mathbf{x},\lambda)-\bar{m}_S(\mathbf{x},\lambda)\right\|^2 \le \frac{1}{N}\sum_{k=1}^{N}\left\|m_{k,S}(\mathbf{x},\lambda)\right\|^2 \le C^2.$$

The law of total variance now gives

$$\mathrm{Var}\left[\mathscr{G}_{\mathbf{x}}^{\mathcal{J}}(\mathbf{x},\lambda)\right] = \mathbb{E}_{\mathcal{J}}\left[\mathrm{Var}_{\mathrm{MC}}\left[\mathscr{G}_{\mathbf{x}}^{\mathcal{J}}(\mathbf{x},\lambda)\middle|\mathcal{J}\right]\right] + \mathrm{Var}_{\mathcal{J}}\left[\mathbb{E}_{\mathrm{MC}}\left[\mathscr{G}_{\mathbf{x}}^{\mathcal{J}}(\mathbf{x},\lambda)\middle|\mathcal{J}\right]\right],$$

which proves the announced bound for the $\mathbf{x}$-component.

For the λ-component, the estimator is a constant ϱ minus a mini-batch average of self-normalized averages of $c(\widehat{\xi}_k,\zeta)$. Since $0\le c(\xi,\zeta)\le C_c$, the same proof applies with C replaced by C_c. Summing the componentwise bounds gives the stated bound for the full gradient estimator. □

From the previous lemma, the proof of the convergence theorem then follows from existing convergence results [Braun et al., 2025].

Proof of Theorem 1. Use Theorem 4.12 in [Braun et al., 2025]. The assumptions needed there are satisfied because F is convex and $\mathbf{L}_F$-smooth, $\mathcal{X}\times[0,\lambda_{\max}]$ has diameter D, and the stochastic gradients have uniformly bounded variance by Lemma 2. As Corollary 1 states, we can bound each λ_t by $\lambda_{\max}\le 2\|f\|_\infty/(\varrho-\mathbf{L}_c\sigma^2 d)$. Setting this as an upper bound for the iterates gives the result. □

4 Numerical illustrations

We illustrate our framework on our two running examples, the Traffic Assignment and Quadratic Minimum Spanning Tree. We first describe the common experimental protocol and the way the WDRO and ERM baselines are compared. We then study the two applications separately, highlighting how the distributionally robust approach adapts to different structural environments: the continuous Traffic Assignment in Section 4.2 and the relaxed Quadratic Minimum Spanning Tree in Section 4.3.

4.1 Experimental setting

We compare the robustness of solutions obtained with our WDRO framework against solutions obtained by standard empirical risk minimization. Given a training data set $\widehat{\xi}_1,\dots,\widehat{\xi}_{N_{\mathrm{train}}}$, both approaches use the same feasible region and the same linear minimization oracle; they differ only in the objective and in the gradient information used by the first-order method.

	ERM	WDRO
Objective	$\frac{1}{N_{\mathrm{train}}}\sum_{k=1}^{N_{\mathrm{train}}} f(\mathbf{x},\widehat{\xi}_k)$	$F(\mathbf{x},\lambda)$ as in (4)
Gradient	$\frac{1}{N_{\mathrm{train}}}\sum_{k=1}^{N_{\mathrm{train}}} \nabla_{\mathbf{x}} f(\mathbf{x},\widehat{\xi}_k)$	$\mathscr{G}^{\mathcal{J}}(\mathbf{x},\lambda)$ as in (13)
LMO	Same linear oracle	
Algorithm	*Classical Frank-Wolfe*	Algorithm 2

Once both solutions are computed, we evaluate their out-of-sample behavior on shifted test scenarios $\widetilde{\xi}_1,\dots,\widetilde{\xi}_{N_{\mathrm{test}}}$. More precisely, we compare the parametrized losses $f(\mathbf{x}_{\mathrm{WDRO}},\cdot)$ and $f(\mathbf{x}_{\mathrm{ERM}},\cdot)$ on data drawn from a distribution that differs from the training distribution. The instance generators are chosen to make this shift controlled: for the

Quadratic Minimum Spanning Tree, we build synthetic graphs and cost matrices; for the Traffic Assignment, we perturb a real base network. This lets us stress-test ERM while keeping the experimental setting interpretable.

Throughout the experiments, we use the squared Euclidean ground cost $c(\xi,\zeta) \overset{\text{def}}{=} \|\xi-\zeta\|_2^2$ in the Wasserstein distance. The robust model also depends on the Wasserstein radius $\varrho > 0$, the sampling variance σ^2, and the smoothing temperature $\varepsilon > 0$. These parameters shape the robust objective itself, so we tune them empirically through validation experiments.

The remaining numerical issue is the upper bound $\lambda_{\max}$ for the dual variable. The bound must be large enough not to exclude the relevant minimizer, but excessively large values can create numerical instabilities in the exponential weights when $\lambda\|\widehat{\xi}_k-\zeta\|_2^2$ dominates $f(\mathbf{x},\zeta)$. We therefore calibrate $\lambda_{\max}$ so that $\lambda_{\max}\overline{c}$ has the same order of magnitude as the typical dispersion Δ_f of the loss over sampled scenarios, where

$$\Delta_f \overset{\text{def}}{=} \mathbb{E}_{\mathbf{x}}\left[\operatorname{diam}(f(\mathbf{x},\Xi))\right] = \mathbb{E}_{\mathbf{x}}\left[\max_{\xi\in\Xi} f(\mathbf{x},\xi) - \min_{\zeta\in\Xi} f(\mathbf{x},\zeta)\right] \qquad \text{and} \qquad \overline{c} \overset{\text{def}}{=} \mathbb{E}_{\widehat{\xi}\sim\widehat{\mathbb{P}}_N}\left[\mathbb{E}_{\zeta}\left[\|\widehat{\xi}-\zeta\|_2^2\right]\right].$$

Both quantities are only needed as calibration scales, so we estimate them by the following sampling heuristic.

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

Input: training samples $\{\widehat{\xi}_1,\ldots,\widehat{\xi}_N\}$ of the learning problem
for $k \in [\![N]\!]$ **do**
Sample $\zeta_1^{(k)},\ldots,\zeta_S^{(k)} \sim \mathcal{N}(\widehat{\xi}_k,\sigma^2\mathbf{I})$
$\mathbf{x}_k \leftarrow \mathrm{LMO}(\zeta_k)$ 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
$\widetilde{\Delta_f} \leftarrow \dfrac{1}{N}\displaystyle\sum_{k=1}^{N}\left(\max_{s\in[\![S]\!]}\left\{f(\mathbf{x}_k,\zeta_s^{(k)})\right\} - \min_{s'\in[\![S]\!]}\left\{f(\mathbf{x}_k,\zeta_s^{(k)})\right\}\right)$ ▷ Approximation of the average dispersion
return $\lambda_{\max} \overset{\text{def}}{=} \widetilde{\Delta_f}/2\widetilde{c}$

For all experiments, the primal variable $\mathbf{x}_0$ is initialized at random in $\mathcal{X}$, and the dual variable is initialized at the midpoint of the calibrated interval, $\lambda_0 \overset{\text{def}}{=} \lambda_{\max}/2$. Both WDRO and ERM solvers are run with a maximum budget of 5000 iterations.

4.2 Uncertain Traffic Assignment

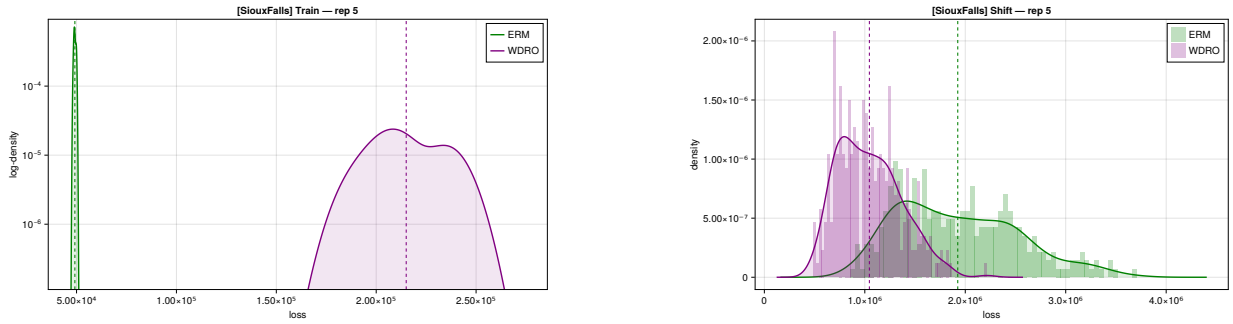
We construct our instances based on the "SiouxFalls" network, with 24 nodes and 76 links, from the *TransportationNetwork* dataset [Xu et al., 2024]. For this problem, our linear minimization oracle is solving a shortest path problem on the network, with weights given by the current gradient [Mitradjieva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

$$\mathbf{x}_{\mathrm{WDRO}} \overset{\text{def}}{=} \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}_{\mathrm{flow}}} \min_{\lambda\in[0,\lambda_{\max}]} F(\mathbf{x},\lambda) \qquad \text{and} \qquad \mathbf{x}_{\mathrm{ERM}} \overset{\text{def}}{=} \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}_{\mathrm{flow}}} \frac{1}{N_{\mathrm{train}}}\sum_{k=1}^{N_{\mathrm{train}}}\sum_{a\in A}\int_0^{\mathbf{x}_a}(t_k)_a^{(0)}\left(1+\alpha_k\left(\frac{s}{c_a}\right)^{\beta_k}\right)\mathrm{d}s.$$

We build a shifted distribution $\widetilde{\mathbf{P}}_{\mathbf{Inst}}$, whose construction is detailed in Appendix A.1, from which we sample a shifted set $\Xi_{\mathrm{test}} \overset{\text{def}}{=} \{\widetilde{\xi}_1,\ldots,\widetilde{\xi}_{N_{\mathrm{test}}}\}$ with $\widetilde{\xi}_1,\ldots,\widetilde{\xi}_{N_{\mathrm{test}}} \sim \widetilde{\mathbf{P}}_{\mathbf{Inst}}$. We evaluate the losses on the training dataset and the test dataset

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

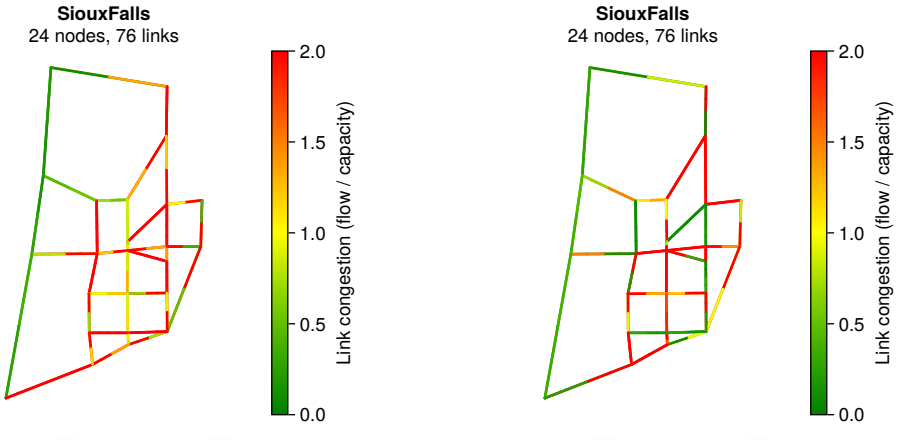
For this experiment, the chosen WDRO parameters are: $\varepsilon = 10^{-3}$ for the smoothing temperature, the Wasserstein radius has been set to $\varrho = 17$, and we estimated each integral with $S = 100$ samples and a mini-batch size of $b = 10$.



(a) training data (b) shift test

Figure 1: example of Traffic Assignment on "SiouxFalls". Purple histograms represent the losses obtained with the robust solution; green histograms represent the losses obtained with the ERM. For better readability, in Figure 1a we represented the log-density of the Kernel density of the losses on the training data.

Figures 1a and 1b 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 2.18×10^5. 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.



(a) shift ERM flow (b) shift robust flow

Figure 2: Illustrate the behavior of the ERM solution (left) and the robust solution (right) on "SiouxFalls". On each arc, a red shade indicates a high congestion-to-capacity ratio, while a green shade indicates that the considered link is uncongested.

In Figures 2a and 2b, we visualize the routing profiles yielded by the ERM and the robust solution on a common instance with a shifted distribution. We see that the robust solution alleviates congestion on many arcs in the city center, providing a solution with a reduced objective value.

4.3 Uncertain Quadratic Minimum Spanning Tree

For the Quadratic Minimum Spanning Tree, we evaluate the robustness and out-of-sample performance of the fractional solutions produced by the convex relaxation. The instances are generated with varying degrees of freedom, through changes in the number of nonzero entries, the variance of the entries, and the support pattern; the full generation procedure is described in Appendix A.2 and Algorithm 5. The linear minimization oracle is solved by Kruskal's algorithm [Schrijver, 2002], with complexity $\mathcal{O}(m\log(m))$.

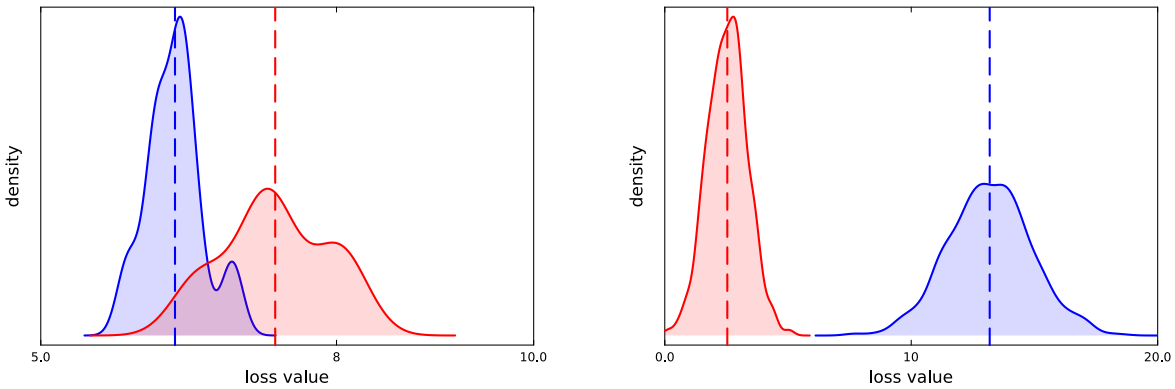
To test robustness, we build a shifted distribution $\widetilde{\mathbf{P}}_G$ using the same construction as the training distribution, but with a modified ground cost $\widetilde{\mu}_G$, a modified mask $\widetilde{M}$, and a larger variance on the entries. This shift is designed to represent a more unstable regime, where the test instances have more degrees of freedom than the training instances. We then compute the relaxed WDRO and ERM solutions

$$\mathbf{x}_{\mathrm{WDRO}} \in \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}_{\mathrm{tree}}} \min_{\lambda\in[0,\lambda_{\max}]} F(\mathbf{x},\lambda), \qquad \mathbf{x}_{\mathrm{ERM}} \in \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}_{\mathrm{tree}}} \frac{1}{N_{\mathrm{train}}} \sum_{k=1}^{N_{\mathrm{train}}} \mathbf{x}^\top \widehat{\xi}_k \mathbf{x}.$$

In practice, we set $\varepsilon = 10^{-4}$, $\varrho = 10$, and use a sampling budget $S = 10$ with mini-batches of size $b = 10$. The shifted test set is sampled as $\Xi_{\mathrm{test}} \overset{\mathrm{def}}{=} \{\widetilde{\xi}_1, \ldots, \widetilde{\xi}_{N_{\mathrm{test}}}\}$ with $\widetilde{\xi}_1, \ldots, \widetilde{\xi}_{N_{\mathrm{test}}} \sim \widetilde{\mathbf{P}}_G$. We compare ERM and WDRO on both the training and test sets through the loss collections

$$\mathcal{L}_\Xi(\mathbf{x}) \overset{\mathrm{def}}{=} \left\{ f(\mathbf{x},\xi) = \mathbf{x}^\top \xi \mathbf{x} \,\middle|\, \xi \in \Xi \right\}, \qquad \mathbf{x} \in \{\mathbf{x}_{\mathrm{ERM}}, \mathbf{x}_{\mathrm{WDRO}}\}, \quad \Xi \in \{\Xi_{\mathrm{train}}, \Xi_{\mathrm{test}}\}.$$

The histograms in Figures 3a and 3b show the resulting empirical loss distributions, with kernel-density estimates added for readability.



(a) training data (b) shift test

Figure 3: Quadratic Minimum Spanning Tree example on a graph G on $n = 50$ vertices and $m = 331$ edges. 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.

On the training data, Figure 3a shows the expected advantage of ERM: it is optimized directly on the empirical distribution and therefore obtains the lowest average training losses, while the WDRO solution is deliberately more conservative. Under the shifted distribution, however, the ranking reverses[3]. As shown in Figure 3b, the WDRO solution achieves smaller losses than ERM, illustrating the intended tradeoff: WDRO sacrifices empirical performance to obtain a solution that is less sensitive to distributional changes.

For a second experiment, we focus on a density test with fixed graph size $n = 50$. For densities spread from sparse to complete graphs, we consider edge counts m_k such that $d(m_k, n) \approx k/10$, $k \in [\![10]\!]$. For each density, we generate $n_{\mathrm{inst}} = 10$ independent instances on the same graph size but with different training dynamics, namely different base costs, masks, and generation parameters. We aggregate the shifted-test advantage of WDRO over ERM through

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

where $\mathbf{x}^{(m,i)}_{\mathrm{ERM}}$ and $\mathbf{x}^{(m,i)}_{\mathrm{WDRO}}$ denote the ERM and WDRO solutions for the i^{th} instance at edge count m.

[3] This behavior should be interpreted in light of Appendix C. In highly constrained combinatorial problems, the feasible set may contain only a limited number of solutions, so ERM can already generalize well even under moderate distributional shifts. In particular, Theorems 2 and 3 establish the good performance of ERM in finite settings. The robust objective is more useful in looser settings, where computing an integer solution is in practice intractable, or where the feasible set is richer, and ERM has more opportunity to overfit the empirical sample. Consistently with Theorem 2, we study how the relative performance of WDRO changes as the problem becomes less constrained, using the graph size n and the density $d(m,n) \overset{\mathrm{def}}{=} 2m/(n(n-1))$ as complexity parameters.

Figure 4 reports these aggregated shifted losses. The distributions remain instance-dependent, and their averages do not follow a perfectly monotone trend, reflecting the sensitivity of robustness gains to the particular shifted distribution $\widetilde{\mathbf{P}}_G$. Nevertheless, the averages of $\mathcal{D}^{(m)}_{\text{test}}$ are positive across densities, meaning that WDRO improves shifted-test performance on average. The dispersion also narrows as density increases: in denser, less constrained graphs, the advantage of the robust solution becomes more systematic. Overall, the experiment suggests that the benefit of WDRO appears when the shifted distribution is sufficiently different from the empirical distribution, especially when the training set has limited variability or when the scenario variance is large enough to expose different behaviors in the gradient estimates. This is precisely the regime where distributional shifts can change the effective decision problem, and where the WDRO formulation is expected to be most relevant.

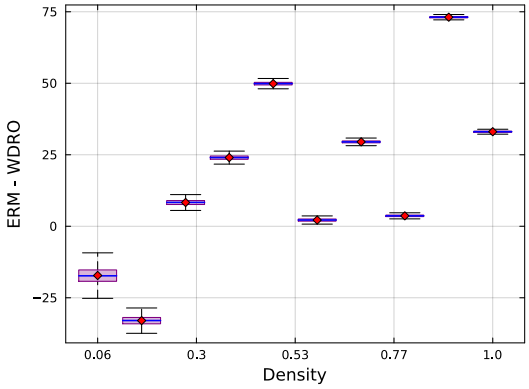


Figure 4: aggregated losses for instances of the Quadratic Minimum Spanning Tree on a graph G on $n = 50$ vertices and $m \in \{75, 203, \ldots, 1225\}$ edges.

5 Conclusion

This work proposes a general algorithmic framework bringing Wasserstein distributional robustness to constrained stochastic optimization. The main ingredient is to replace the nonsmooth inner worst-case term by an entropic approximation, which yields a smooth objective with computable stochastic gradients. Coupled with a momentum stochastic Frank-Wolfe scheme, this makes the approach compatible with feasible sets described through a linear minimization oracle, and therefore with many continuous constraint sets, as well as convex relaxations of mixed-integer and combinatorial problems. We prove the consistency and unbiasedness of the gradient estimators and derive convergence guarantees for the resulting method. The numerical experiments on traffic assignment and quadratic spanning trees show the expected tradeoff: ERM remains competitive when the constraints restrict the solutions, whereas WDRO becomes valuable when the test distribution departs from the training data and the feasible set is rich enough for overfitting. This supports the use of smooth WDRO as a practical robustness tool for structured decision problems, and opens many questions, notably on geometric properties of the feasible set which determine the empirical gains of robustness.

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. We thank Changhyun Kwon and Guillaume Dalle for the development of `TrafficAssignment.jl`. We also thank Mathis Azéma for relevant discussions on the unconstrained setting, and extend our sincere thanks to Waïss Azizian for his feedback and comments on the paper.

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.

Dimitris Bertsimas, David B. Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011. doi: 10.1137/080734510. URL `https://doi.org/10.1137/080734510`.

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.

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.

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

Vincent Florian, Waïss Azizian, Franck Iutzeler, and Jérôme Malick. `skwdro`: a library for Wasserstein distributionally robust machine learning. *Journal of Machine Learning Research*, 27(8):1–7, 2026.

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.

Zaid Harchaoui, Anatoli Juditsky, and Arkadi Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, 152(1):75–112, 2015.

Deborah Hendrych, Hannah Troppens, Mathieu Besançon, and Sebastian Pokutta. Convex mixed-integer optimization with Frank-Wolfe methods. *Mathematical Programming Computation*, 17(4):731–757, 2025.

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.

Tam Le and Jérôme Malick. Universal generalization guarantees for wasserstein distributionally robust models. *ICLR*, 2025:20887–20921, 2025.

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.

Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. Stochastic conditional gradient methods: From convex minimization to submodular maximization. *Journal of machine learning research*, 21(105):1–49, 2020.

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

Art B. Owen. *Monte Carlo Theory, Methods and Examples*. `https://artowen.su.domains/mc/`, 2013. Chapter 9: Importance Sampling.

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.

Sebastian Pokutta. Symmetric extension complexity of the spanning tree polytope. *arXiv preprint arXiv:2606.17017*, 2026.

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.

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.

Jianzhe Zhen, Daniel Kuhn, and Wolfram Wiesemann. A unified theory of robust and distributionally robust optimization via the primal-worst-equals-dual-best principle. *Operations Research*, 73(2):862–878, 2025.

A Generation of instances

This appendix details the instance-generation procedures used in the numerical section. Its purpose is twofold: first, to make explicit which part of each problem instance is kept fixed and which part is random; second, to clarify how the training and shifted distributions differ in the experiments. The two procedures below follow the same pattern: a nominal structure is fixed, and random scenarios are generated around it. In Section 4, the robust and ERM solutions are then compared on samples drawn either from the training distribution or from a shifted distribution.

A.1 Instances for the Traffic Assignment

The instances used for the Traffic Assignment experiments come from the *TransportationNetwork* dataset (https://github.com/bstabler/TransportationNetworks). Each nominal instance provides a network $G = (V, A)$ with capacities $c_a > 0$, free-flow times $t_a^{(0)} > 0$ for all $a \in A$, along with a congestion multiplier α and a power parameter $\beta > 0$ in the BPR-type travel-time function.

The network topology and capacities are kept fixed throughout the experiment. Uncertainty is introduced only through the travel-time parameters: free-flow times are multiplied arc-wise, while α and β are sampled around shifted values. Positive values of μ_α and μ_β make the training scenarios optimistic relative to the nominal parameters, which creates a controlled mismatch with the more pessimistic shifted scenarios used for out-of-sample evaluation. The uncertain training distribution is generated as follows:

Algorithm 4 *Generation of Uncertain Traffic Assignment 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 $\mathbf{P}_{\mathbf{Inst}}$ defined as
    function $\mathbf{P}_{\mathbf{Inst}}$:
      for $a \in A$ **do**
        Sample $m_a \sim \mathcal{U}([0.75, 1])$ ▷ change the free flow time of arc a
      end for
      Sample $\widetilde{\alpha} > 0$ with $\widetilde{\alpha} \sim \mathcal{N}\left(\alpha - \mu_\alpha, \sigma_\alpha^2\right)$ ▷ normal perturbation
      Sample $\widetilde{\beta} > 0$ with $\widetilde{\beta} \sim \mathcal{U}\left(\left[\left(\beta - \mu_\beta\right) \pm B_\beta/2\right]\right)$ ▷ 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 $\mathbf{P}_{\mathbf{Inst}}$

This procedure yields uncertain Traffic Assignment scenarios in which the objective parameters vary while the underlying road network remains unchanged. We optimize the robust objective and the ERM objective using training scenarios $\widehat{\xi}_1, \ldots, \widehat{\xi}_{N_{\text{train}}} \sim \mathbf{P}_{\mathbf{Inst}}$. 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 1a and 1b and the illustration of Figures 2a and 2b are obtained with generating parameters $\sigma_\alpha^2 = 0.3 \times \alpha$, $B_\beta = 0.9$, $\mu_\alpha = 0.15 \times \alpha$, $\mu_\beta = 1$.

A.2 Instances for the Quadratic Minimum Spanning Tree

For the *Uncertain Quadratic Minimum Spanning Tree*, the graph determines the combinatorial structure, while the random cost matrix determines the quadratic interaction between selected edges. We first generate a connected graph and then sample a base cost matrix on its fixed edge set. A binary mask is used to introduce sparsity and missing interactions, so that each scenario contains both noisy costs and incomplete information. The normalization step keeps the scale of the generated matrices comparable across instances. The full generation protocol is given in Algorithm 5. Each instance is described by a graph G with n vertices and m edges, together with a "true" distribution $\mathbf{P}_G$ from which the training data are sampled: $\widehat{\xi}_1, \ldots, \widehat{\xi}_{N_{\text{train}}} \sim \mathbf{P}_G$.

Algorithm 5 *Generation of Uncertain Quadratic Minimum Spanning Tree 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 $M_{ij} \sim \mathscr{B}(0,7), \forall i, j \in [\![m]\!]$ ▷ 30% chance of missing data
  Build distribution $\mathbf{P}_G$ defined as
    function $\mathbf{P}_G$:
      Sample $\mathbf{C}$ with $\mathbf{C}_{ij} \sim \mathcal{N}(0,1)$ ▷ perturbation
      $N \leftarrow M \otimes \left(\mu_G + 0,1 \times \mathbf{C}\right)$ ▷ Base cost + noise + missing data
      return $N/\|N\|_2$ ▷ normalized to facilitate comparison across instances
    end
  return $(G, \mathbf{P}_G)$

B Comparison between exact and smoothed WDRO

This appendix compares the smoothed WDRO formulation (4) and the initial dual WDRO formulation (3) in a case where the inner supremum over ζ can be evaluated explicitly. The aim of this appendix is to show that both the exact and smoothed WDRO approaches may provide improved robustness under distributional shift compared with ERM, but that neither approach uniformly dominates the other.

We focus on the Quadratic Minimum Spanning Tree objective with unconstrained cost matrices, namely $\Xi = \mathbb{R}^{m\times m}$, and use the squared Euclidean ground cost $c(\xi,\zeta) = \|\xi - \zeta\|_2^2$. This setting is simpler than the one considered in Section 4.3, but it provides a useful reference point: the inner supremum in the dual WDRO problem admits a closed-form solution.

For a fixed spanning-tree decision $\mathbf{x}$ and a dual parameter $\lambda > 0$, the sample-wise adversarial term is

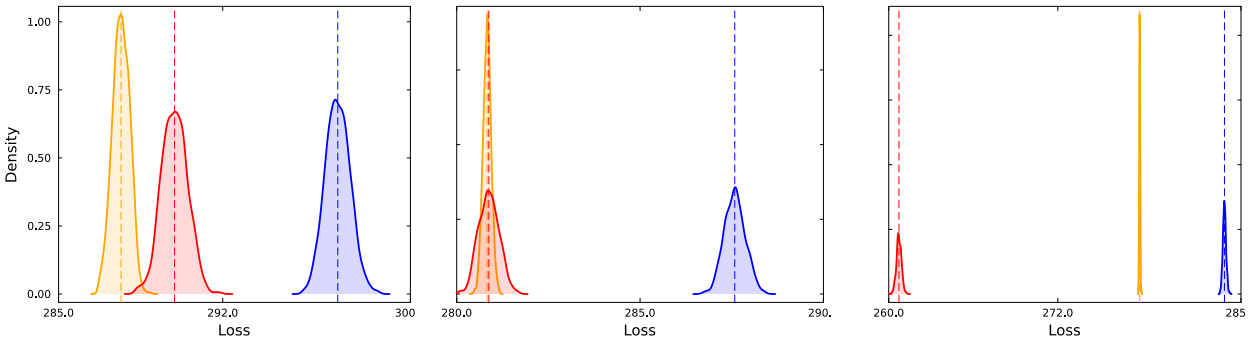
$$\sup_{\zeta\in\mathbb{R}^{m\times m}} \left\{\mathbf{x}^\top \zeta \mathbf{x} - \lambda \|\widehat{\xi} - \zeta\|_2^2\right\}.$$

The function inside the supremum is strictly concave in ζ. Its gradient vanishes at $\zeta^\star = \widehat{\xi} + \frac{1}{2\lambda}\mathbf{x}\mathbf{x}^\top$, which is therefore the unique maximizer. Substituting this optimal value into the dual formulation gives $\mathrm{WDRO}_{\mathrm{ST}} \leq \min_{\mathbf{x}\in\mathcal{X}_{\mathrm{tree}}} \inf_{\lambda>0} \lambda\varrho + \mathbb{E}_{\widehat{\xi}\sim\widehat{\mathbb{P}}_N}\left[\mathbf{x}^\top\widehat{\xi}\mathbf{x} + \frac{1}{4\lambda}\|\mathbf{x}\mathbf{x}^\top\|^2\right] = \min_{\mathbf{x}\in\mathcal{X}_{\mathrm{tree}}} \mathbb{E}_{\widehat{\xi}\sim\widehat{\mathbb{P}}_N}\left[\mathbf{x}^\top\widehat{\xi}\mathbf{x} + \inf_{\lambda>0}\left\{\lambda\varrho + \frac{\|\mathbf{x}\|^4}{4\lambda}\right\}\right]$. The remaining one-dimensional minimization is explicit: the minimizer is $\lambda^\star = \|\mathbf{x}\|^2/(2\sqrt{\varrho})$ and yields

$$\mathrm{WDRO}_{\mathrm{ST}} \leq \min_{\mathbf{x}\in\mathcal{X}_{\mathrm{tree}}} \mathbb{E}_{\widehat{\xi}\sim\widehat{\mathbb{P}}_N}\left[\mathbf{x}^\top\widehat{\xi}\mathbf{x} + \sqrt{\varrho}\|\mathbf{x}\|^2\right] = \min_{\mathbf{x}\in\mathcal{X}_{\mathrm{tree}}} \mathbb{E}_{\widehat{\xi}\sim\widehat{\mathbb{P}}_N}\left[\mathbf{x}^\top\left(\widehat{\xi} + \sqrt{\varrho}\,\mathbf{I}\right)\mathbf{x}\right].$$

Note that this closed-form expression relies on the unconstrained scenario space $\Xi = \mathbb{R}^{m\times m}$ and does not directly apply to the constrained scenario sets considered in the main experiments (see Appendix A.2). In contrast, our smoothed approach can accommodate such constraints through sampling feasible scenarios, *e.g.*, using rejection sampling (see Remark 1).

Thus, in this unconstrained setting, the exact robust counterpart can be solved as an empirical problem with shifted training matrices $\widehat{\xi}_k + \sqrt{\varrho}\,\mathbf{I}$. We denote the corresponding solution by $\mathbf{x}_{\mathrm{exact}}$ and compare it with the smoothed WDRO solution and the ERM solution.



(a) (b) (c)

Figure 5: Losses for the Quadratic Minimum Spanning Tree with unconstrained cost matrices on unseen data, for 3 different instances (a),(b),(c). Yellow histograms correspond to the exact robust solution, red histograms to the smoothed robust solution, and blue histograms to ERM. We see that both exact/smoothed robust approaches improve over ERM.

Figure 5 shows that both robust variants improve over ERM and that the comparison between the two is instance-dependent. In some cases, the exact reformulation gives the best shifted performance (a); in others, the smoothed WDRO solution is comparable (b) or better (c). This is consistent with the role of smoothing in the main framework: it is not intended to reproduce the exact worst-case solution in every special case, but to provide a tractable and broadly applicable surrogate when the exact inner supremum is unavailable or too costly to exploit.

C Properties of the empirical risk minimization in the discrete case

In this section, we quantify the quality of a solution to (ERM) for the underlying expectation minimization problem when the number of solutions is finite (i.e., $\mathcal{X}$ is a finite subset of $\mathbb{Z}^n$. This highlights a complementary aspect of the behavior of ERM: it tends to perform well in generalization and robustness to moderate shifts in highly constrained settings, similarly to what we observe empirically in the main text for continuous problems.

Theorem 2 formalizes the generalization guarantee for finite solution sets. The analysis is similar to the ERM guarantees provided in Kleywegt et al. [2002], with a few differences. We use the boundedness of f as a simplifying assumption, which, although stronger than their setting using only finiteness of the moment-generating function of $f(x,\cdot)$, is realistic for many operations research problems. This allows us to derive a non-asymptotic result that remains valid for any N. 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 $\mathcal{X}$ be a finite set of points. Let $\mathbf{P}$ be an unknown probability distribution over Ξ and let $\widehat{\xi}_1,\ldots,\widehat{\xi}_N$ be drawn i.i.d. from $\mathbf{P}$. Define the true risk and empirical risk as*

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

Define the ERM solution $\mathbf{x}_{\mathrm{ERM}} \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} \widehat{\mathbf{F}}_N(\mathbf{x})$ and the true optimal value $\mathbf{F}^\star \overset{def}{=} \min_{\mathbf{x}\in\mathcal{X}} \mathbf{F}(\mathbf{x})$. Then, for a threshold $\delta\in(0,1)$, with probability at least $1-\alpha$:

$$\mathbf{F}(\mathbf{x}_{\mathrm{ERM}}) - \mathbf{F}^\star \le 2\|f\|_\infty \sqrt{\frac{2\log\left(2|\mathcal{X}|/\alpha\right)}{N}} + \delta. \tag{19}$$

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

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

By applying the union bound to the event of a deviation t between the empirical and true risks over all $\mathbf{x} \in \mathcal{X}$,

$$\mathbb{P}\left(\sup_{\mathbf{x}\in\mathcal{X}} |\widehat{\mathbf{F}}_N(\mathbf{x}) - \mathbf{F}(\mathbf{x})| \geq t\right) \leq \sum_{\mathbf{x}\in\mathcal{X}} \mathbb{P}\left(|\widehat{\mathbf{F}}_N(\mathbf{x}) - \mathbf{F}(\mathbf{x})| \geq t\right) \tag{20}$$
$$\leq \sum_{\mathbf{x}\in\mathcal{X}} 2\exp\left(-\frac{Nt^2}{2\|f\|_\infty^2}\right) = 2|\mathcal{X}|\exp\left(-\frac{Nt^2}{2\|f\|_\infty^2}\right).$$

We can derive the requirement to upper-bound the right-hand side by α:

$$2|\mathcal{X}|\exp\left(-\frac{Nt^2}{2\|f\|_\infty^2}\right) \leq \alpha \quad \Rightarrow \quad t \geq \|f\|_\infty \sqrt{\frac{2\log\left(2|\mathcal{X}|/\alpha\right)}{N}}.$$

Define the uniform convergence event:

$$\mathscr{E} = \left\{ \sup_{\mathbf{x}\in\mathcal{X}} |\widehat{\mathbf{F}}_N(\mathbf{x}) - \mathbf{F}(\mathbf{x})| \leq \|f\|_\infty \sqrt{\frac{2\log\left(2|\mathcal{X}|/\alpha\right)}{N}} \right\} \tag{21}$$

From complementing the event in the probability in (20), we obtain $\mathbb{P}(\mathscr{E}) \geq 1-\alpha$.

Denote with $\mathbf{x}_{\text{ERM}}$ a δ-approximate optimizer of $\widehat{\mathbf{F}}_N$ over $\mathcal{X}$: $\widehat{\mathbf{F}}_N(\mathbf{x}_{\text{ERM}}) \leq \delta + \min_{\mathbf{x}\in\mathcal{X}} \widehat{\mathbf{F}}_N(\mathbf{x})$, and with $\mathbf{x}^\star$ an optimizer of $\mathbf{F}$ over $\mathcal{X}$. We have, with probability $1-\alpha$:

$$\mathbf{F}(\mathbf{x}_{\text{ERM}}) \leq \widehat{\mathbf{F}}_N(\mathbf{x}_{\text{ERM}}) + \|f\|_\infty \sqrt{\frac{2\log\left(2|\mathcal{X}|/\alpha\right)}{N}} \tag{22}$$
$$\leq \widehat{\mathbf{F}}_N(\mathbf{x}^\star) + \delta + \|f\|_\infty \sqrt{\frac{2\log\left(2|\mathcal{X}|/\alpha\right)}{N}} \tag{23}$$
$$\leq \mathbf{F}(\mathbf{x}^\star) + \delta + 2\|f\|_\infty \sqrt{\frac{2\log\left(2|\mathcal{X}|/\alpha\right)}{N}}. \tag{24}$$

Inequalities (22) and (24) are obtained from the uniform convergence event valid at $\mathbf{x}_{\text{ERM}}$ and $\mathbf{x}^\star$ respectively, and (23) comes from $\mathbf{x}_{\text{ERM}}$ being a minimizer of $\widehat{\mathbf{F}}_N$. Rearranging provides (19) with probability $1-\alpha$. □

We now state a technical lemma bounding the change in expectation of a continuous function of a random variable under distributional shift. The result is 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 3 (Bounding deviations of expectations)**.** *Under Assumption 3 and Assumption 2 (i), suppose $h : \Xi \to \mathbb{R}$ satisfies $|h(\xi) - h(\xi')| \leq \mathbf{L}_h \left\|\xi - \xi'\right\|_2$ for all $\xi, \xi' \in \Xi$. Then for any two distributions $\mathbb{Q}_1, \mathbb{Q}_2$ over Ξ:*

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

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

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

Using Jensen's inequality, $\mathbb{E}_{(\xi,\xi')\sim\pi}[\left\|\xi - \xi'\right\|_2] \leq \sqrt{\mathbb{E}_{(\xi,\xi')\sim\pi}[\|\xi - \xi'\|_2^2]}$. Combining with the lower bound on the cost from assumption 3, we obtain

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

The left-hand side does not depend of the chosen coupling π; taking the infimum of the right-hand side over possible couplings, and using the fact that $\sqrt{\cdot}$ is nondecreasing, we have

$$\inf_\pi \sqrt{\mathbb{E}[c]} = \sqrt{\inf_\pi \mathbb{E}_\pi[c]} = \sqrt{W_c(\mathbb{Q}_1, \mathbb{Q}_2)},$$

concluding the proof. □

We can now extend the generalization bound to a performance guarantee of ERM under distributional shift. For simplicity of exposition, we will consider $\mathbf{x}_{\text{ERM}}$, the exact minimizer of the ERM problem, i.e., with $\delta = 0$ following the notation of theorem 2.

Theorem 3 (ERM robustness to distribution shift)**.** *Retaining the notation of Theorem 2, suppose Assumption 2(i) and Assumption 3 hold, and let* $\mathbf{L}_f$ *be the Lipschitz constant of* $f(\mathbf{x},\cdot)$ *valid at all* $\mathbf{x}\in\mathcal{X}$*:*

$$|f(\mathbf{x},\xi)-f(\mathbf{x},\xi')|\leq \mathbf{L}_f\left\|\xi-\xi'\right\|_2 \quad \forall \xi,\xi'\in\Xi,\mathbf{x}\in\mathcal{X}.$$

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

$$\mathbf{F}_{\mathbb{Q}}(\mathbf{x}_{\mathrm{ERM}})-\mathbf{F}_{\mathbb{Q}}(\mathbf{x}_{\mathbb{Q}})\leq 2\|f\|_\infty\sqrt{\frac{2\log\left(2|\mathcal{X}|/\alpha\right)}{N}}+2\mathbf{L}_f\sqrt{\frac{\varrho}{\mu_c}}. \tag{25}$$

Proof. For any fixed $\mathbf{x}\in\mathcal{X}$, applying Lemma 3 with distributions $\mathbb{Q},\mathbb{P}$ with $W_c(\mathbb{P},\mathbb{Q})\leq\varrho$ and $h=f(\mathbf{x},\cdot)$,

$$|\mathbf{F}(\mathbf{x})-\mathbf{F}_{\mathbb{Q}}(\mathbf{x})|=|\mathbb{E}_{\xi'\sim\mathbb{P}}[f(\mathbf{x},\xi')]-\mathbb{E}_{\xi\sim\mathbb{Q}}[f(\mathbf{x},\xi)]|\leq\mathbf{L}_f\sqrt{\frac{W_c(\mathbb{P},\mathbb{Q})}{\mu_c}}\leq\mathbf{L}_f\sqrt{\frac{\varrho}{\mu_c}}\quad\forall\mathbf{x}\in\mathcal{X}. \tag{26}$$

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

$$\mathscr{E}=\left\{\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{\mathbf{F}}_N(\mathbf{x})-\mathbf{F}(\mathbf{x})|\leq\|f\|_\infty\sqrt{\frac{2\log\left(2|\mathcal{X}|/\alpha\right)}{N}}\right\} \tag{27}$$

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

$$\begin{aligned}
\mathbf{F}_{\mathbb{Q}}(\mathbf{x}_{\mathrm{ERM}})\leq&\mathbf{F}(\mathbf{x}_{\mathrm{ERM}})+\mathbf{L}_f\sqrt{\frac{\varrho}{\mu_c}} && \text{(28a)}\\
\leq&\widehat{\mathbf{F}}_N(\mathbf{x}_{\mathrm{ERM}})+\|f\|_\infty\sqrt{\frac{2\log\left(2|\mathcal{X}|/\alpha\right)}{N}}+\mathbf{L}_f\sqrt{\frac{\varrho}{\mu_c}} && \text{(28b)}\\
\leq&\widehat{\mathbf{F}}_N(\mathbf{x}_{\mathbb{Q}})+\|f\|_\infty\sqrt{\frac{2\log\left(2|\mathcal{X}|/\alpha\right)}{N}}+\mathbf{L}_f\sqrt{\frac{\varrho}{\mu_c}} && \text{(28c)}\\
\leq&\mathbf{F}(\mathbf{x}_{\mathbb{Q}})+2\|f\|_\infty\sqrt{\frac{2\log\left(2|\mathcal{X}|/\alpha\right)}{N}}+\mathbf{L}_f\sqrt{\frac{\varrho}{\mu_c}} && \text{(28d)}\\
\leq&\mathbf{F}_{\mathbb{Q}}(\mathbf{x}_{\mathbb{Q}})+2\|f\|_\infty\sqrt{\frac{2\log\left(2|\mathcal{X}|/\alpha\right)}{N}}+2\mathbf{L}_f\sqrt{\frac{\varrho}{\mu_c}} && \text{(28e)}
\end{aligned}$$

where (28a) comes from applying (26) at $\mathbf{x}_{\mathrm{ERM}}$, (28b) uses the uniform convergence event bound holding at $\mathbf{x}_{\mathrm{ERM}}$ for $\mathbb{P}$, (28c) uses optimality of $\mathbf{x}_{\mathrm{ERM}}$ for $\widehat{\mathbf{F}}_N$, (28d) applies the uniform convergence event bound, this time at $\mathbf{x}_{\mathbb{Q}}$, and (28e) uses (26) a second time at $\mathbf{x}_{\mathbb{Q}}$. □

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 $\mathcal{O}\left(\sqrt{k\log(n)}N^{-1/2}\right)$. This analysis heavily relies on the finiteness of $\mathcal{X}$ and on the union bound, and it leaves open the question of a similar bound for mixed-integer sets.

D Supporting lemmas

For the sake of completeness, we provide proofs of elementary lemmas used in the convergence analysis.

Lemma 4. *Let* $x_1,\ldots,x_N\in\mathbb{R}^d$*. Let* $\mathcal{J}\subseteq[\![N]\!]$ *be a random subset of size* b*, chosen uniformly and independently among all subsets of 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:

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

□

Lemma 5. *Let Assumption 2(i) hold, let* $\mathbb{Q} \in \mathrm{Proba}(\Xi)$ *and* $g : \Xi \to \mathbb{R}$ *a bounded* $\mathbb{Q}$*-measurable 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. 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 its tangent at 1:

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

Applying the inequality at $t = 0$ provides the desired inequality. □