Primal-dual algorithm for contextual stochastic combinatorial optimization

Louis Bouvier, Thibault Prunet, Axel Parmentier, Vincent Leclère

Abstract

The field of contextual stochastic optimization has recently emerged from the renewed interest of combining combinatorial optimization and machine learning techniques to solve decision-making problems in the face of uncertainty. Traditional solution methods for stochastic optimization problems are not able to fully exploit the available contextual information at decision-making time, prompting the need for new algorithmic developments. In this paper, we address this gap with a "learning by experience" perspective. We aim at learning a policy that *minimizes the empirical risk* from a dataset consisting only of past realizations of the uncertain parameter and context. To that end, we introduce a generic primal-dual algorithm for a broad class of combinatorial stochastic optimization problems. In particular, we make minimal assumptions on the structure of the deterministic problem per scenario, and no assumption at all on the objective function.

We extend classical results on Fenchel-Young losses with the introduction of a new regularization by sparse perturbation on the distribution simplex. This allows us to derive tractable updates for the algorithm that only involve computations in the original space, while being able to handle a generic objective function. The linear convergence of the primal-dual algorithm is established in a restricted setting, provided a convexity assumption related to the Jensen gap of the regularizer. Finally, we bound the non-optimality of the returned policy with respect to the empirical risk. Computational experiments on a contextual stochastic version of the minimum weight spanning tree problem show that our algorithm is practically efficient and scalable, achieving similar performance than a sophisticated Lagrangian-based heuristic for a fraction of the computational load.

1 Introduction

1.1 Setting

Contextual stochastic optimization. Let us consider a decision maker whose choice is affected by some random noise $\boldsymbol{\xi} \in \Xi$. The decision maker does not know $\boldsymbol{\xi}$ when he takes his decision, but has access to a realization x of a *context* variable $\mathbf{x}$ correlated to $\boldsymbol{\xi}$. We denote by $\mathcal{X}$ the context space, meaning that $\mathbf{x}$ takes value in $\mathcal{X}$. Based on a context realization x, the decision maker takes a decision y in $\mathcal{Y}(x)$. To that purpose, he chooses a policy π that maps a context realization x to a decision $y \in \mathcal{Y}(x)$. We do not require the policy to be deterministic and can, therefore, see it as a conditional distribution $\pi(y|x)$ over $\mathcal{Y}(x)$ given x. Given a hypothesis class $\mathcal{H}$ for our policy π, our *contextual stochastic optimization* problem (Sadana et al. 2024) aims at finding a policy π that minimizes the *risk* $\mathcal{R}$, which is the expected cost under π.

$$\min_{\pi \in \mathcal{H}} \mathcal{R}(\pi) \quad \text{where} \quad \mathcal{R}(\pi) = \mathbb{E}_{(\mathbf{x},\boldsymbol{\xi}),\mathbf{y}\sim\pi(\cdot|\mathbf{x})}\big[c(\mathbf{x},\mathbf{y},\boldsymbol{\xi})\big]. \tag{1}$$

The expectation is taken with respect to the distribution over $(\mathbf{x}, \mathbf{y}, \boldsymbol{\xi})$ that derives from the joint distribution over $(\mathbf{x}, \boldsymbol{\xi})$ and the policy π. Since the decision maker does not have access to $\boldsymbol{\xi}$, the decision $\mathbf{y}$ is independent of $\boldsymbol{\xi}$ given the context $\mathbf{x}$. We place ourselves in a *learning* setting: we do not know the joint distribution over $(\mathbf{x}, \boldsymbol{\xi})$. But we have access to a training set $(x_1, \xi_1), \ldots, (x_N, \xi_N)$ of independent samples of $(\mathbf{x}, \boldsymbol{\xi})$. In many situations, the noise ξ is observed once the decision has been taken, and the training set comes from historical data. In this work, we focus on the combinatorial case where, for any context realization $x \in \mathcal{X}$, the set of admissible decisions $\mathcal{Y}(x)$ is finite but potentially combinatorially large, as formalized in the following assumption that holds throughout the paper.

Assumption 1. *For every possible context x, the set of admissible decision $\mathcal{Y}(x) \subset \mathbb{R}^{d(x)}$ is finite. Further, we assume that $\mathcal{Y}(x)$ is the set of exposed vertices of its convex envelope $\mathcal{C}(x) = \mathrm{conv}\,(\mathcal{Y}(x))$, i.e., there are no $y \in \mathcal{Y}(x)$ is a strict convex combination of other points in $\mathcal{Y}(x)$.*

This allows us to work with stochastic decision, seen as an element of $\mathcal{C}(x)$, while remaining in finite dimension.

Stochastic optimization policies. Using a stochastic optimization approach, one would typically build a policy π by solving the stochastic optimization problem that arises by taking the conditional expectation over $\boldsymbol{\xi}$ given $\mathbf{x} = x$.

$$\min_{y \in \mathcal{Y}(x)} \mathbb{E}_{\boldsymbol{\xi}}\big[c(\mathbf{x}, y, \boldsymbol{\xi}) \big| \mathbf{x} = x\big]. \tag{2}$$

Practical approaches typically solve a sample average approximation of (2). Decomposition-coordination methods such as progressive hedging or the L-shaped method solve thousands of instances of deterministic single scenario problem

$$\min_{y \in \mathcal{Y}(x)} c\big(x(\omega), y, \xi(\omega)\big) + \langle \theta | y \rangle, \tag{3}$$

where θ is a dual vector, such as a vector of Lagrange multipliers. Our combinatorial and large dimensional setting brings two challenges. First, we do not know the distribution over $(\mathbf{x}, \boldsymbol{\xi})$. We may learn a model, but large dimensional $\mathbf{x}$ and $\boldsymbol{\xi}$ require a large training set, which we do not always have in industrial settings. Second, the computational burden required by such algorithms becomes significant and prevents their application online in a contextual setting, where the computing time is limited.

Learning approach. We therefore propose a change of paradigm. We instead define a hypothesis class $\mathcal{H}_{\mathcal{W}}$ of policies π_w parameterized by w in $\mathcal{W}$. Policies in $\mathcal{H}_{\mathcal{W}}$ are chosen to be fast enough to be used online. And we use the training set within a learning algorithm that seeks a policy π_w in $\mathcal{H}_{\mathcal{W}}$ with a low risk $\mathcal{R}(\pi_w)$. Practically we solve the *empirical risk minimization problem.*

$$\min_{w \in \mathcal{W}} R_N(\pi_w) \quad \text{where} \quad R_N(\pi_w) := \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\mathbf{y} \sim \pi_w(\cdot|\mathbf{x}_i)} \Big[c(x_i, \mathbf{y}, \xi_i) \Big| \pi_w\Big]. \tag{4}$$

Policies based on a combinatorial optimization layer. Working with a combinatorial solution space $\mathcal{Y}(x)$ makes the choice of π challenging. Indeed, there are few statistically relevant while computationally tractable models from a combinatorial set to another. We rely on a combinatorial optimization layer to build such policies. We build upon recent contributions (Blondel et al. 2020; Berthet et al. 2020; Dalle et al. 2022) that derive from the regularized linear optimization problem

$$\max_{y \in \mathcal{C}(x)} \langle \theta | y \rangle - \Omega_{\mathcal{C}(x)}(y), \tag{5}$$

a conditional distribution $p_{\Omega_{\mathcal{C}(x)}}(\cdot|\theta)$ on $\mathcal{Y}(x)$ (see Section 4.1). Here, $\Omega_{\mathcal{C}(x)} : \mathrm{dom}(\Omega_{\mathcal{C}(x)}) \to \mathbb{R}$ is a proper convex lower-semicontinuous regularization function, with $\mathcal{C}(x) \subseteq \mathrm{cl}\left(\mathrm{dom}(\Omega_{\mathcal{C}(x)})\right)$. The simplest such regularization is $\Omega_{\mathcal{C}(x)} = 0$, in which case we obtain a Dirac on one of the argmax. This model is parameterized by θ, the direction of the linear term. In our policies, we use a statistical model φ_w to predict θ from the context x. Model φ_w is typically a neural network parameterized by $w \in \mathcal{W}$ where $\mathcal{W} \subset \mathbb{R}^{n_w}$. In other words, we seek policies in the hypothesis class

$$\mathcal{H}_{\mathcal{W}} = \Big\{\pi_w \colon w \in \mathcal{W}\Big\} \quad \text{where} \quad \pi_w(y|x) = p_{\Omega_{\mathcal{C}(x)}}\big(y|\varphi_w(x)\big), \tag{6}$$

where $\varphi_w : x \in \mathcal{X} \mapsto \theta \in \mathbb{R}^{d(x)}$ is a machine learning predictor.

Alternating minimization algorithm. As in stochastic optimization, we suppose having a reasonably efficient algorithm to solve the deterministic single scenario problem (3). However, instead of using it online within a decomposition-coordination algorithm solving the conditional stochastic optimization problem (2), we use it offline to solve the learning problem (4), which makes the approach scalable.

Our learning algorithm works as follows. It introduces a *surrogate objective*

$$\mathcal{S}\big(w, y_\otimes\big) = \frac{1}{N}\sum_{i=1}^{N} c(x_i, y_i, \xi_i) + \kappa \mathcal{L}(\theta_i, y_i), \quad \text{with} \quad \begin{cases} \theta_i = \varphi_w(x_i), \\ y_\otimes = (y_i)_{i\in[N]} \in \mathcal{Y}_\otimes, \end{cases} \tag{7}$$

where $[N]$ denotes the set $\{1, \ldots, N\}$, $\kappa > 0$ is a positive constant, and $\mathcal{L}$ is a Fenchel-Young loss (Blondel et al. 2020) and

$$\mathcal{Y}_\otimes = \Big\{(y_i)_{i\in[N]} \colon y_i \in \mathcal{Y}(x_i) \text{ for each } i\Big\}.$$

The loss $\mathcal{L}$ measures the difference $\mathcal{L}(\theta, \bar{y})$ between the target $\bar{y}$ and the prediction of y according to $p_\Omega(\cdot|\theta)$. Such a loss is typically convex in θ, non-negative, and equal to 0 if and only if $p_\Omega(\cdot|\theta)$ is a Dirac in $\bar{y}$. Our learning algorithm is an alternating minimization algorithm of the form

$$y_\otimes^{(t+1)} = \operatorname*{argmin}_{y_\otimes} \mathcal{S}(w^{(t)}, y_\otimes), \qquad \text{(Decomposition)}, \tag{8a}$$

$$w^{(t+1)} \in \operatorname*{argmin}_{w\in\mathcal{W}} \mathcal{S}(w, y_\otimes^{(t+1)}), \qquad \text{(Coordination)}, \tag{8b}$$

where $y_\otimes^{(t+1)} = (y_i^{(t+1)})_{i\in[N]}$. In (8a), we do not write that $y_\otimes$ belongs to $\mathcal{Y}_\otimes$ as we optimize in practice on a continuous space that contains $\mathcal{Y}_\otimes$. Indeed, to make this algorithm practical on combinatorial spaces, we need to work on the space of distribution over $\mathcal{Y}(x_i)$, which requires some technical preliminaries. We therefore postpone the precise definition to Section 4. Suffices to say at this point that Step (8a) decomposes per scenario and requires solving deterministic single scenario problems of the form (3), and that the coordination step (8b) amounts to a supervised learning problem in Fenchel-Young loss for which efficient algorithms exist.

1.2 Contributions and paper outline

Our main contribution is to introduce an alternating minimization algorithm (see Section 4) for the surrogate minimization problem, which has several nice properties.

1. This algorithm relies on sampling, stochastic gradient descent, and automatic differentiation to update a model φ_w, and is therefore deep learning-compatible.

2. It is *generic*, and can be applied to any setting (given Remark 1) for which we have algorithms for the deterministic problem (3). It notably provides a generic algorithm to train policies based on neural networks with combinatorial optimization layers for contextual stochastic optimization problems.

3. If the main goal of this algorithm is to learn policies, it can also be applied to find solutions to stochastic optimization problems. We show in Section 4.4 that, in this restricted setting, our algorithm coincides with mirror descent for a surrogate problem. Provided some regularity assumptions, we show in Section 4.5 that the updates converge in value to the optimum of the surrogate in this restricted setting. Provided some regularity assumptions, we also bound the difference between the empirical risk of the solution to the surrogate problem and the optimum of the empirical risk, and thus the non-optimality of the policy returned for the initial problem.

4. Our first numerical experiments on the contextual stochastic version of the two-stage minimum weight spanning tree problem show that our algorithm is practically efficient and scalable in the size of the statistical model φ_w, the size N of the training set, and the dimension of the combinatorial optimization problem. The limiting factor being the size of the deterministic instances that can be solved in Equation (3).

The key challenge we face to define practical versions of these algorithms is to develop regularizations on non-full dimensional polytopes $\mathcal{C}(x)$, and on the distribution simplex over $\mathcal{Y}(x)$. They are detailed in Section 3.

5. Based on the work of Berthet et al. (2020), we introduce a new sparse regularization by perturbation on the distribution simplex over $\mathcal{Y}(x)$. This new regularization is perhaps the key element to obtain a tractable and generic learning algorithm for large combinatorial problems.

6. We highlight several results on Fenchel Young losses (Blondel et al. 2020) on non-full dimensional polytopes $\mathcal{C}(x)$, and on the distribution simplex over $\mathcal{Y}(x)$. We analyse their links with Legendre-type functions, mirror maps and regularizers.

Paper outline. Section 2 is dedicated to a literature review. In Section 3, we study Fenchel-Young losses for rich combinatorial problems. Section 4 focuses on the primal-dual algorithm. In Section 5, we study a convexity property that stems from our analysis of the link between our primal-dual algorithm and mirror descent. We provide some computational experiments in Section 6.

2 Related works

Below, we make several bibliographical remarks. The first ones in Sections 2.1-2.2 are related to the tools we use to derive policies for combinatorial optimization problems. Then, in Section 2.3, we consider some literature linked to the primal-dual algorithms we introduce in this paper to learn such policies.

2.1 Policies with CO layers

Actually, policies π_w in the form of Equation (6) can be useful beyond contextual stochastic optimization. Indeed, they can be seen as mappings between an input instance $x \in \mathcal{X}$, and a solution in a combinatorial space $y \in \mathcal{Y}(x)$. They may lead to efficient heuristics for hard CO problems, such as:

$$\min_{y \in \mathcal{Y}(x)} c^0(x; y). \tag{9}$$

As soon as the (finite) combinatorial space $\mathcal{Y}(x)$ corresponds to the vertices of its convex hull, denoted as $\mathcal{C}(x) = \operatorname{conv}(\mathcal{Y}(x))$, the linear program $\max_{y' \in \mathcal{C}(x)} \langle y' | \theta \rangle$ can be parameterized to output a "good" solution $y \in \mathcal{Y}(x)$ to Problem (9). This assumption is not needed when the cost function c^0 in Equation (9) reaches its minimum on the vertices of $\mathcal{C}(x)$. It is particularly the case when it is linear. Therefore, provided that updates are tractable, the algorithms presented in this work can be applied to learn the weights w in this context, to minimize a slightly different empirical risk defined in Equation (10).

$$\min_{w} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\mathbf{y}_i \sim \pi_w(\cdot | x_i)} \left[c^0\left(x_i; \mathbf{y}_i\right) \right]. \tag{10}$$

2.2 Fenchel-Young losses: non-linearity and non-full dimension

Fenchel-Young losses are a primal-dual version of Bregman divergences. Our algorithms rely on the geometry they induce on the decision space.

The theory of Fenchel-Young losses has been introduced by the structured machine learning community (Blondel et al. 2020). Most applications in this literature rely on linear optimization layers on the polytope $\mathcal{C} = \operatorname{conv}(\mathcal{Y})$. Ranking or top-$k$ applications fall within this scope for instance. If some theory covers the case of non-full dimensional $\mathcal{C}$ (Blondel et al. 2020), some important approaches for us have been studied only in the linear and full dimensional case (Berthet et al. 2020).

In operations research, a broader class of problems may be considered as layers. The objective function may be non-linear, and the combinatorial space $\mathcal{Y}$ may not be easily reduced to a (continuous) polytope. To deal with these difficulties, we consider Fenchel-Young losses on the simplex $\Delta^{\mathcal{Y}}$ of exponentially large dimension $|\mathcal{Y}|$, and study some links with the corresponding "moment" space $\mathcal{C} = \operatorname{conv}(\mathcal{Y})$. We notably extend some results of Blondel et al. (2020) to the non-linear case, and results of Berthet et al. (2020) to the non-linear and non-full dimensional case. Actually, our less stringent setting coincides with the notion of *regularizer* of A. Juditsky et al. (2023). In order to deal with combinatorial problems, we consider this methodology to the case of sub-dimensional space $\mathcal{C}$, and analyse the links with Legendre-type functions (Rockafellar 1970, Chapter 26).

2.3 Mirror descent and alternating minimization algorithms

Alternating minimization algorithms. Considering a two-variable function ϕ and the optimization problem

$$(M) \qquad \min_{y \in \mathcal{Y}, z \in \mathcal{Z}} \phi(y, z),$$

the standard approach to solve (M) is via the alternating minimization algorithm, also known as the Gauss-Seidel iteration scheme. That is, starting from an initial point $(y^{(0)}, z^{(0)})$, the algorithm iteratively minimizes ϕ along one of its coordinates while the other is fixed, leading to the updates

$$y^{(t+1)} \in \operatorname*{argmin}_{y \in \mathcal{Y}} \phi(y, z^{(t)}),$$
$$z^{(t+1)} \in \operatorname*{argmin}_{z \in \mathcal{Z}} \phi(y^{(t+1)}, z).$$

The study of alternating minimization algorithms is not new (Auslender 1971/1972), but the wide range of applications in machine learning and signal processing lead to a revived interest in the recent years. We refer to Wright (2015) for a literature review on the topic. Convergence results for alternating minimization algorithms can be found in various settings. One key assumption to prove convergence is that the minimum is uniquely attained in each step, otherwise the algorithm may cycle without approaching any stationary point (Powell 1973).

Generally, the alternating minimization algorithm is studied in a structured context where $\mathcal{Y}$ and $\mathcal{Z}$ are Euclidean and ϕ is convex. In this setting, Beck and Tetruashvili (2013) were the first to prove a sublinear rate of convergence when ϕ is assumed L-smooth. When the convexity assumption is removed, convergence results exist for proximal version of the algorithm (see, e.g., Attouch et al. (2013) and Bolte et al. (2014)). These results often rely on the objective function satisfying the so-called Kurdyka-Łojasiewicz property. A more general setting, without any structural assumption on $\mathcal{Y}$, $\mathcal{Z}$ nor ϕ is studied by Léger et al. (2023). They proved the convergence of the alternating minimization algorithm when ϕ satisfies the five-point property, first introduced by Csiszár et al. (1984), that is a non-local inequality involving ϕ evaluated at different points.

Mirror descent. The mirror descent algorithm, introduced by Nemirovsky et al. (1983), and analysed by Beck and Teboulle (2003) and Bubeck (2015) to name just a few, aims at solving the following problem:

$$\min_{y \in \mathcal{C}} c(y), \tag{12}$$

where the space $\mathcal{C} \subset \mathbb{R}^d$ is a closed convex set with non-empty interior, the objective function c is a convex Lipschitz-continuous function with respect to a given norm $||\cdot||$, the set of minimizers is not empty, and sub-gradients of c can be easily computed. To do so, it uses a certain *mirror map* Ψ between the primal space of points $y \in \mathcal{C}$ and the dual space of sub-gradients, that exploits the geometry of Problem (12). It can be seen as an extension of the (projected) gradient descent algorithm, recovered by choosing $\Psi = \frac{1}{2}||\cdot||_2^2$ as mirror map. In order to deal with our combinatorial applications, it is convenient to use maps on non-full dimensional polytopes, which requires a slightly less stringent definition than the classic mirror map. We deal with these aspects using tools close to the notion of regularizer, considered recently by A. Juditsky et al. (2023) for their unified mirror descent algorithm. The term "unified" is employed to highlight that the resulting algorithm covers mirror descent and its variant named dual averaging (A. B. Juditsky et al. 2005; Nesterov 2009).

Remark 1. *Sometimes Ψ is called a Bregman potential, and the term of mirror map is used for its gradient $\nabla\Psi$.*

Stochastic mirror descent. The stochastic mirror descent algorithm is a variant of mirror descent, adapted to solve the following kind of problem, which is close to ours, although it is neither combinatorial nor contextual

$$\min_{y \in \mathcal{C}} \mathbb{E}_{\boldsymbol{\xi}}\left[c(y, \boldsymbol{\xi})\right], \tag{13}$$

where it is assumed that we can easily sample the random variable $\boldsymbol{\xi}$. It is based on technical assumptions that vary in the literature. We refer to D'Orazio et al. (2023) and Zhou et al. (2020) for the details, and to Dang et al. (2015) for its block variant. It has many applications, notably in deep learning (Azizan et al. 2022), since it extends the popular stochastic gradient descent. Roughly, at each iteration t of the algorithm, the noise is sampled $\xi^{(t)}$, and mirror descent updates are applied using an estimator of the sub-gradient of the objective function based on $\xi^{(t)}$.

Mirror descent and distributed optimization. To the best of our knowledge, there is not much literature on mini-batch stochastic mirror descent, where several data-points, in our case context-noise pairs, are handled simultaneously. Related questions are however considered in distributed optimization (Li et al. 2018; Yuan et al. 2018; Duchi et al. 2012), where a variant of mirror descent can be applied to solve the following problem:

$$\min_{y \in \mathcal{C}} \sum_{i=1}^{N} c_i(y), \tag{14}$$

where a set of N agents correspond to the nodes $V = \{1 \ldots, N\}$ of a communication graph denoted as $G^{(t)} = (V, E^{(t)})$, where $E^{(t)}$ is a set of communication edges that varies through time. Each agent i has access to a local contribution of the cost function c_i, and can communicate to its direct neighbors in $G^{(t)}$. Forgetting the time-varying communication graph, this problem shares some structure with the sample average approximation of a stochastic optimization problem, or even with our empirical risk in Equation (4). When applying mirror descent in this setting, a local version of the decision variable $y_i^{(t)}$ is updated per node i. Nonetheless, to the best of our knowledge, it is not done in a combinatorial setting.

3 Fenchel Young loss for rich combinatorial problems

Let Ω be a proper l.s.c. convex function. In this section, we consider the *regularized prediction* problem defined as

$$\hat{y}_\Omega(\theta) \in \operatorname*{argmax}_{\mu \in \operatorname{dom}(\Omega)} \langle \theta | \mu \rangle - \Omega(\mu), \tag{15}$$

and introduce some old and new geometric results related to it. We start with some notations.

3.1 Notations

Sets. We denote by $\mathbb{R}$ the set of real numbers, and by $\mathbb{R}_{++}$ the set of positive real numbers. Let E be an Euclidean space, and $\mathcal{X} \subset E$ be a set. We denote by $\operatorname{span}(\mathcal{X})$ the span of $\mathcal{X}$, $\operatorname{aff}(\mathcal{X})$ its affine hull, $\operatorname{int}(\mathcal{X})$ its interior, $\operatorname{cl}(\mathcal{X})$ its closure, $\operatorname{bdry}(\mathcal{X})$ its boundary, and $\operatorname{rel\,int}(\mathcal{X})$ its relative interior. We introduce $\mathbb{I}_{\mathcal{X}} : E \to [-\infty, +\infty]$ the indicator function of the set $\mathcal{X}$, with value 0 over $\mathcal{X}$ and $+\infty$ elsewhere. For two sets $\mathcal{X}_1$ and $\mathcal{X}_2$, we denote by $\mathcal{X}_1 \times \mathcal{X}_2$ their Cartesian product space, and we introduce $\mathcal{X}_1 + \mathcal{X}_2 := \{x_1 + x_2 \,|\, x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2\}$. When in addition $\mathcal{X}_1$ and $\mathcal{X}_2$ are vector subspaces of E, with $\mathcal{X}_1 \cap \mathcal{X}_2 = \{0\}$, we have a direct sum written as $\mathcal{X}_1 \oplus \mathcal{X}_2$. We extend this notation to $S_1 \oplus S_2$ to denote $\{s_1 + s_2 \colon s_1 \in S_1,\, s_2 \in S_2\}$ given two subsets S_1 and S_2 of E (not necessarily vector spaces) such that $\langle s_1 | s_2 \rangle = 0$ for any $s_1 \in S_1$ and $s_2 \in S_2$.

Functions. For E an Euclidean space with inner product $\langle \cdot | \cdot \rangle$ and associated norm $|| \cdot ||$, we denote by $\Gamma_0(E)$ the set of proper lower-semicontinuous (l.s.c.) convex functions from E to $]-\infty, +\infty]$. For a function $\Psi \in \Gamma_0(E)$, we denote by $\operatorname{dom}(\Psi)$ the domain of Ψ, by $\operatorname{argmin} \Psi$ and $\operatorname{argmax} \Psi$ the sets of global minimizers and maximizers of Ψ (possibly empty), by Ψ^* its Fenchel-conjugate function,

$$\begin{array}{cccl} \Psi^* & E & \to & \mathbb{R} \\ & y & \mapsto & \sup_{x \in \operatorname{dom}(\Psi)} \{\langle x | y \rangle - \Psi(x)\}, \end{array}$$

and by $\partial \Psi$ its subdifferential

$$\begin{array}{cccl} \partial \Psi & E & \to & 2^E \\ & x & \mapsto & \{g \in E \,|\, \forall y \in E, \langle y - x | g \rangle + \Psi(x) \leq \Psi(y)\}. \end{array}$$

Let $x \in E$, Ψ is subdifferentiable at x if $\partial\Psi(x) \neq \emptyset$; the elements of $\partial\Psi(x)$ are the subgradients of Ψ at x. If Ψ is differentiable at x, we name $\nabla\Psi(x)$ the gradient of Ψ at x.

3.2 Background on Legendre-type functions and Fenchel-Young losses

Legendre-type functions (Rockafellar 1970, Section 26) A function $\Psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is *Legendre-type* if the following holds

- Ψ is *essentially smooth*, *i.e.*,
    1. $\mathrm{int}(\mathrm{dom}(\Psi))$ is non-empty;
    2. Ψ is differentiable through $\mathrm{int}(\mathrm{dom}(\Psi))$;
    3. $\lim_{i\to\infty} ||\nabla\Psi(\mu_i)|| = +\infty$ for any sequence $(\mu_i)_{i\in\mathbb{N}} \in \big(\mathrm{int}(\mathrm{dom}(\Psi))\big)^{\mathbb{N}}$ converging to a boundary point of $\mathrm{int}(\mathrm{dom}(\Psi))$.
- Ψ is stricly convex over $\mathrm{int}(\mathrm{dom}(\Psi))$.

The following theorem given by Rockafellar (1970, Theorem 26.5) highlights a convenient property of Legendre-type functions.

Theorem 1 (Rockafellar (1970))**.** *Let $\Psi \in \Gamma_0(\mathbb{R}^d)$ be a proper convex l.s.c. function with Fenchel conjugate Ψ^*. Let $\mathcal{D} := \mathrm{int}\,\big(\mathrm{dom}(\Psi)\big)$ and $\mathcal{D}^* := \mathrm{int}\,\big(\mathrm{dom}(\Psi^*)\big)$. Then, Ψ is a convex function of Legendre type if and only if Ψ^* is a convex function of Legendre type. When these conditions hold, the gradient mapping $\nabla\Psi$ is one-to-one from the open convex set $\mathcal{D}$ onto the open convex set $\mathcal{D}^*$, continuous in both directions, and $\nabla\Psi^* = (\nabla\Psi)^{-1}$.*

Mirror maps (A. Juditsky et al. 2023, Definition 2.1). Let $\Psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a function and $\mathcal{C} \subset \mathbb{R}^d$ be a closed convex set. We say that Ψ is a *$\mathcal{C}$-compatible mirror map* if

1. Ψ is lower-semicontinuous and strictly convex,
2. Ψ is differentiable on $\mathrm{int}(\mathrm{dom}(\Psi))$,
3. the gradient of Ψ takes all possible values, *i.e.*, $\nabla\Psi\big(\mathrm{int}(\mathrm{dom}(\Psi))\big) = \mathbb{R}^d$.
4. $\mathcal{C} \subset \mathrm{cl}\,\big(\mathrm{int}(\mathrm{dom}(\Psi))\big)$,
5. $\mathrm{int}(\mathrm{dom}(\Psi)) \cap \mathcal{C} \neq \emptyset$.

Remark 2. *Let $\mathcal{C} \subset \mathbb{R}^d$ be closed convex set, and Ψ be a Legendre-type function such that $\mathcal{C} \subset \mathrm{cl}\,\big(\mathrm{int}(\mathrm{dom}(\Psi))\big)$, $\mathrm{int}(\mathrm{dom}(\Psi)) \cap \mathcal{C} \neq \emptyset$, and $\mathrm{dom}(\Psi^*) = \mathbb{R}^d$. Then Ψ is a $\mathcal{C}$-compatible mirror map.*

Regularizers (A. Juditsky et al. 2023, Definition 2.8). Let $\mathcal{C} \subset \mathbb{R}^d$ be a closed convex set. A function $\Omega : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a *$\mathcal{C}$-pre-regularizer* if it is strictly convex, lower-semicontinuous, and if $\mathrm{cl}(\mathrm{dom}(\Omega)) = \mathcal{C}$. If in addition $\mathrm{dom}(\Omega^*) = \mathbb{R}^d$, then Ω is said to be a *$\mathcal{C}$-regularizer*.

Remark 3. *The previous definition is less restrictive than the one of mirror maps. In particular, regularizers are not necessarily differentiable, and their domains may be sub-dimensional.*

Fenchel-Young losses (Blondel et al. 2020, Definition 2). Let Ω be a proper l.s.c. convex function in $\Gamma_0(\mathbb{R}^d)$, such that the domain of its Fenchel conjugate is $\mathrm{dom}(\Omega^*) = \mathbb{R}^d$. The Fenchel-Young loss generated by Ω, and denoted as $\mathcal{L}_\Omega : \mathrm{dom}(\Omega^*) \times \mathrm{dom}(\Omega) \to \mathbb{R}_+$ is defined as:

$$\mathcal{L}_\Omega(\theta; y) := \Omega^*(\theta) + \Omega(y) - \langle \theta | y \rangle. \tag{16}$$

The following proposition summarizes known properties that are direct consequences of classic convex optimization results. Apart from the last point, it corresponds to Proposition 2 of the study by Blondel et al. (2020).

Proposition 1. *(Blondel et al. 2020) Let $\Omega \in \Gamma_0(\mathbb{R}^d)$ be a proper l.s.c. convex function, such that the domain of its Fenchel conjugate is $\mathrm{dom}(\Omega^*) = \mathbb{R}^d$. Let (y, θ) be in $(\mathbb{R}^d)^2$. We have the following properties.*

1. $\mathcal{L}_\Omega(\theta; y) \geq 0$.
2. $\mathcal{L}_\Omega(\theta; y) = 0 \iff \theta \in \partial\Omega(y) \iff y \in \partial\Omega^*(\theta)$.
3. *$\mathcal{L}_\Omega(\cdot; y)$ is convex, and $\hat{y}_\Omega(\theta) - y \in \partial_\theta \mathcal{L}_\Omega(\theta; y)$.*
4. *If in addition Ω is strictly convex, the* argmax *in Equation (15) is unique, Ω^* is differentiable, $\mathcal{L}_\Omega$ is differentiable with respect to θ, and we have*

$$\mathcal{L}_\Omega(\theta; y) = 0 \iff y = \nabla\Omega^*(\theta) = \hat{y}_\Omega(\theta). \tag{17}$$

5. *If in addition Ω is Legendre-type, we have*

$$\nabla\Omega(\nabla\Omega^*(\theta)) = \theta, \quad \forall\theta \in \mathrm{dom}(\Omega^*). \tag{18}$$

Elements of proof. Point 1 is known as the Fenchel-Young inequality. See e.g. the work of Rockafellar (1970, Theorem 23.5) for point 2. Point 3 follows from the definition of $\hat{y}_\Omega$. Point 4 follows from point 2 and the unicity of the argmax. The last point follows from point 2 in the case both Ω and Ω^* are differentiable, since Ω is a Legendre-type function. □

Examples. We give several examples of regularization functions Ω, and associated regularized predictions.

1. Let us consider the case when $\Omega(y) = \frac{1}{2}||y||_2^2$ and $\mathrm{dom}(\Omega) = \mathbb{R}^d$. The regularized prediction function is computed as in Equation (15) and it is easy to see that it produces $\hat{y}_\Omega(\theta) = \theta$. The associated Fenchel-Young loss is then $\mathcal{L}_\Omega(\theta; y) = \frac{1}{2}||y - \theta||_2^2$.

2. We may typically want to restrict our prediction $\hat{y}_\Omega(\theta)$ to some polytope $\mathcal{C}$. In that case, we can use $\Omega(y) = \psi(y) + \mathbb{I}_\mathcal{C}(y)$, where $\psi \in \Gamma_0(\mathbb{R}^d)$ is a proper l.s.c. convex function, and $\mathbb{I}_\mathcal{C}(y)$ is the indicator function of $\mathcal{C}$, equal to 0 if y belongs to $\mathcal{C}$ and $+\infty$ otherwise. We get

$$\hat{y}_\Omega(\theta) \in \operatorname*{argmax}_{\mu \in \mathrm{dom}(\Omega)} \langle \theta | \mu \rangle - \Omega(\mu) = \operatorname*{argmax}_{\mu \in \mathcal{C}} \langle \theta | \mu \rangle - \psi(\mu).$$

3. As an example of the previous case, we may want to use the regularization $\frac{1}{2}\|y\|_2^2$ to make predictions on the simplex $\Delta^d := \{q \in \mathbb{R}^d, q \geq 0, \sum_{i \in [d]} q_i = 1\}$. Remark that this set is sometimes denoted as Δ^{d-1}, to highlight that it is of dimension $d - 1$ in $\mathbb{R}^d$. To be consistent with the machine learning literature, and with the notation $\Delta^\mathcal{Y}$ below, we keep

the notation Δ^d. We define $\Omega(y) = \frac{1}{2}||y||_2^2 + \mathbb{I}_{\Delta^d}(y)$, the domain under consideration is $\text{dom}(\Omega) = \Delta^d$, and the regularized prediction coincides with the *sparsemax*

$$\hat{y}_\Omega(\theta) = \underset{y \in \Delta^d}{\text{argmin}}\, ||\theta - y||_2 = \text{sparsemax}(\theta).$$

The associated Fenchel-Young loss is then

$$\mathcal{L}_\Omega(\theta; y) = \frac{1}{2}||y - \theta||_2^2 - \frac{1}{2}||\theta - \hat{y}_\Omega(\theta)||_2^2.$$

4. Last, consider the negative entropy defined as $\Omega(y) = \sum_{i=1}^d y_i \log(y_i) + \mathbb{I}_{\Delta^d}(y)$ with domain $\text{dom}(\Omega) = \Delta^d$. The corresponding regularized prediction is $\hat{y}_\Omega(\theta) = \text{softmax}(\theta)$, where

$$\text{softmax}(\theta)_j = \frac{e^{\theta_j}}{\sum_{i \in [d]} e^{\theta_i}}.$$

We stress that the choice of the domain affects the computation of the corresponding regularized prediction function. This fact will be useful later on.

3.3 Regularized prediction on non-full-dimensional spaces.

To the best of our knowledge, most of the theory of Fenchel-Young losses has been made under the assumption that either $\text{dom}(\Omega) = \mathcal{C}$ is full-dimensional, and Ω is Legendre-type, or $\Omega := \Psi + \mathbb{I}_\mathcal{C}$, where Ψ is Legendre-type. In many applications in operations research, we consider polytopes that are not full dimensional (take the simplex as a case in point). When defining Ω directly on the polytope (using a perturbation (Berthet et al. 2020) for instance), the decomposition $\Omega := \Psi + \mathbb{I}_\mathcal{C}$ is not given. Therefore, when the polytope is not full dimensional, we do not have access to a Legendre-type function. Instead, we have at our disposal a $\mathcal{C}$-regularizer. We further study this case in terms of convex analysis, and show that actually a certain Legendre-type function Ψ can be exhibited.

Proposition 2. *Let $\mathcal{C} \subset \mathbb{R}^d$ be a convex compact set. We consider a proper l.s.c. convex regularization function $\Omega \in \Gamma_0(\mathbb{R}^d)$ with domain $\text{dom}(\Omega) = \mathcal{C}$. We assume that the restriction of Ω to $H = \text{aff}(\mathcal{C})$, denoted as $\Omega_{|H}$, is Legendre-type (with respect to the metric of H, and not the one of $\mathbb{R}^d$). These assumptions imply that in particular, Ω is a $\mathcal{C}$-pre-regularizer. We denote by V the direction of H in $\mathbb{R}^d$, and we have the orthogonal sum $\mathbb{R}^d = V \oplus V^\perp$. We introduce Π_V, the linear orthogonal projection onto V in $\mathbb{R}^d$. We show the following results:*

1. *The Fenchel conjugate of Ω, has full domain, i.e., $\text{dom}(\Omega^*) = \mathbb{R}^d$. Therefore, Ω is a $\mathcal{C}$-regularizer. The function Ω^* is differentiable over $\mathbb{R}^d$, and we have the property:*

$$\nabla\Omega^*(\partial\Omega(y)) = y, \quad \forall y \in \text{rel}\,\text{int}(\mathcal{C}). \tag{19}$$

2. *Let $\theta \in \mathbb{R}^d$, decomposed as $\theta = \theta_V + \theta_{V^\perp}$, where $\theta_V = \Pi_V(\theta)$ and $\theta_{V^\perp} = \theta - \theta_V$, and $y_0 \in \mathcal{C}$ be any point in $\mathcal{C}$. The Fenchel conjugate of Ω, denoted as Ω^*, has an affine component over $V^\perp$:*

$$\Omega^*(\theta) = \Omega^*(\theta_V) + \langle\theta_{V^\perp}|y_0\rangle. \tag{20}$$

3. *Let $y \in H$, the subdifferential of Ω at y is given by:*

$$\partial\Omega(y) = \partial(\Omega_{|H})(y) + V^\perp, \tag{21}$$

where we have omitted the canonical injection from H to $\mathbb{R}^d$ for notational simplicity. In particular, for $y \in \operatorname{rel\,int}(\operatorname{dom}(\Omega))$, we have:

$$\partial\Omega(y) = \{\nabla\Omega_{|H}(y)\} + V^{\perp}. \tag{22}$$

We illustrate Proposition 2 in Figure 1, in the case $d = 3$, H is an affine hyperplane, and $V^{\perp}$ a straight line. Arrows represent the links between primal and dual variables, involving the subdifferential of Ω and the gradient of Ω^*.

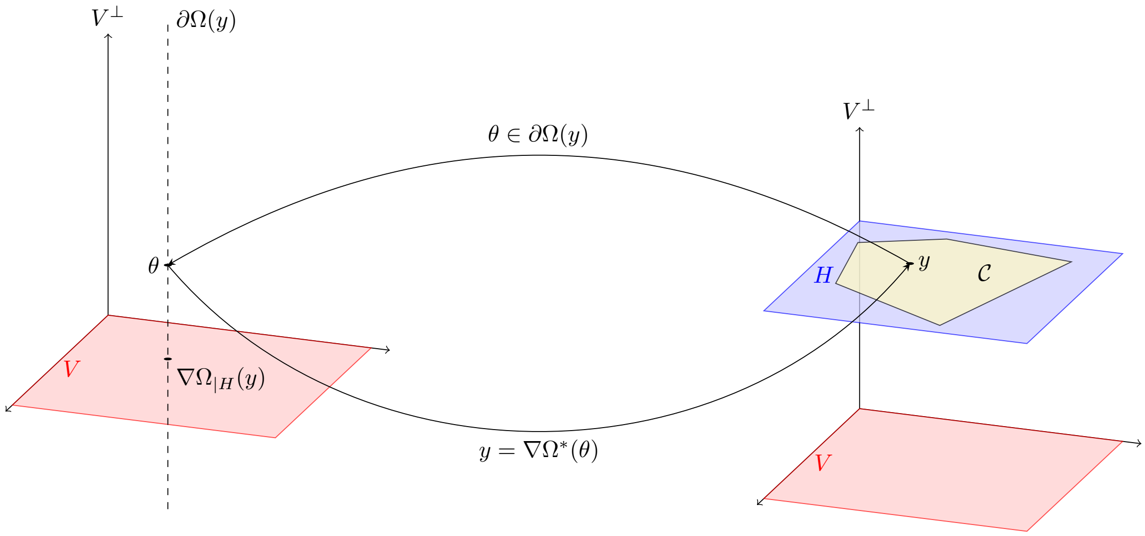

Figure 1: Primal-dual maps for non-full-dimensional $\operatorname{dom}(\Omega)$.

Proof of Proposition 2. Given the assumptions in the preamble of the proposition,

1. The function Ω belongs to the set of proper l.s.c. convex functions $\Gamma_0(\mathbb{R}^d)$, thus for any $\theta \in \mathbb{R}^d$, the supremum over the compact $\mathcal{C}$ of $\langle\theta|\cdot\rangle - \Omega(\cdot)$ is finite and attained, thus $\operatorname{dom}(\Omega^*) = \mathbb{R}^d$. Recall that, as Ω is in $\Gamma_0(\mathbb{R}^d)$, using the computations of Rockafellar (1970, Theorem 23.5), $\partial\Omega^*(\theta) = \operatorname{argmax}_y \langle\theta|y\rangle - \Omega(y)$. As Ω is strictly convex, the argmax is reduced to a single point, and Ω^* is differentiable over $\mathbb{R}^d$. Therefore, we have by Proposition 1.2, for $(y, \theta) \in (\mathbb{R}^d)^2$:

$$\theta \in \partial\Omega(y) \iff y \in \partial\Omega^*(\theta) \iff y = \nabla\Omega^*(\theta).$$

   Therefore, for $y \in \operatorname{rel\,int}(\mathcal{C})$, we have $\nabla\Omega^*(\partial\Omega(y)) = y$.

2. Let $y_0 \in \mathcal{C}$ and $\theta \in \mathbb{R}^d$ be decomposed as $\theta = \theta_V + \theta_{V^{\perp}}$. Note that for all $y \in \mathcal{C}$ we have $\langle\theta_{V^{\perp}}|y\rangle = \langle\theta_{V^{\perp}}|y_0\rangle$, since $\theta_{V^{\perp}}$ is orthogonal to the direction of the affine hull of $\mathcal{C}$. Thus,

$$\Omega^*(\theta) = \sup_{y\in\mathcal{C}} \langle\theta_V + \theta_{V^{\perp}}|y\rangle - \Omega(y) = \langle\theta_{V^{\perp}}|y_0\rangle + \sup_{y\in\mathcal{C}} \langle\theta_V|y\rangle - \Omega(y),$$

   which yields the result.

3. Let $(y, y') \in H^2$, $\theta \in \partial\Omega(y)$, by definition of the subgradients, we have

$$\Omega(y') - \Omega(y) \geq \langle\theta|y' - y\rangle = \langle\Pi_V(\theta)|y' - y\rangle,$$

   since $y' - y$ belongs to V. Therefore we have shown that $\Pi_V(\theta) \in \partial(\Omega_{|H})(y)$.

Conversely, let $y \in H$ and $\theta \in V$ be an element of $\partial(\Omega_{|H})(y)$, for $y' \in \mathbb{R}^d$ and $\tilde{\theta} \in V^\perp$,

$$\Omega(y') \geq \Omega(y) + \langle \theta + \tilde{\theta} | y' - y \rangle,$$

since either $y' \notin H$ and $\Omega(y') = +\infty$, or $y' \in H$ and $y' - y \in V$ therefore $\langle \tilde{\theta} | y' - y \rangle = 0$. We have shown the first equality in Equation (21).

Equation (22) comes from the fact that $\Omega_{|H}$ is Legendre-type, thus differentiable, and for $y \in \operatorname{rel\,int}(\operatorname{dom}(\Omega))$, $\partial(\Omega_{|H})(y) = \{\nabla(\Omega_{|H})(y)\}$.

□

The following proposition, which is new, enables to connect the non-full-dimensional case to results studied by Blondel et al. (2020).

Proposition 3. *Let $\mathcal{C} \subset \mathbb{R}^d$ be a convex compact set, and Ω be a proper l.s.c. convex regularization function in $\Gamma_0(\mathbb{R}^d)$, with domain $\operatorname{dom}(\Omega) = \mathcal{C}$. We assume that the restriction of Ω to $H = \operatorname{aff}(\mathcal{C})$ is Legendre-type (with respect to the metric of H). Then, there exists a Legendre-type function Ψ, with $\mathcal{C} \subset \operatorname{cl}\big(\operatorname{int}(\operatorname{dom}(\Psi))\big)$, $\operatorname{int}(\operatorname{dom}(\Psi)) \cap \mathcal{C} \neq \emptyset$, and such that*

$$\Omega = \Psi + \mathbb{I}_{\mathcal{C}}, \quad \textit{and} \quad \operatorname{dom}(\Psi^*) = \mathbb{R}^d.$$

Besides, let V be the direction of H in $\mathbb{R}^d$, we have the direct sum $\mathbb{R}^d = V \oplus V^\perp$. Given a vector $\theta \in \mathbb{R}^d$, there exists a vector $z \in V^\perp$ such that

$$\nabla\Psi\big(\nabla\Omega^*(\theta)\big) = \theta + z.$$

This result can be seen as the converse, in a restricted setting (with more assumptions on Ω), of a proposition by A. Juditsky et al. (2023, Proposition 2.11), where given a $\mathcal{C}$-compatible mirror map Ψ, a $\mathcal{C}$-regularizer Ω is defined as $\Omega := \Psi + \mathbb{I}_{\mathcal{C}}$. Indeed, we know with Proposition 2 that Ω defined in the preamble of Proposition 3 is (in particular) a $\mathcal{C}$-regularizer. We use Proposition 3 in Section 4.4 to study some links between our primal-dual algorithms and mirror descent.

Proof. Let Π_V be the linear orthogonal projection onto V. W.l.o.g., we consider the case $H = V$, which means the affine hull of the domain of Ω is actually a vector subspace in $\mathbb{R}^d$. Extending to the affine case involves a translation. We define the following application:

$$\begin{aligned} \Psi :\ & \mathbb{R}^d \to \mathbb{R} \\ & y \mapsto \Omega(\Pi_V(y)) + \frac{1}{2}||y - \Pi_V(y)||_2^2, \end{aligned}$$

where $\Pi_V(y)$ is seen as an element of $\mathbb{R}^d$ here. When it is the input of the restriction of Ω to V, we see it as an element of V. Notice that by definition, $\Omega = \Psi + \mathbb{I}_{\mathcal{C}}$. We are going to prove that Ψ is a Legendre-type function. We first show that Ψ defined as such is essentially smooth by checking the three properties of the definition.

1. Since $\operatorname{dom}(\Omega) = \mathcal{C} \subset V$ and $\|\cdot\|_2^2$ is defined over $\mathbb{R}^d$, the domain of Ψ is $\operatorname{dom}(\Psi) = \mathcal{C} \oplus V^\perp$, and $\operatorname{int}(\operatorname{dom}(\Psi)) = \operatorname{rel\,int}(\mathcal{C}) \oplus V^\perp$, which is not empty. We therefore have

$$\mathcal{C} \subset \operatorname{cl}\big(\operatorname{int}(\operatorname{dom}(\Psi))\big), \quad \text{and} \quad \operatorname{int}(\operatorname{dom}(\Psi)) \cap \mathcal{C} \neq \emptyset.$$

2. By composition with linear projections and sum, Ψ is differentiable over $\operatorname{int}(\operatorname{dom}(\Psi))$. We denote by J_{Π_V} the Jacobian of Π_V, that can be seen as the canonical injection of V into $\mathbb{R}^d$. Let now $(y,h) \in \operatorname{int}(\operatorname{dom}(\Psi)) \times \mathbb{R}^d$ be two vectors such that $y+h \in \operatorname{dom}(\Psi)$,

$$\begin{aligned}
\Psi(y+h) &= \Omega(\Pi_V(y+h)) + \frac{1}{2}||y+h-\Pi_V(y+h)||_2^2, \\
&= \Omega(\Pi_V(y)+\Pi_V(h)) + \frac{1}{2}||y-\Pi_V(y)+h-\Pi_V(h)||_2^2, \\
&= \Omega_{|V}(\Pi_V(y)) + \langle J_{\Pi_V}\nabla(\Omega_{|V})(\Pi_V(y))|\Pi_V(h)\rangle + o(||\Pi_V(h)||), \\
&+ \frac{1}{2}||y-\Pi_V(y)||_2^2 + \langle y-\Pi_V(y)|h-\Pi_V(h)\rangle + o(||h-\Pi_V(h)||), \\
&= \Omega(\Pi_V(y)) + \langle J_{\Pi_V}\nabla(\Omega_{|V})(\Pi_V(y))|\Pi_V(h)+h-\Pi_V(h)\rangle + o(||h||), \\
&+ \frac{1}{2}||y-\Pi_V(y)||_2^2 + \langle y-\Pi_V(y)|h-\Pi_V(h)+\Pi_V(h)\rangle + o(||h||), \\
&= \Psi(y) + \langle J_{\Pi_V}\nabla(\Omega_{|V})(\Pi_V(y)) + y - \Pi_V(y)|h\rangle + o(||h||).
\end{aligned}$$

In the computations above we use the linearity of Π_V, the fact that Ω and $\Omega_{|V}$ coincide over V, that $\Omega_{|V}$ is Legendre-type thus differentiable, and the orthogonal sum $\mathbb{R}^d = V \oplus V^\perp$. Therefore, we have shown that the gradient of Ψ is given by:

$$\nabla\Psi(y) = \underbrace{J_{\Pi_V}}_{\substack{\text{Canonical}\\ \text{injection}\\ V\to\mathbb{R}^d}} \nabla\Omega_{|V}(\Pi_V(y)) + y - \Pi_V(y).$$

3. The boundary of $\operatorname{int}(\operatorname{dom}(\Psi))$ is

$$\operatorname{bdry}\big(\operatorname{int}(\operatorname{dom}(\Psi))\big) = \operatorname{cl}\big(\operatorname{rel\,int}(\mathcal{C})\big)\setminus \operatorname{rel\,int}(\mathcal{C}) \oplus V^\perp.$$

Indeed, $\operatorname{int}(\operatorname{dom}(\Psi)) = \operatorname{rel\,int}(\mathcal{C}) \oplus V^\perp$ is isomorphic to $\operatorname{rel\,int}(\mathcal{C}) \times V^\perp$, $\operatorname{bdry}(V^\perp) = \emptyset$, $\operatorname{cl}(V^\perp) = V^\perp$, and for two sets S_1 and S_2

$$\operatorname{bdry}(S_1 \times S_2) = \big(\operatorname{bdry}(S_1) \times \operatorname{cl}(S_2)\big) \cup \big(\operatorname{cl}(S_1) \times \operatorname{bdry}(S_2)\big),$$

where the boundaries in the right-hand side above are computed with respect to the topology corresponding to each set.

Let now μ be in $\operatorname{bdry}\big(\operatorname{int}(\operatorname{dom}(\Psi))\big)$, and let $(\mu_i)_{i\in\mathbb{N}}$ be a sequence in $\big(\operatorname{int}(\operatorname{dom}(\Psi))\big)^{\mathbb{N}}$, such that

$$\lim_{i\to+\infty} \mu_i = \mu = \underbrace{\Pi_V(\mu)}_{\in \operatorname{cl}\big(\operatorname{rel\,int}(\mathcal{C})\big)\setminus \operatorname{rel\,int}(\mathcal{C})} + \underbrace{\mu - \Pi_V(\mu)}_{\in V^\perp}.$$

Since Π_V is continuous,

$$\lim_{i\to+\infty} \underbrace{\Pi_V(\mu_i)}_{\in \operatorname{rel\,int}(\mathcal{C})} = \Pi_V(\mu), \quad \text{and} \quad \lim_{i\to+\infty} \underbrace{\mu_i - \Pi_V(\mu_i)}_{\in V^\perp} = \mu - \Pi_V(\mu).$$

Now, using the fact that $\Omega_{|V}$ is Legendre-type, the expression of $\nabla\Psi$ above, and the reverse triangular inequality,

$$\begin{aligned}
||\nabla\Psi(\mu_i)|| &= ||J_{\Pi_V}\nabla(\Omega_{|V})(\Pi_V(\mu_i)) + \mu_i - \Pi_V(\mu_i)||, \\
&\geq \Big|\underbrace{||J_{\Pi_V}\nabla(\Omega_{|V})(\Pi_V(\mu_i))||}_{\to+\infty} - \underbrace{||\mu_i - \Pi_V(\mu_i)||}_{\to||\mu-\Pi_V(\mu)||}\Big|.
\end{aligned}$$

Therefore, we have shown that $\lim_{i\to+\infty} ||\nabla\Psi(\mu_i)|| = +\infty$.

Let us finally show that Ψ is strictly convex. We first remark that $\mathrm{dom}(\Psi)$ is convex since both $\mathcal{C}$ and $V^\perp$ are. Let (y_1, y_2) be in $(\mathrm{dom}(\Psi))^2$, with $y_1 \neq y_2$, and let t be in $(0, 1)$. To ease notations, we denote $y_i^V = \Pi_V(y_i)$, and $y_i^\perp = y_i - \Pi_V(y_i)$,

$$\begin{aligned}\Psi(ty_1 + (1-t)y_2) &= \Omega(ty_1^V + (1-t)y_2^V) + \frac{1}{2}||ty_1^\perp + (1-t)y_2^\perp||_2^2,\\ &< t\Omega(y_1^V) + (1-t)\Omega(y_2^V) + t\frac{1}{2}\|y_1^\perp\|^2 + (1-t)\frac{1}{2}\|y_1^\perp\|^2,\\ &= t\Psi(y_1) + (1-t)\Psi(y_2).\end{aligned}$$

The first line is by linearity of the orthogonal projection onto V. Further, as $y_1 \neq y_2$ we have $y_1^V \neq y_2^V$ or $y_1^\perp \neq y_2^\perp$, thus the strict convexity of Ω over $\mathcal{C}$ and of $\|\cdot\|_2^2$ yields the second line. We have therefore shown that Ψ is a Legendre-type function.

We now consider its Fenchel conjugate Ψ^*, and study its domain. Let $\theta \in \mathbb{R}^d$, decomposed as $\theta = \theta_V + \theta_{V^\perp}$, where $\theta_V = \Pi_V(\theta)$ and $\theta_{V^\perp} = \theta - \theta_V$,

$$\Psi^*(\theta) = \sup_{y\in\mathbb{R}^d}\{\langle\theta|y\rangle - \Psi(y)\}, \tag{23a}$$

$$= \sup_{y\in\mathbb{R}^d}\Big\{\langle\theta_V|\Pi_V(y)\rangle - \Omega(\Pi_V(y)) + \langle\theta_{V^\perp}|y - \Pi_V(y)\rangle - \frac{1}{2}||y - \Pi_V(y)||_2^2\Big\}, \tag{23b}$$

$$= \sup_{\substack{y_V\in V,\\ y_{V^\perp}\in V^\perp}}\Big\{\langle\theta_V|y_V\rangle - \Omega(y_V) + \langle\theta_{V^\perp}|y_{V^\perp}\rangle - \frac{1}{2}||y_{V^\perp}||_2^2\Big\}, \tag{23c}$$

$$= \Omega^*(\theta_V) + \frac{1}{2}||\theta_{V^\perp}||_2^2. \tag{23d}$$

Now, since Ω^* has full domain using Proposition 2, we have $\mathrm{dom}(\Psi^*) = \mathbb{R}^d$.

We eventually show the property on the composition of the gradients of Ψ and Ω^*. First, we highlight that by Proposition 2, Ω^* is indeed differentiable over $\mathbb{R}^d$. Its gradient corresponds to the regularized prediction defined by Equation (15), and belongs to the relative interior of the convex compact set $\mathcal{C}$. Let $\theta \in \mathbb{R}^d$ be a vector decomposed as $\theta = \theta_V + \theta_{V^\perp}$, where $\theta_V = \Pi_V(\theta)$ and $\theta_{V^\perp} = \theta - \theta_V$. Then we have

$$\nabla\Omega^*(\theta) = \nabla\Omega^*(\theta_V) = \underbrace{J_{\Pi_V}}_{\substack{\text{Canonical}\\ \text{injection}\\ V\to\mathbb{R}^d}} \nabla\Omega^*_{|V}(\theta_V).$$

Therefore, applying the gradient of Ψ leads to

$$\begin{aligned}\nabla\Psi\big(\nabla\Omega^*(\theta)\big) &= \nabla\Psi\big(\underbrace{J_{\Pi_V}\nabla\Omega^*_{|V}(\theta_V)}_{\in\mathrm{rel\,int}(\mathcal{C})} + \underbrace{0}_{\in V^\perp}\big),\\ &= J_{\Pi_V}\nabla\Omega_{|V}\big(\nabla\Omega^*_{|V}(\theta_V)\big) + 0,\\ &= \theta_V = \theta \underbrace{-\theta_{V^\perp}}_{z\in V^\perp}.\end{aligned}$$

In the computations above, we use the expression of the gradient of Ψ, and the fact that the restriction $\Omega_{|V}$ of Ω to V is a Legendre-type function with Fenchel conjugate $\Omega^*_{|V}$. Therefore, we have shown that there exists a vector $z \in V^\perp$ such that $\nabla\Psi\big(\nabla\Omega^*(\theta)\big) = \theta + z$. □

Fenchel-Young losses and Bregman divergences. The *Bregman divergence* B_Ψ generated by a strictly convex and differentiable function Ψ is defined as the difference at y between Ψ and its linearization around μ.

$$B_\Psi\colon \begin{array}{rcl} \operatorname{dom}(\Psi) \times \operatorname{rel\,int}(\operatorname{dom}(\Psi)) & \to & \mathbb{R}_+ \\ (y, \mu) & \mapsto & B_\Psi(y\|\mu) = \Psi(y) - \Psi(\mu) - \langle \nabla\Psi(\mu) | y - \mu \rangle. \end{array} \tag{24}$$

Bregman divergences measure the difference between y and μ. They generalize the squared Euclidean distance, which corresponds to $\Psi = \frac{1}{2}\|\cdot\|_2^2$.

Consider the case of a non-full-dimensional space $\mathcal{C}$, where Ω is defined as in Proposition 3. The function Ω defined as such is strictly convex over its domain, but not Legendre-type generally, because it is not essentially smooth (criterion 3). Besides, when $\mathcal{C}$ is not full-dimensional, the criterion 1 does not hold either. Nonetheless, we can reinterpret the regularized prediction as a Bregman projection using the following result (Blondel et al. 2020, Proposition 3).

Proposition 4 (Blondel et al. (2020))**.** *Let $\mathcal{C} \subset \mathbb{R}^d$ be a convex compact set, and Ψ be a Legendre-type function with $\operatorname{dom}(\Psi^*) = \mathbb{R}^d$, such that $\mathcal{C} \subset \operatorname{dom}(\Psi)$. Consider the proper l.s.c. convex function $\Omega \in \Gamma_0(\mathbb{R}^d)$ defined as*

$$\Omega := \Psi + \mathbb{I}_{\mathcal{C}}.$$

Let eventually θ be in $\mathbb{R}^d$ and y in $\mathcal{C}$.

1. *The regularized prediction in Equation (15) is a Bregman projection,*

$$\hat{y}_\Omega(\theta) = \operatorname*{argmax}_{\mu \in \mathcal{C}} \langle \theta | \mu \rangle - \Psi(\mu) = \operatorname*{argmin}_{\mu \in \mathcal{C}} B_\Psi(\mu || \hat{y}_\Psi(\theta)). \tag{25}$$

2. *The Fenchel-Young loss generated by Ω is a difference of Bregman divergences,*

$$\mathcal{L}_\Omega(\theta; y) = B_\Psi(y || \hat{y}_\Psi(\theta)) - B_\Psi(\hat{y}_\Omega(\theta) || \hat{y}_\Psi(\theta)). \tag{26}$$

3. *The Fenchel-Young loss generated by Ω admits the following lower bound,*

$$0 \leq B_\Psi(y || \hat{y}_\Omega(\theta)) \leq \mathcal{L}_\Omega(\theta; y). \tag{27}$$

4. *Eventually, the Fenchel-Young loss characterizes when the Bregman divergence is* 0*,*

$$\hat{y}_\Omega(\theta) = y \iff \mathcal{L}_\Omega(\theta; y) = 0 \iff B_\Psi(y || \hat{y}_\Omega(\theta)) = 0. \tag{28}$$

3.4 Regularizing in the distribution space

Until now in this paper, the Fenchel-Young loss has only been introduced in the context of the regularization (see Equation (15)) of a linear optimization problem $\max_{\mu \in \mathcal{C}} \langle \theta | \mu \rangle$. However, to deal with arbitrary minimization problems

$$\min_{y \in \mathcal{Y}} c(y),$$

on a finite but combinatorial set $\mathcal{Y}$, it is convenient to consider regularization on distributions. Contrary to Sections 3.2-3.3, we now base the definition of the convex compact set $\mathcal{C}$ on $\mathcal{Y}$.

Moment polytope. Let $\mathcal{Y}$ be a finite combinatorial set in $\mathbb{R}^d$, and $\mathcal{C} = \operatorname{conv}(\mathcal{Y})$ be its convex hull. As in the introduction, we make the assumption that no element of $\mathcal{Y}$ is a strict convex combination of other elements of $\mathcal{Y}$. In other words, $\mathcal{Y}$ is the set of vertices of the polytope $\mathcal{C}$. We denote by $H = \operatorname{aff}(\mathcal{Y})$ the affine hull of $\mathcal{Y}$, and by V the direction of H, a sub-vector space in $\mathbb{R}^d$. We have the orthogonal sum $\mathbb{R}^d = V \oplus V^\perp$, and we denote by Π_V the orthogonal projection onto V in $\mathbb{R}^d$. We name Y the wide matrix with vectors $y \in \mathcal{Y}$ as columns.

Distribution polytope. Let $\Delta^{\mathcal{Y}} := \{q \in \mathbb{R}^{|\mathcal{Y}|}, q \geq 0, \sum_{y\in\mathcal{Y}} q_y = 1\}$ be the probability simplex whose vertices are indexed by $\mathcal{Y}$, and H_Δ its affine hull $H_\Delta = \operatorname{aff}(\Delta^{\mathcal{Y}})$. We denote by V_Δ the vector subspace (hyperplane) in $\mathbb{R}^{|\mathcal{Y}|}$ which is the direction of H_Δ. As previously, we rely on the orthogonal sum $\mathbb{R}^{|\mathcal{Y}|} = V_\Delta \oplus V_\Delta^\perp$, where here $V_\Delta^\perp = \operatorname{span}(\mathbf{1})$. Let $\theta \in \mathbb{R}^d$ be a cost vector and $q \in \Delta^{\mathcal{Y}}$ be a probability distribution, then $s_\theta = Y^\top\theta \in \mathbb{R}^{|\mathcal{Y}|}$ is the vector $(y^\top\theta)_{y\in\mathcal{Y}}$, and $\mu_q = Yq = \sum_y q_y y = \mathbb{E}(\mathbf{y}|q)$ is the moment vector of the random variable $\mathbf{y}$ on $\mathcal{Y}$ with distribution q. The proposition below explores regularization on the distribution polytope.

Proposition 5. *Let $\Omega_{\Delta^{\mathcal{Y}}} \in \Gamma_0(\mathbb{R}^{|\mathcal{Y}|})$ be a proper l.s.c. convex function with domain $\Delta^{\mathcal{Y}}$. We drop the $\mathcal{Y}$ in the notation Ω_Δ when $\mathcal{Y}$ is clear from context. We assume that the restriction of Ω_Δ to H_Δ, denoted as $\Omega_{\Delta|H_\Delta}$ is Legendre-type (with respect to the metric of H_Δ, and not the one of $\mathbb{R}^{|\mathcal{Y}|}$). For $\mu \in \mathcal{C}$, we define*

$$\Omega_{\mathcal{C}}(\mu) := \min\{\Omega_\Delta(q) \colon Yq = \mu\}. \tag{29}$$

Let $\theta \in \mathbb{R}^d$ and $q \in \Delta^{\mathcal{Y}}$, we have the following properties:

1. *$\langle s_\theta | q\rangle = \langle Y^\top\theta | q\rangle = \theta^\top Y q = \langle \theta | Yq\rangle = \theta^\top \mu_q$.*
2. *$\Omega_{\mathcal{C}}^*(\theta) = \Omega_\Delta^*(Y^\top\theta)$, therefore $\Omega_{\mathcal{C}}^*$ has domain $\operatorname{dom}(\Omega_{\mathcal{C}}^*) = \mathbb{R}^d$, it is differentiable over its domain and affine over $V^\perp$.*
3. *$\Omega_\Delta(q) \geq \Omega_{\mathcal{C}}(\mu_q)$ and $\mathcal{L}_{\Omega_\Delta}(s_\theta; q) \geq \mathcal{L}_{\Omega_{\mathcal{C}}}(\theta; \mu_q)$, both with equality if and only if*
$$q = \operatorname*{argmin}_{q' \in \Delta^{\mathcal{Y}} \colon Yq' = \mu_q} \Omega_\Delta(q').$$
4. *$\min_{\theta'} \mathcal{L}_{\Omega_\Delta}(s_{\theta'}; q) \geq \min_{\theta'} \mathcal{L}_{\Omega_{\mathcal{C}}}(\theta'; \mu_q)$, with equality if and only if $q = \operatorname*{argmin}_{q' \in \Delta^{\mathcal{Y}} \colon Yq' = \mu_q} \Omega_\Delta(q')$.*
5. *$\nabla_\theta \mathcal{L}_{\Omega_\Delta}(s_\theta; q) = \nabla_\theta \mathcal{L}_{\Omega_{\mathcal{C}}}(\theta; \mu_q) = Y(\nabla\Omega_\Delta^*(s_\theta) - q) = \nabla\Omega_{\mathcal{C}}^*(\theta) - \mu_q$.*

This notion of regularization on distributions, and the way it induces a regularization on the moment space have already been introduced by Blondel et al. (2020, Section 7.1), but in the case of Ω_Δ being a generalized entropy (Grünwald et al. 2004).

Proof. 1. Immediate.

2. By definition,
$$\Omega_{\mathcal{C}}^*(\theta) = \max_\mu \left(\theta^\top\mu + \max_{q \colon Yq = \mu} -\Omega(q)\right) = \max_{\mu, q \colon \mu = Yq} \underbrace{\theta^\top\mu}_{s_\theta^\top q} - \Omega(q) = \max_q s_\theta^\top q - \Omega(q) = \Omega_\Delta^*(s_\theta).$$

Applying Proposition 2.1 to Ω_Δ, we get that $\operatorname{dom}(\Omega_\Delta) = \mathbb{R}^{|\mathcal{Y}|}$, and it is differentiable over $\mathbb{R}^{|\mathcal{Y}|}$. Composing with a linear map gives the domain and differentiability results. Applying Proposition 2.2 to Ω_Δ^* with $V_\Delta^\perp = \operatorname{span}(\mathbf{1})$, we have Ω_Δ^* affine over $\operatorname{span}(\mathbf{1})$. Besides,
$$s_\theta \in \operatorname{span}(\mathbf{1}) \iff \exists \alpha \in \mathbb{R}, \forall y \in \mathcal{Y}, \langle\theta|y\rangle = \alpha \iff \theta \in V^\perp.$$

3. An immediate consequence of the definition of $\Omega_{\mathcal{C}}$ and the previous points.
4. It follows from properties 1 and 2 that the two minimization problems are equivalent up to a constant.
5. Consequence of the equality of the losses up to a constant that does not depend on θ.

$\square$

3.5 Structured perturbation

We use the notations defined in Section 3.4 for both the variable and distribution spaces. Explicitly defining a proper l.s.c. convex regularization function $\Omega \in \Gamma_0(\mathbb{R}^d)$ with domain $\mathrm{dom}(\Omega) = \mathcal{C}$, and computing the regularized predictions $\hat{y}_\Omega(\theta)$ defined in Equation (15) can be challenging. It may rely on Frank-Wolfe (Frank et al. 1956) algorithm in practice. We follow another approach pioneered by Berthet et al. (2020), defining instead $\Omega^*_{\mathcal{C}}$ and Ω^*_{Δ} directly. More precisely, let $\varepsilon \in \mathbb{R}_{++}$, we introduce:

$$F_{\varepsilon,\mathcal{C}}(\theta) = \mathbb{E}[\max_{y \in \mathcal{Y}} (\theta + \varepsilon \mathbf{Z})^\top y] = \mathbb{E}[\max_{y \in \mathcal{C}} (\theta + \varepsilon \mathbf{Z})^\top y], \tag{30}$$

$$F_{\varepsilon,\Delta}(s) = \mathbb{E}[\max_{y \in \mathcal{Y}} s(y) + \varepsilon \mathbf{Z}^\top y] = \mathbb{E}[\max_{q \in \Delta^{\mathcal{Y}}} (s + \varepsilon Y^\top \mathbf{Z})^\top q], \tag{31}$$

where $\mathbf{Z}$ is a centred random variable on $\mathbb{R}^d$ from an exponential family with positive density, typically a standard multivariate normal distribution. The perturbed linear program in Equation (30) is introduced by Berthet et al. (2020), while Equation (31) is new to the best of our knowledge. We denote by $\Omega_{\varepsilon,\mathcal{C}}$ and $\Omega_{\varepsilon,\Delta}$ their respective Fenchel conjugates. We extend from the work of Berthet et al. (2020, Proposition 2.2) the following properties for $F_{\varepsilon,\mathcal{C}}$ to the case when $\mathcal{C}$ is not full-dimensional.

Proposition 6. *Let $\varepsilon \in \mathbb{R}_{++}$, the function $F_{\varepsilon,\mathcal{C}}$ defined above has the following properties:*

1. *The domain of $F_{\varepsilon,\mathcal{C}}$ is $\mathbb{R}^d$, and $F_{\varepsilon,\mathcal{C}}$ is a proper l.s.c. convex function in $\Gamma_0(\mathbb{R}^d)$.*
2. *$F_{\varepsilon,\mathcal{C}}$ is strictly convex over V, and affine over $V^\perp$. Let $\theta \in \mathbb{R}^d$, such that $\theta = \theta_V + \theta_{V^\perp}$, where $\theta_V = \Pi_V(\theta)$ and $\theta_{V^\perp} = \theta - \theta_V$, and $y_0 \in \mathcal{C}$ be any point in $\mathcal{C}$,*

$$F_{\varepsilon,\mathcal{C}}(\theta) = \langle y_0 | \theta_{V^\perp} \rangle + F_{\varepsilon,\mathcal{C}}(\theta_V).$$

3. *$F_{\varepsilon,\mathcal{C}}$ is twice differentiable, with gradient given by:*

$$\nabla_\theta F_{\varepsilon,\mathcal{C}}(\theta) = \mathbb{E}[\operatorname*{argmax}_{y \in \mathcal{C}} (\theta + \varepsilon \mathbf{Z})^\top y]. \tag{32}$$

4. *The Fenchel conjugate $\Omega_{\varepsilon,\mathcal{C}} := F^*_{\varepsilon,\mathcal{C}}$ has domain $\mathcal{C}$, and its restriction to H is a Legendre-type function.*

Contrary to the full dimension case considered by Berthet et al. (2020), $F_{\varepsilon,\mathcal{C}}$ is not strictly convex over the whole space $\mathbb{R}^d$. Therefore, it is not a Legendre-type function, but its restriction to V is. Besides, its conjugate is not Legendre-type, but the restriction of its conjugate to the affine subspace H is.

Proof. Let $\varepsilon \in \mathbb{R}_{++}$,

1. Let $(\theta, Z) \in (\mathbb{R}^d)^2$, since $\mathcal{C}$ is compact, the maximum in the definition (30) of $F_{\varepsilon,\mathcal{C}}$ is well-defined. The expectation with respect to the Gaussian distribution remains finite, and thus $\mathrm{dom}(F_{\varepsilon,\mathcal{C}}) = \mathbb{R}^d$. Let $Z \in \mathbb{R}^d$, the function $\theta \mapsto \max_{y\in\mathcal{C}}(\theta + \varepsilon Z)^\top y$ is convex since it is the maximum of affine functions in θ. Since the distribution of $\mathbf{Z}$ is non-negative and the expectation linear, $F_{\varepsilon,\mathcal{C}}$ is convex. It is proper since $\mathrm{dom}(F_{\varepsilon,\mathcal{C}}) = \mathbb{R}^d$ and $\mathcal{C}$ is not degenerate. We show in point 3 that $F_{\varepsilon,\mathcal{C}}$ is twice differentiable, it is therefore lower-semicontinuous and $F_{\varepsilon,\mathcal{C}}$ is a proper l.s.c. convex function in $\Gamma_0(\mathbb{R}^d)$.

2. The strict convexity of $F_{\varepsilon,\mathcal{C}}$ over V stems directly from the proof of Proposition 2.2 in the appendix of the study by Berthet et al. (2020). Let now $\theta = \theta_V + \theta_{V^\perp}$ be any vector in $\mathbb{R}^d$, and y_0 be any vector in $\mathcal{C}$,

$$\begin{aligned} F_{\varepsilon,\mathcal{C}}(\theta) = \mathbb{E}[\max_{y\in\mathcal{C}}(\theta_V + \theta_{V^\perp} + \varepsilon\mathbf{Z})^\top y] &= \mathbb{E}[\theta_{V^\perp}^\top y_0 + \max_{y\in\mathcal{C}}(\theta_V + \varepsilon\mathbf{Z})^\top y], \\ &= \theta_{V^\perp}^\top y_0 + F_{\varepsilon,\mathcal{C}}(\theta_V). \end{aligned}$$

   Therefore $F_{\varepsilon,\mathcal{C}}$ is affine over $V^\perp$.

3. As highlighted by Berthet et al. (2020, Proposition 3.1), we can apply the technique of Abernethy et al. (2016, Lemma 1.5) using the smoothness of the distribution of the noise variable $\mathbf{Z}$ to show the smoothness of $F_{\varepsilon,\mathcal{C}}$. It involves a simple change of variable $u = \theta + \varepsilon Z$. The expression of the gradient comes from Danskin's lemma and swap of integration and differentiation.

4. We first show that the domain of $F^*_{\varepsilon,\mathcal{C}}$ is $\mathcal{C}$. Let $y \in \mathbb{R}^d$, by definition

$$F^*_{\varepsilon,\mathcal{C}}(y) = \sup_{\theta\in\mathbb{R}^d}\{\theta^\top y - \mathbb{E}[\max_{y'\in\mathcal{C}}(\theta + \varepsilon\mathbf{Z})^\top y']\}.$$

   - If $y \in \mathbb{R}^d \backslash \mathcal{C}$, we can apply the hyperplane separation theorem given by Rockafellar (1970, Corollary 11.4.2) to $\{y\}$ and $\mathcal{C}$ which are both closed, convex and bounded. There exists a vector $\bar{\theta} \in \mathbb{R}^d$, and a positive real number $\eta \in \mathbb{R}_{++}$ such that,

$$\langle\bar{\theta}|y - y'\rangle \geq \eta, \quad \forall y' \in \mathcal{C}.$$

     Let now $Z \in \mathbb{R}^d$ be any realization of our random vector, and $\lambda \in \mathbb{R}_{++}$ a positive scalar,

$$\begin{aligned} \langle\lambda\bar{\theta} + \varepsilon Z|y - y'\rangle &\geq \lambda\eta + \langle\varepsilon Z|y - y'\rangle, \quad \forall y' \in \mathcal{C}, \\ &\geq \lambda\eta - |\langle\varepsilon Z|y - y'\rangle|, \quad \forall y' \in \mathcal{C}, \\ &\geq \lambda\eta - ||\varepsilon Z||||y - y'||, \quad \forall y' \in \mathcal{C}. \end{aligned}$$

     For the last line we apply Cauchy-Schwarz inequality. Since $\mathcal{C}$ is compact in $\mathbb{R}^d$, we can consider $D_{\mathcal{C},y} := \sup_{y'\in\mathcal{C}} ||y - y'|| < +\infty$.

$$\langle\lambda\bar{\theta} + \varepsilon Z|y - y'\rangle \geq \lambda\eta - ||\varepsilon Z||D_{\mathcal{C},y}, \quad \forall y' \in \mathcal{C}.$$

     Considering the minimum of the left-hand side with respect to $y' \in \mathcal{C}$,

$$\langle\lambda\bar{\theta} + \varepsilon Z|y\rangle - \max_{y'\in\mathcal{C}}\langle\lambda\bar{\theta} + \varepsilon Z|y'\rangle \geq \lambda\eta - ||\varepsilon Z||D_{\mathcal{C},y}.$$

Now, we recall that $\mathbf{Z}$ is centered with distribution ν (typically a multivariate standard normal distribution) with finite variance, therefore

$$N_\nu := \mathbb{E}_{\mathbf{Z}\sim\nu}[||\mathbf{Z}||] < +\infty, \quad \text{and} \quad \mathbb{E}[||\varepsilon\mathbf{Z}||] = |\varepsilon| N_\nu < +\infty.$$

Taking the expectation with respect to ν of the inequality above,

$$\langle\lambda\bar{\theta}|y\rangle - \mathbb{E}[\max_{y'\in\mathcal{C}}\langle\lambda\bar{\theta} + \varepsilon\mathbf{Z}|y'\rangle] \geq \lambda\eta - |\varepsilon| N_\nu D_{\mathcal{C},y}.$$

Therefore, since $F^*_{\varepsilon,\mathcal{C}}(y) \geq \langle\lambda\bar{\theta}|y\rangle - \mathbb{E}[\max_{y'\in\mathcal{C}}\langle\lambda\bar{\theta}+\varepsilon\mathbf{Z}|y'\rangle]$ and $|\varepsilon|N_\nu D_{\mathcal{C},y}$ is finite and does not depend on λ, considering the limit $\lambda \to +\infty$ gives us $F^*_{\varepsilon,\mathcal{C}}(y) = +\infty$.

- If $y \in \mathcal{C}$, since $\mathbf{Z}$ is centered,

$$\begin{aligned} F^*_{\varepsilon,\mathcal{C}}(y) &= \sup_{\theta\in\mathbb{R}^d}\{\theta^\top y - \mathbb{E}[\max_{y'\in\mathcal{C}}(\theta+\varepsilon\mathbf{Z})^\top y']\}, \\ &= \sup_{\theta\in\mathbb{R}^d}\{\mathbb{E}[(\theta+\varepsilon\mathbf{Z})^\top y] - \mathbb{E}[\max_{y'\in\mathcal{C}}(\theta+\varepsilon\mathbf{Z})^\top y']\}, \\ &= \sup_{\theta\in\mathbb{R}^d}\{\mathbb{E}\underbrace{[(\theta+\varepsilon\mathbf{Z})^\top y - \max_{y'\in\mathcal{C}}(\theta+\varepsilon\mathbf{Z})^\top y']}_{\leq 0 \text{ since } y\in\mathcal{C}}\} < +\infty. \end{aligned}$$

Therefore, we have shown that $\operatorname{dom}(F^*_{\varepsilon,\mathcal{C}}) = \mathcal{C}$. The point 2. shows that $(F_{\varepsilon,\mathcal{C}})_{|V}$ is strictly convex over V. Using the computations of point 3., we show that $(F_{\varepsilon,\mathcal{C}})_{|V}$ is differentiable over V. It is thus a Legendre-type function with $\operatorname{dom}\big((F_{\varepsilon,\mathcal{C}})_{|V}\big) = V$. Indeed, point 3 of the essentially smooth definition holds vacuously. To show that $(F^*_{\varepsilon,\mathcal{C}})_{|H}$ is Legendre-type, we use the fact that it is the conjugate (up to a translation) of $(F_{\varepsilon,\mathcal{C}})_{|V}$ in V. Then, Theorem 1 shows that $(F^*_{\varepsilon,\mathcal{C}})_{|H}$ with domain $\mathcal{C}$ is a Legendre-type function (with respect to the metric of H).

This concludes the proof. □

The perturbation in the definition of $F_{\varepsilon,\Delta}$ spans $\operatorname{Im}(Y^\top)$, which is a subspace of dimension $d' \ll |\mathcal{Y}|$. The proofs of Berthet et al. (2020), therefore, do not stand any more. However, perhaps surprisingly, many properties remain valid.

Proposition 7. *Let $\varepsilon \in \mathbb{R}_{++}$, the function $F_{\varepsilon,\Delta}$ defined above has the following properties:*

1. *The domain of $F_{\varepsilon,\Delta}$ is $\mathbb{R}^{|\mathcal{Y}|}$, it is Lipschitz continuous, and it is a proper l.s.c. convex function in $\Gamma_0(\mathbb{R}^{|\mathcal{Y}|})$.*
2. *$F_{\varepsilon,\Delta}$ is strictly convex over V_Δ, and affine over $V_\Delta^\perp = \operatorname{span}(\mathbf{1})$. More precisely, let $s \in \mathbb{R}^{|\mathcal{Y}|}$, decomposed as $s = s_{V_\Delta} + s_{V_\Delta^\perp}$, where $s_{V_\Delta} = \Pi_{V_\Delta}(s)$ and $s_{V_\Delta^\perp} = s - s_{V_\Delta}$, and $q_0 \in \Delta^{\mathcal{Y}}$ be any point in $\Delta^{\mathcal{Y}}$,*
$$F_{\varepsilon,\Delta}(s) = \langle s_{V_\Delta^\perp}|q_0\rangle + F_{\varepsilon,\Delta}(s_{V_\Delta}).$$
3. *$F_{\varepsilon,\Delta}$ is differentiable over $\mathbb{R}^{|\mathcal{Y}|}$, with gradient given by:*
$$\nabla_s F_{\varepsilon,\Delta}(s) = \mathbb{E}[\operatorname*{argmax}_{q\in\Delta^{\mathcal{Y}}}(s + \varepsilon Y^\top\mathbf{Z})^\top q]. \tag{33}$$
4. *The Fenchel conjugate $\Omega_{\varepsilon,\Delta} := F^*_{\varepsilon,\Delta}$ has domain $\Delta^{\mathcal{Y}}$, and its restriction to H_Δ is Legendre-type.*

In the definition of $F_{\varepsilon,\Delta}$, the noise $Y^\top \mathbf{Z}$ follows a degenerate multivariate Gaussian distribution. Therefore, the change of variable in the paper by Abernethy et al. (2016) cannot be applied to show smoothness. Besides, the proof of strict convexity by Berthet et al. (2020) does not hold either. We use alternative approaches in the proof.

Proof. 1. The domain of F_Δ, as well as the fact that it is proper and convex, are proved in the same way as in Proposition 6 point 1. We now show the Lipschitz continuity, and the lower-semicontinuity follows.

Let $(s_1, s_2) \in \mathbb{R}^{|\mathcal{Y}|}$,

$$\begin{aligned}
F_\Delta(s_1) - F_\Delta(s_2) &= \mathbb{E}_{\mathbf{Z}}\big[\max_{q\in\Delta^{\mathcal{Y}}} \langle s_1 + \varepsilon Y^\top \mathbf{Z} | q\rangle - \max_{q\in\Delta^{\mathcal{Y}}} \langle s_2 + \varepsilon Y^\top \mathbf{Z} | q\rangle\big], \\
&\leq \mathbb{E}_{\mathbf{Z}}\big[\max_{q\in\Delta^{\mathcal{Y}}} \langle s_1 - s_2 | q\rangle\big] = \max_{q\in\Delta^{\mathcal{Y}}} \langle s_1 - s_2 | q\rangle, \\
&\leq \max_{q\in\Delta^{\mathcal{Y}}} ||s_1 - s_2|| ||q|| = ||s_1 - s_2||.
\end{aligned}$$

We use, in turn, the fact that the maximum of a sum is smaller or equal to the sum of maxima, the non-negativity of the density of $\mathbf{Z}$, Cauchy-Schwarz inequality, and the definition of the simplex $\Delta^{\mathcal{Y}}$. By symmetry, we have shown that F_Δ is 1-Lipschitz continuous.

2. The proof relies on the following technical lemma proved in Appendix 8.1.

Lemma 2. *Let $\mathcal{Y} \subset \mathbb{R}^d$ be a finite set. We make the assumption that no element of $\mathcal{Y}$ is a strict convex combination of other elements of $\mathcal{Y}$. In other words, $\mathcal{Y}$ is the set of vertices of the polytope $\mathcal{C} = \mathrm{conv}(\mathcal{Y})$. Let $\Delta^{\mathcal{Y}}$ be the probability simplex whose vertices are indexed by $\mathcal{Y}$, and H_Δ its affine hull $H_\Delta = \mathrm{aff}(\Delta^{\mathcal{Y}})$. We denote by V_Δ the vector subspace of $\mathbb{R}^{|\mathcal{Y}|}$ which is the direction of H_Δ, and we have the orthogonal sum $\mathbb{R}^{|\mathcal{Y}|} = V_\Delta \oplus \mathrm{span}(\mathbf{1})$. For any vector $s \in \mathbb{R}^{|\mathcal{Y}|}$, we denote by $s(y) \in \mathbb{R}$ the component of s indexed by $y \in \mathcal{Y}$. We also consider $\varepsilon \in \mathbb{R}_{++}$ a positive real number, and $\mathbf{Z}$, a random variable with standard multivariate Gaussian distribution over $\mathbb{R}^d$.*

Then, for two vectors $(s_1, s_2) \in (V_\Delta)^2$, $s_1 \neq s_2$,

$$\mathbb{P}_{\mathbf{Z}}\left(\operatorname*{argmax}_{y\in\mathcal{Y}} f_1(y;\mathbf{Z}) \cap \operatorname*{argmax}_{y\in\mathcal{Y}} f_2(y;\mathbf{Z}) = \emptyset\right) > 0.$$

Where f_1 and f_2 are defined as:

$$f_1(y; Z) := s_1(y) + \varepsilon Z^\top y, \quad f_2(y; Z) := s_2(y) + \varepsilon Z^\top y.$$

Let $t \in (0, 1)$, the lemma above leads to

$$\mathbb{P}_{\mathbf{Z}}\left[\max_{y\in\mathcal{Y}}\Big(tf_1(y;\mathbf{Z}) + (1-t)f_2(y;\mathbf{Z})\Big) < \max_{y\in\mathcal{Y}} tf_1(y;\mathbf{Z}), \right.$$
$$\left. + \max_{y\in\mathcal{Y}}(1-t)f_2(y;\mathbf{Z})\right] > 0.$$

Since with point 1. F_Δ is convex, this strict inequality and the fact that the distribution of $\mathbf{Z}$ is non-negative leads to the strict convexity of F_Δ over V_Δ. Last, using the decomposition $\mathbb{R}^{|\mathcal{Y}|} = V_\Delta + \mathrm{span}(\mathbf{1})$, we get the affine property over $V_\Delta^\perp = \mathrm{span}(\mathbf{1})$ with the same arguments as for Proposition 6 Point 2.

3. Since we are in the setting given in the preamble of Lemma 2, for $s \in \mathbb{R}^{|\mathcal{Y}|}$, the argmax in the definition of $F_{\varepsilon,\Delta}$ is reduced to a singleton almost surely:

$$P_{\mathbf{Z}}\big[\big|\operatorname*{argmax}_{y\in\mathcal{Y}}\{s(y)+\varepsilon\mathbf{Z}^\top y\}\big|>1\big]=0. \tag{34}$$

Indeed, because the standard multivariate Gaussian distribution has $\mathbb{R}^d$ as support, proving Equation (34) reduces to prove that for any pair of distinct vectors $(y,y')\in\mathcal{Y}^2, y\neq y'$,

$$P_{\mathbf{Z}}\big[s(y)+\varepsilon\mathbf{Z}^\top y=s(y')+\varepsilon\mathbf{Z}^\top y'\big]=0,$$

which is true, otherwise $\langle\cdot|y-y'\rangle$ is constant on a ball of radius $r>0$ in $\mathbb{R}^d$, leading to the contradiction $y=y'$. Remark that the uniqueness of the argmax in $\mathcal{Y}$ implies the uniqueness of the corresponding argmax in the distribution space $\Delta^{\mathcal{Y}}$. Now, using Equation (34), Danskin's lemma (Bertsekas 2009, Proposition A.3.2), and swapping integration with respect to the density of $\mathbf{Z}$ and differentiation with respect to s, we get:

$$\nabla_s\mathbb{E}[\max_{q\in\Delta^{\mathcal{Y}}}(s+\varepsilon Y^\top\mathbf{Z})^\top q]=\mathbb{E}[\operatorname*{argmax}_{q\in\Delta^{\mathcal{Y}}}(s+\varepsilon Y^\top\mathbf{Z})^\top q].$$

4. The domain property is proved in the same way as for Proposition 6 point 4, the rest also yields similarly.

□

Eventually, we show that $\Omega_{\varepsilon,\mathcal{C}}$ is the structured regularization corresponding to $\Omega_{\varepsilon,\Delta}$.

Proposition 8. *$\Omega_{\varepsilon,\mathcal{C}}(\mu)=\min_{q\colon Yq=\mu}\Omega_{\varepsilon,\Delta}(q)$, and hence all the properties of Proposition 5 are true in the sparse perturbation case.*

Proof. By definition, $F_{\varepsilon,\mathcal{C}}(\theta)=F_{\varepsilon,\Delta}(Y^\top\theta)$. Besides, since $\Omega_{\varepsilon,\mathcal{C}}$ and $\Omega_{\varepsilon,\Delta}$ are proper convex lower-semicontinuous, they are bi-conjugate by Fenchel-Moreau theorem. Applying the computations of Bauschke et al. (2017, Corollary 15.28) with $g=F_{\varepsilon,\Delta}$ and $L: x\mapsto Y^\top x$ leads to the result. □

4 Primal-dual algorithm

In this section, we return to the problem of learning structured policies framed in the introduction. We start by defining policies π_w using the distribution space. As mentioned in the introduction, we look for the best policies, i.e., the ones that minimize the empirical risk $R_N(\pi_w)$. Our strategy is the following. In Section 4.1 we precisely define the empirical risk $R_N(\pi_w)$ that we would like to minimize. We do not face this minimization problem directly, since it is not easily tractable in practice. Instead, we define a surrogate function $\mathcal{S}_{\Delta,N}$ in Section 4.2, and an alternating minimization algorithm to solve

$$\min_{\substack{w\in\mathcal{W},\\ q_\otimes\in\Delta_\otimes}}\mathcal{S}_{\Delta,N}\Big(\big(Y(x_i)^\top\varphi_w(x_i)\big)_{i\in[N]},q_\otimes\Big).$$

We show in Section 4.3 that the alternating minimization algorithm leads to tractable approximate updates based on decomposition, sampling and stochastic gradient descent. Then, we show in Section 4.4 that in a restricted setting, the primal updates $q_\otimes^{(t)}$ coincide with the ones of mirror descent applied to the partial minimum $\min_{s_\otimes}\mathcal{S}_{\Delta,N}\big(s_\otimes,q_\otimes\big)$. In Section 4.5 we study the convergence of the primal-dual algorithm. We show in Theorem 7 that, provided some regularity assumption on the regularizer, the algorithm converges with a linear rate in a restricted setting.

Last, in Section 4.6, we show that in a restricted setting and provided some technical assumptions, the miminizers of the surrogate

$$\theta_{\mathcal{S}} \in \operatorname*{argmin}_{\theta} \min_{q_\otimes \in \Delta_\otimes} \mathcal{S}_{\Delta,N}\Big(\big(Y^\top \theta\big)_{i\in[N]}, q_\otimes\Big)$$

have a bounded non-optimality with respect to the empirical risk.

4.1 Empirical risk based on regularized policies

We come back to the setting described in Section 1, and denote by (x,ξ) a context-noise pair observation. Recall that we have access to a dataset $(x_1,\xi_1),\dots(x_N,\xi_N)$ of such pairs. We now have all the tools to build the conditional probability distributions $p_{\Omega_{\mathcal{C}(x)}}(\cdot|x)$ mentioned in the introduction. Indeed, since optimizing over a finite set is equivalent to optimizing over distributions on this set, the linear optimization problem

$$\max_{y\in\mathcal{Y}(x)} \langle\theta|y\rangle$$

is equivalent to the moment problem

$$\max_{q\in\Delta^{\mathcal{Y}(x)}} \langle Y(x)^\top\theta|q\rangle, \tag{35}$$

where $Y(x)=(y)_{y\in\mathcal{Y}(x)}$ is the wide matrix of points in $\mathcal{Y}(x)$, and $\Delta^{\mathcal{Y}(x)}$ is the probability simplex over $\mathcal{Y}(x)$. Any element q of $\Delta^{\mathcal{Y}(x)}$ is therefore a probability distribution over $\mathcal{Y}(x)$. However, with the non-regularized problem (35), there is always a basic optimal solution that is a Dirac over one of the points in $\mathcal{Y}(x)$. We therefore introduce a regularization function $\Omega_{\Delta^{\mathcal{Y}(x)}}$ that is proper convex l.s.c. with domain $\operatorname{dom}(\Omega_{\Delta^{\mathcal{Y}(x)}})=\Delta^{\mathcal{Y}(x)}$, and such that its restriction to the affine hull of $\Delta^{\mathcal{Y}(x)}$ is Legendre type. In practice, we use below the perturbed maxima and their conjugates defined in Section 3.5. Based on this regularization function, we define the conditional distribution $p_{\Omega_{\Delta(x)}}$ as the optimum of the regularized problem

$$\big(p_{\Omega_{\Delta(x)}}(y|\theta)\big)_{y\in\mathcal{Y}(x)} = \operatorname*{argmax}_{q\in\Delta^{\mathcal{Y}(x)}} \langle Y(x)^\top\theta|q\rangle - \Omega_{\Delta^{\mathcal{Y}(x)}}(q) = \nabla\Omega^*_{\Delta^{\mathcal{Y}(x)}}\big(Y(x)^\top\theta\big).$$

The uniqueness of the argmax and the second equality come from Proposition 1. As stated in the introduction, we consider policies of the form

$$\pi_w(\cdot|x) = p_{\Omega_{\Delta(x)}}\big(\,\cdot\,|\varphi_w(x)\big),$$

where φ_w is a machine learning predictor that takes the context observation as input. Using the distribution $p_{\Omega_{\Delta(x)}}$, we can define the *expected cost* given θ as

$$\mathbb{E}\Big[c(x,\mathbf{y},\xi)\Big|\theta\Big] = \langle\varsigma(x,\xi)|\nabla\Omega^*_{\Delta^{\mathcal{Y}(x)}}\big(Y(x)^\top\theta\big)\rangle \quad \text{where} \quad \varsigma(x,\xi)=\big(c(x,y,\xi)\big)_{y\in\mathcal{Y}(x)}.$$

We now introduce the product spaces

$$S_\otimes = \{(s_i)_{i\in[N]}\,|\,\forall i\in[N], s_i\in\mathbb{R}^{|\mathcal{Y}(x_i)|}\} \quad \text{and} \quad \Delta_\otimes = \{(q_i)_{i\in[N]}\,|\,\forall i\in[N], q_i\in\Delta^{\mathcal{Y}(x_i)}\}.$$

Let $q\in\Delta^{\mathcal{Y}(x)}$ be a distribution over $\mathcal{Y}(x)$, and $s_\otimes\in S_\otimes$ be a product of score vectors in large dimension, we define

$$R_\Delta(q;x,\xi) = \langle\varsigma(x,\xi)|q\rangle \quad \text{where} \quad \varsigma(x,\xi)=\big(c(x,y,\xi)\big)_{y\in\mathcal{Y}(x)}, \tag{36a}$$

$$\mathcal{R}_{\Delta,N}(s_\otimes) = \frac{1}{N}\sum_{i=1}^N R_\Delta(\nabla\Omega^*_{\Delta^{\mathcal{Y}(x_i)}}(s_i);x_i,\xi_i) \quad \text{where} \quad s_\otimes=(s_i)_{i\in[N]}. \tag{36b}$$

Since we focus on a single cost function c below, when considering product variables, we also shorten the notation of $\varsigma(x_i, \xi_i)$ as ς_i. The empirical risk minimization problem (4) becomes

$$\min_{\pi_w \in \mathcal{H}_{\mathcal{W}}} R_N(\pi_w) = \min_{w \in \mathcal{W}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\mathbf{y} \sim \pi_w(\cdot|x_i)} \Big[c(x_i, \mathbf{y}, \xi_i) \Big| \varphi_w(x_i) \Big], \tag{37a}$$

$$= \min_{w \in \mathcal{W}} \mathcal{R}_{\Delta,N} \Big(\big(Y(x_i)^\top \varphi_w(x_i) \big)_{i \in [N]} \Big), \tag{37b}$$

$$= \min_{w \in \mathcal{W}} \frac{1}{N} \sum_{i=1}^{N} \langle \varsigma_i | \nabla \Omega^*_{\Delta^{\mathcal{Y}(x_i)}} \big(\underbrace{Y(x_i)^\top \varphi_w(x_i)}_{s_{\theta_i}} \big) \rangle. \tag{37c}$$

We now introduce a damped primal-dual algorithm to minimize a surrogate of the empirical regret in Equation (37).

4.2 Damped primal-dual algorithm for a surrogate learning problem

We recall that the Fenchel-Young loss generated by $\Omega_{\Delta^{\mathcal{Y}(x)}}$ is defined over $\mathbb{R}^{|\mathcal{Y}(x)|} \times \Delta^{\mathcal{Y}(x)}$, and written as $\mathcal{L}_{\Omega_{\Delta(x)}}(s; q) = \Omega_{\Delta^{\mathcal{Y}(x)}}(q) + \Omega^*_{\Delta^{\mathcal{Y}(x)}}(s) - \langle s|q \rangle$. Let $\kappa > 0$ be a positive constant, we introduce the surrogate functions for Equation (36) as

$$S_{\kappa,\Delta}(s, q; x, \xi) = \langle \varsigma(x,\xi) | q \rangle + \kappa \mathcal{L}_{\Omega_{\Delta(x)}}(s; q), \quad \text{where} \quad \varsigma(x,\xi) = \big(c(x,y,\xi) \big)_{y \in \mathcal{Y}(x)}, \tag{38a}$$

$$\mathcal{S}_{\Delta,N}(s_\otimes, q_\otimes) = \frac{1}{N} \sum_{i=1}^{N} S_{\kappa,\Delta}(s_i, q_i; x_i, \xi_i), \qquad \text{where} \quad s_\otimes = (s_i)_{i \in [N]},\ q_\otimes = (q_i)_{i \in [N]}. \tag{38b}$$

Notice that if $S_{\kappa,\Delta}(\cdot,\cdot; x, \xi)$ defined in Equation (38a) takes as inputs $s \in \mathbb{R}^{|\mathcal{Y}(x)|}$ and its dual variable $q = \nabla \Omega^*_{\Delta^{\mathcal{Y}(x)}}(s)$, by Proposition 1 the Fenchel-Young loss reaches zero, and we recover $R_\Delta(q; x, \xi)$ defined in Equation (36a). In general, we have the Fenchel-Young inequality.

$$S_{\kappa,\Delta}(s, \nabla \Omega^*_{\Delta^{\mathcal{Y}(x)}}(s); x, \xi) = R_\Delta(\nabla \Omega^*_{\Delta^{\mathcal{Y}(x)}}(s); x, \xi), \tag{39a}$$

$$S_{\kappa,\Delta}(s, q; x, \xi) = \underbrace{\langle \varsigma(x,\xi) | q \rangle}_{R_\Delta(q;x,\xi)} + \kappa \underbrace{\mathcal{L}_{\Omega_{\Delta(x)}}(s; q)}_{\geq 0}. \tag{39b}$$

This is why we use the term surrogate functions for Equation (38). Our *surrogate learning problem* can then be written as

$$\min_{\substack{w \in \mathcal{W},\\ q_\otimes \in \Delta_\otimes}} \mathcal{S}_{\Delta,N} \Big(\big(Y(x_i)^\top \varphi_w(x_i) \big)_{i \in [N]}, q_\otimes \Big) = \min_{\substack{w \in \mathcal{W},\\ q_\otimes \in \Delta_\otimes}} \frac{1}{N} \sum_{i=1}^{N} S_{\kappa,\Delta} \big(Y(x_i)^\top \varphi_w(x_i), q_i; x_i, \xi_i \big). \tag{40}$$

In order to derive solutions to Equation (40), we rely on the following *primal-dual alternating minimization scheme*, initialized by some weights values $\bar{w}^{(0)}$

$$q_\otimes^{(t+1)} = \operatorname*{argmin}_{q_\otimes \in \Delta_\otimes} \mathcal{S}_{\Delta,N} \Big(\big(Y(x_i)^\top \varphi_{\bar{w}^{(t)}}(x_i) \big)_{i \in [N]}, q_\otimes \Big), \qquad \text{(decomposition)}, \tag{41a}$$

$$\bar{w}^{(t+\frac{1}{2})} \in \operatorname*{argmin}_{w \in \mathcal{W}} \mathcal{S}_{\Delta,N} \Big(\big(Y(x_i)^\top \varphi_w(x_i) \big)_{i \in [N]}, q_\otimes^{(t+1)} \Big), \qquad \text{(coordination)}, \tag{41b}$$

$$\bar{w}^{(t+1)} = \alpha \bar{w}^{(t+\frac{1}{2})} + (1-\alpha) \bar{w}^{(t)}, \qquad \text{(damping)}, \tag{41c}$$

where $\alpha \in (0,1)$ is a damping parameter. The primal update in Equation (41a) is also named decomposition step, since the minimization can be done per term i and variable q_i as we highlight below. The dual update in Equation (41b) is named coordination, because this time a single vector of weights is used, coupling the N terms of the sum in $\mathcal{S}_{\Delta,N}$, and coordinating the primal updates. As we discuss below, it corresponds to a simple supervised learning problem with a Fenchel-Young loss. The last update in Equation (41c) is a mere convex combination, also called damping step.

In the sequel, we further study the tractability of the primal-dual algorithm (41) in Section 4.3. We then highlight some links with mirror descent algorithm in Section 4.4, paving the way for convergence results in a restricted setting and provided some regularity assumptions. Last, we study the performance of the resulting solution to the surrogate problem with respect to the initial empirical risk in Section 4.6.

4.3 Tractable approximate primal-dual updates

We come back to the primal-dual algorithm given in Equation (41), and exploit the structure of the surrogate function $\mathcal{S}_{\Delta,N}$ to study the tractability of each step.

Proposition 9. *Using the notations of Section 4.2, Equations (41a)-(41b) can be recast as*

$$q_i^{(t+1)} = \underset{q_i \in \Delta^{\mathcal{Y}(x_i)}}{\operatorname{argmin}} \; S_{\kappa,\Delta}\big(Y(x_i)^\top \varphi_{\bar{w}^{(t)}}(x_i), q_i; x_i, \xi_i\big), \tag{42a}$$

$$= \nabla\Omega^*_{\Delta^{\mathcal{Y}(x_i)}}\big(Y(x_i)^\top \varphi_{\bar{w}^{(t)}}(x_i) - \frac{1}{\kappa}\varsigma_i\big), \tag{42b}$$

$$\bar{w}^{(t+\frac{1}{2})} \in \underset{w \in \mathcal{W}}{\operatorname{argmin}} \; \frac{1}{N}\sum_{i=1}^N \mathcal{L}_{\Omega_{\mathcal{C}(x_i)}}\big(\varphi_w(x_i); Y(x_i)q_i^{(t+1)}\big), \tag{42c}$$

where $\mathcal{L}_{\Omega_{\mathcal{C}(x)}}$ is the Fenchel-Young loss on the moment polytope $\mathcal{C}(x) = \operatorname{conv}(\mathcal{Y}(x))$ generated by

$$\Omega_{\mathcal{C}(x)}(\mu) := \min_{q \in \Delta^{\mathcal{Y}(x)}} \{\Omega_{\Delta^{\mathcal{Y}(x)}}(q) \,|\, Y(x)q = \mu\}. \tag{43}$$

The dual update in Equation (42c) corresponds to a supervised learning problem with a Fenchel-Young loss, using as learning dataset $\big(x_i, Y(x_i)q_i^{(t+1)}\big)_{i\in[N]}$, where each target point $Y(x)q_i^{(t+1)}$ belongs to the corresponding convex compact set $\mathcal{C}(x_i)$. This update can thus be computed in the tractable dimension of the moment polytope.

Proof. To derive Equation (42b) remark the following. First, $\mathcal{S}_{\Delta,N}$ is defined as the sum of $S_{\kappa,\Delta}$, and since $\bar{w}^{(t)}$ is fixed, we obtain N independent problems. Now, using the expression of $S_{\kappa,\Delta}$, we see that $q_i^{(t+1)}$ belongs to the minimizers of

$$\Omega_{\Delta^{\mathcal{Y}(x_i)}}(\cdot) - \langle Y(x_i)^\top \varphi_{\bar{w}^{(t)}}(x_i) - \frac{1}{\kappa}\varsigma_i | \cdot \rangle.$$

Using Proposition 1 Point 4, we recognize $\nabla\Omega^*_{\Delta^{\mathcal{Y}(x_i)}}\big(Y(x_i)^\top \varphi_{\bar{w}^{(t)}}(x_i) - \frac{1}{\kappa}\varsigma_i\big)$.

For the dual update in Equation (42c), using the expression of $S_{\kappa,\Delta}$, and omitting the term that does not depend on w, we can first recast Equation (41b) as

$$\bar{w}^{(t+\frac{1}{2})} \in \underset{w \in \mathcal{W}}{\operatorname{argmin}} \; \frac{1}{N}\sum_{i=1}^N \mathcal{L}_{\Omega_{\Delta(x_i)}}\big(Y(x_i)^\top \varphi_w(x_i); q_i^{(t+1)}\big).$$

Now, we leverage the computations of Section 3.4. In particular, based on the regularization function $\Omega_{\Delta^{\mathcal{Y}(x)}}$ on distributions in $\Delta^{\mathcal{Y}(x)}$, we define a regularization function $\Omega_{\mathcal{C}(x)}$ on the moment space $\mathcal{C}(x) = \operatorname{conv}(\mathcal{Y}(x))$ as in Equation (43). We then use Proposition 5, more precisely Points 1-2, to get Equation (42c). $\square$

At this point, we can base ourselves on all the literature on Fenchel-Young losses and the regularization functions considered in Section 3 for the dual update corresponding to a supervised learning problem. Nonetheless, the way we can evaluate the regularization function in small dimension $\Omega_{\mathcal{C}(x)}$ is still unclear. Besides, the primal update written as Equation (42b) requires working with distributions. Ideally, we would like to do all the computations in the space $\mathcal{Y}(x)$ of small-dimension. This is where the perturbation setting of Section 3.5 becomes useful.

Proposition 10. *Let $\varepsilon > 0$ be a positive constant. In the case the regularization functions $\Omega_{\mathcal{C}(x)}$ and $\Omega_{\Delta^{\mathcal{Y}(x)}}$ are defined as the Fenchel conjugates of the perturbed maxima $\Omega_{\mathcal{C}(x)} := F^*_{\varepsilon,\mathcal{C}(x)}$ and $\Omega_{\Delta^{\mathcal{Y}(x)}} := F^*_{\varepsilon,\Delta(x)}$ introduced in Equations (30)-(31), the primal-dual updates in Equations (42b)-(42c) become*

$$\mu_i^{(t+1)} = \mathbb{E}_{\mathbf{Z}}\big[\operatorname*{argmin}_{y_i \in \mathcal{Y}(x_i)} c(x_i, y_i, \xi_i) - (\varphi_{\bar{w}^{(t)}}(x_i) + \varepsilon \mathbf{Z})^\top y_i\big], \tag{44a}$$

$$\bar{w}^{(t+\frac{1}{2})} \in \operatorname*{argmin}_{w} \frac{1}{N} \sum_{i=1}^{N} F_{\varepsilon,\mathcal{C}(x_i)}\big(\varphi_w(x_i)\big) - \langle \varphi_w(x_i) | \mu_i^{(t+1)} \rangle, \tag{44b}$$

where, given the primal update in Equation (42b), the moment primal variables $\mu_i^{(t+1)}$ correspond to $\mu_i^{(t+1)} = Y(x_i) q_i^{(t+1)}$, and can be seen as expectations under the distributions $q_i^{(t+1)}$. They belong to the convex compact sets $\mathcal{C}(x_i)$.

As mentioned above, the dual update in Equation (44b) corresponds to a supervised learning problem on a dataset made of $\big(x_i, \mu_i^{(t+1)}\big)_{i \in [N]}$. In practice, it is approximately solved using stochastic gradient descent, with Monte-Carlo estimates and sampling $\mathbf{Z}$. The details are derived by Berthet et al. (2020). The primal update in Equation (44a) has now a convenient expression. Indeed, we recall that in Section 1, we assume that we have access to an efficient algorithm to solve Equation (3). We can therefore leverage the latter to derive Monte-Carlo estimates of the μ_i, again by sampling the random variable $\mathbf{Z}$.

Proof. For the primal update in Equation (44a), we can reformulate Equation (42b) in this perturbation setting

$$\begin{aligned}
\mu_i^{(t+1)} &= Y(x_i) q_i^{(t+1)} = Y(x_i) \nabla F_{\varepsilon,\Delta(x_i)}\big(Y(x_i)^\top \varphi_{\bar{w}^{(t)}}(x_i) - \frac{1}{\kappa}\varsigma_i\big), \\
&= Y(x_i) \mathbb{E}_{\mathbf{Z}}\big[\operatorname*{argmax}_{q_i \in \Delta^{\mathcal{Y}(x_i)}} \langle Y(x_i)^\top \varphi_{\bar{w}^{(t)}}(x_i) - \frac{1}{\kappa}\varsigma_i + \varepsilon Y(x_i)^\top \mathbf{Z} | q_i \rangle\big], \\
&= \mathbb{E}_{\mathbf{Z}}\big[Y(x_i) \operatorname*{argmin}_{q_i \in \Delta^{\mathcal{Y}(x_i)}} \langle \big(\frac{1}{\kappa} c(x_i, y, \xi_i) - (\varphi_{\bar{w}^{(t)}}(x_i) + \varepsilon \mathbf{Z})^\top y\big)_{y \in \mathcal{Y}(x_i)} | q_i \rangle\big], \\
&= \mathbb{E}_{\mathbf{Z}}\big[\operatorname*{argmin}_{y_i \in \mathcal{Y}(x_i)} c(x_i, y_i, \xi_i) - \kappa(\varphi_{\bar{w}^{(t)}}(x_i) + \varepsilon \mathbf{Z})^\top y_i\big].
\end{aligned}$$

In the computations above, we use Proposition 7 for the expression of the gradient of the function $F_{\varepsilon,\Delta(x_i)}$, and the fact that the minimum of the linear optimization problem in q_i is attained (almost surely) at a vertex of the simplex $\Delta^{\mathcal{Y}(x_i)}$, corresponding to a Dirac on a point $y_i \in \mathcal{Y}(x_i)$. We see that the two constants κ and ε play similar roles in this setting. Up to a re-normalization of the ML predictor φ_w, we keep the hyper-parameter ε to tune the regularization scale.

For the dual update in Equation (44b), we simply use the expression of the Fenchel-Young loss in the perturbation setting, and omit the terms that do not depend on w. □

Instead of the distribution variables q_i, our primal-dual algorithm in Equation (44) now only relies on the moment variables μ_i. Although these variables belong to the convex hulls of the combinatorial sets $\mathcal{C}(x_i) = \mathrm{conv}(\mathcal{Y}(x_i))$, the updates in Equation (44) only require to call the cost function c at points that belong to the combinatorial sets $\mathcal{Y}(x_i)$. This is particularly useful when this cost function cannot be directly applied to points in the convex hulls, which happens quite often in operations research.

4.4 Relationship to mirror descent algorithm

In this section, we place ourselves in the setting of (non-contextual) stochastic optimization. Therefore, we can drop x from the notations for the sets $\mathcal{Y}$, $\mathcal{C}$, $\Delta^{\mathcal{Y}}$, $\mathbb{R}^{|\mathcal{Y}|}$, and regularization functions Ω_Δ and $\Omega_{\mathcal{C}}$. We are going to make a link between a special case of our primal-dual updates in Equation (41) and mirror descent. For that purpose, we start with some definitions and lemmas. We introduce the following regularization function $\Omega_{\Delta_\otimes}$, that takes as input a product of distributions $q_\otimes \in \Delta_\otimes$

$$\Omega_{\Delta_\otimes}(q_\otimes) := \sum_{i=1}^{N} \Omega_\Delta(q_i). \tag{46}$$

Using Proposition 3, we define a $\Delta^{\mathcal{Y}}$-compatible mirror map Ψ, such that $\Omega_\Delta = \Psi + \mathbb{I}_{\Delta^{\mathcal{Y}}}$. Similarly, for $q_\otimes \in \Delta_\otimes$ in the product distribution space, we define

$$\Psi_\otimes(q_\otimes) := \sum_{i=1}^{N} \Psi(q_i). \tag{47}$$

The following lemma is useful to derive mirror descent below.

Lemma 3. *The function $\Psi_\otimes$ defined in Equation (47) is a $\Delta_\otimes$-compatible mirror map. Besides, its conjugate $\Psi_\otimes^*$, defined over $(\mathbb{R}^{|\mathcal{Y}|})^N$ is such that, for $s_\otimes = (s_i)_{i\in[N]} \in (\mathbb{R}^{|\mathcal{Y}|})^N$,*

$$\Psi_\otimes^*(s_\otimes) = \sum_{i=1}^{N} \Psi^*(s_i).$$

Proof. Extending the properties of the $\Delta^{\mathcal{Y}}$-compatible mirror map Ψ to a $\Delta_\otimes$-compatible mirror map $\Psi_\otimes$ comes from the fact that $\Psi_\otimes$ is a sum of functions of independent variables. In particular, for $q_\otimes \in \mathrm{rel\,int}(\Delta_\otimes)$, the gradient of $\Psi_\otimes$ is the product of gradients of Ψ

$$\nabla\Psi_\otimes(q_\otimes) = \big(\nabla\Psi(q_i)\big)_{i\in[N]}.$$

For the conjugate, let $s_\otimes \in (\mathbb{R}^{|\mathcal{Y}|})^N$. By definition we have

$$\begin{aligned}
\Psi_\otimes^*(s_\otimes) &= \sup_{q_\otimes \in \mathrm{dom}(\Psi_\otimes)} \{\langle s_\otimes | q_\otimes\rangle - \Psi_\otimes(q_\otimes)\}, &\text{(48a)}\\
&= \sum_{i=1}^{N} \sup_{q_i \in \mathrm{dom}(\Psi)} \Big\{\langle s_i | q_i\rangle - \Psi(q_i)\Big\}, &\text{(48b)}\\
&= \sum_{i=1}^{N} \Psi^*(s_i). &\text{(48c)}
\end{aligned}$$

□

We now introduce the following function, defined on the same product space $\Delta_\otimes$, and based on a partial minimization of our surrogate function $\mathcal{S}_{\Delta,N}$ introduced in Equation (38b)

$$\bar{\mathcal{S}}_{\Delta,N}(q_\otimes) := \min_{s_\otimes \in \bar{S}_\otimes} \mathcal{S}_{\Delta,N}(s_\otimes, q_\otimes), \text{ where } \bar{S}_\otimes = \{(s_i)_{i\in[N]} \in (\mathbb{R}^{|\mathcal{Y}|})^N \mid s_1 = \ldots = s_N\}. \tag{49}$$

Lemma 4. *Let $q_\otimes = (q_i)_{i\in[N]} \in \Delta_\otimes$, the function $\bar{\mathcal{S}}_{\Delta,N}$ defined above is equal to*

$$\bar{\mathcal{S}}_{\Delta,N}(q_\otimes) = \frac{1}{N}\sum_{i=1}^N \langle \varsigma_i | q_i \rangle + \frac{\kappa}{N}\Big[\sum_{i=1}^N \Omega_\Delta(q_i) - N\Omega_\Delta(\frac{1}{N}\sum_{i=1}^N q_i)\Big].$$

Besides, $\bar{\mathcal{S}}_{\Delta,N}$ coincides over $\Delta_\otimes$ with

$$\hat{\mathcal{S}}_{\Delta,N}(q_\otimes) := \frac{1}{N}\sum_{i=1}^N \langle \varsigma_i | q_i \rangle + \frac{\kappa}{N}\Big[\sum_{i=1}^N \Psi(q_i) - N\Psi(\frac{1}{N}\sum_{i=1}^N q_i)\Big]. \tag{50}$$

Note that it is convex if the function $q_\otimes \mapsto \frac{1}{N}\sum_{i=1}^N \Psi(q_i) - \Psi(\frac{1}{N}\sum_{i=1}^N q_i)$ is convex. This latter corresponds to the Jensen gap of Ψ. We postpone the convexity study to Section 5.

Proof. Let $q_\otimes \in \Delta_\otimes$, and $\bar{S}_\otimes = \{(s_i)_{i\in[N]} \in (\mathbb{R}^{|\mathcal{Y}|})^N \mid s_1 = \ldots = s_N\}$,

$$\begin{aligned}
\bar{\mathcal{S}}_{\Delta,N}(q_\otimes) &= \min_{s_\otimes \in \bar{S}_\otimes} \mathcal{S}_{\Delta,N}(s_\otimes, q_\otimes), && (51a)\\
&= \min_{s\in\mathbb{R}^{|\mathcal{Y}|}} \frac{1}{N}\sum_{i=1}^N \langle \varsigma_i | q_i \rangle + \kappa\big(\Omega_\Delta(q_i) + \Omega_\Delta^*(s) - \langle s | q_i \rangle\big), && (51b)\\
&= \frac{1}{N}\sum_{i=1}^N \langle \varsigma_i | q_i \rangle + \frac{\kappa}{N}\Big[\sum_{i=1}^N \Omega_\Delta(q_i) - N\Omega_\Delta(\frac{1}{N}\sum_{i=1}^N q_i)\Big]. && (51c)
\end{aligned}$$

In the computations above, we use the definition (38) of the function $\mathcal{S}_{\Delta,N}$, and the fact that Ω_Δ is the Fenchel conjugate of Ω_Δ^* to derive the last line. □

Now, the following proposition makes a link between a special case of our primal-dual algorithm and mirror descent.

Theorem 5. *Let $\alpha \in (0,1)$ and $\eta = N\frac{\alpha}{\kappa}$. Then, the following special case of algorithm (41)*

$$\begin{aligned}
q_\otimes^{(t+1)} &= \operatorname*{argmin}_{q_\otimes \in \Delta_\otimes} \mathcal{S}_{\Delta,N}\Big(\bar{s}_\otimes^{(t)}, q_\otimes\Big), && \textit{(decomposition)}, && (52a)\\
\bar{s}_\otimes^{(t+\frac{1}{2})} &\in \operatorname*{argmin}_{s_\otimes \in \bar{S}_\otimes} \mathcal{S}_{\Delta,N}\Big(s_\otimes, q_\otimes^{(t+1)}\Big), && \textit{(coordination)}, && (52b)\\
\bar{s}_\otimes^{(t+1)} &= \alpha \bar{s}_\otimes^{(t+\frac{1}{2})} + (1-\alpha)\bar{s}_\otimes^{(t)}, && \textit{(damping)}, && (52c)
\end{aligned}$$

has the same primal iterates $q_\otimes^{(t)}$ as a mirror descent algorithm applied to the function $\hat{\mathcal{S}}_{\Delta,N}$, using the $\Delta_\otimes$-compatible mirror map $\Psi_\otimes$ and step size η

$$\begin{aligned}
s_\otimes^{(t)} &= \nabla\Psi_\otimes(q_\otimes^{(t)}), && (53a)\\
q_\otimes^{(t+1)} &\in \operatorname*{argmin}_{q_\otimes \in \operatorname{rel\,int}(\Delta_\otimes)} B_{\Psi_\otimes}\big(q_\otimes \| \nabla\Psi_\otimes^*(s_\otimes^{(t)} - \eta g^{(t)})\big), \quad \textit{where} \quad g^{(t)} = \nabla\hat{\mathcal{S}}_{\Delta,N}(q_\otimes^{(t)}), && (53b)
\end{aligned}$$

and where $B_{\Psi_\otimes}$ is the Bregman divergence generated by $\Psi_\otimes$. Therefore, provided that $\Psi_\otimes$ is ρ-strongly convex over $\operatorname{rel\,int}(\Delta_\otimes)$ *with respect to the norm $||\cdot||$, and that the function $\hat{\mathcal{S}}_{\Delta,N}$ is convex and L-Lipschitz (has bounded gradients) with respect to the same norm $||\cdot||$, the average primal iterates $q_\otimes^{(t)}$ in Equation (52) converge in value to the minimum of $\hat{\mathcal{S}}_{\Delta,N}$ over $\Delta_\otimes$, which is also the minimum of the function $\bar{\mathcal{S}}_{\Delta,N}$ by definition.*

As we highlight above in Section 4.3, the main reasons why we can derive tractable primal-dual updates in practice are that

- The primal updates decompose per term i and variable q_i. They may be approximated by solving multiple times the deterministic problem (3).
- Compared with Equation (52b), the dual updates in Equation (41b) rely on a dimension reduction based on a machine learning predictor $Y^\top \varphi_w$, parameterizing the score vector. They can be approximated using stochastic gradient descent.

This theorem shows that, provided some regularity assumptions (that we have not shown up to now), the decomposition per term i in the first point actually does not prevent the average primal iterates from converging in value.

Proof. The proof is in three steps. First, we recast the mirror descent updates given in Equation (53). Second, we explicit the primal-dual alternating minimization steps in Equation (52). Last, we show by induction that the primal updates of the two algorithms coincide.

Mirror descent updates. The gradient of $\hat{\mathcal{S}}_{\Delta,N}$ at $q_\otimes \in \operatorname{rel\,int}(\Delta_\otimes)$ is

$$\nabla \hat{\mathcal{S}}_{\Delta,N}(q_\otimes) = \left(\frac{1}{N}\Big[\varsigma_i + \kappa\big[\nabla\Psi(q_i) - \nabla\Psi\big(\frac{1}{N}\sum_{j=1}^N q_j\big)\big]\Big]\right)_{i\in[N]}. \tag{54}$$

Let's now write the mirror descent updates for the function $\hat{\mathcal{S}}_{\Delta,N}$ with $\Delta_\otimes$-compatible mirror map $\Psi_\otimes$, and step size η

$$\begin{aligned} s_\otimes^{(t)} &= \nabla\Psi_\otimes(q_\otimes^{(t)}), \\ q_\otimes^{(t+1)} &\in \underset{q_\otimes \in \operatorname{rel\,int}(\Delta_\otimes)}{\operatorname{argmin}} \; B_{\Psi_\otimes}\big(q_\otimes || \nabla\Psi_\otimes^*(s_\otimes^{(t)} - \eta g^{(t)})\big), \quad \text{where} \quad g^{(t)} = \nabla\hat{\mathcal{S}}_{\Delta,N}(q_\otimes^{(t)}). \end{aligned}$$

Using the definition of $\Psi_\otimes$ in Equation (47),

$$s_\otimes^{(t)} = \Big(\nabla\Psi(q_i^{(t)})\Big)_{i\in[N]}.$$

Using in addition the expression of the gradient of $\hat{\mathcal{S}}_{\Delta,N}$ given in Equation (54),

$$s_\otimes^{(t)} - \eta g^{(t)} = \left(\nabla\Psi(q_i^{(t)}) - \frac{\eta}{N}\Big(\varsigma_i + \kappa\big[\nabla\Psi(q_i^{(t)}) - \nabla\Psi\big(\frac{1}{N}\sum_{j=1}^N q_j^{(t)}\big)\big]\Big)\right)_{i\in[N]}.$$

Recall that we use $\eta = N\frac{\alpha}{\kappa}$. It leads to

$$s_\otimes^{(t)} - \eta g^{(t)} = \left((1-\alpha)\nabla\Psi(q_i^{(t)}) + \alpha\Big(-\frac{1}{\kappa}\varsigma_i + \nabla\Psi\big(\frac{1}{N}\sum_{j=1}^N q_j^{(t)}\big)\Big)\right)_{i\in[N]}. \tag{55}$$

We see with Equation (55) above that we can consider each component $i \in [N]$ of the product variable separately. Using Lemma 3, this property remains true when applying both the gradient of $\Psi^*_\otimes$ and the Bregman divergence generated by $\Psi_\otimes$. Remark now that since $\Omega_\Delta = \Psi + \mathbb{I}_{\Delta^\mathcal{Y}}$, we have for $i \in [N]$, by Proposition 4 point 1

$$\begin{aligned} q_i^{(t+1)} &\in \operatorname*{argmin}_{q_i \in \operatorname{rel\,int}(\Delta^\mathcal{Y})} B_\Psi\big(q_i || \nabla \Psi^*(s_i^{(t)} - \eta g_i^{(t)})\big) = \{\nabla \Omega^*_\Delta\big(s_i^{(t)} - \eta g_i^{(t)}\big)\}, \\ &= \nabla \Omega^*_\Delta \Big((1-\alpha)\nabla\Psi(q_i^{(t)}) + \alpha\Big(-\frac{1}{\kappa}\varsigma_i + \nabla\Psi\big(\frac{1}{N}\sum_{j=1}^N q_j^{(t)}\big)\Big)\Big). \end{aligned}$$

We now define the following iterates, for an additional notation

$$\bar{s}^{(0)} = 0, \quad \bar{s}^{(t+1)} = \alpha \nabla \Psi\big(\frac{1}{N}\sum_{i=1}^N q_i^{(t+1)}\big) + (1-\alpha)\bar{s}^{(t)}.$$

Finally, we recast mirror descent updates (with the additional variable $\bar{s}^{(t)}$) as

$$q_i^{(t+1)} = \nabla \Omega^*_\Delta \Big((1-\alpha)\nabla\Psi(q_i^{(t)}) + \alpha\Big(-\frac{1}{\kappa}\varsigma_i + \nabla\Psi\big(\frac{1}{N}\sum_{j=1}^N q_j^{(t)}\big)\Big)\Big), \quad \forall i \in [N], \tag{56a}$$

$$\bar{s}^{(t+1)} = \alpha \nabla \Psi\big(\frac{1}{N}\sum_{i=1}^N q_i^{(t+1)}\big) + (1-\alpha)\bar{s}^{(t)}. \tag{56b}$$

Primal-dual alternating minimization updates. Now, we are going to detail the primal-dual updates of Equation (52). Let $\tilde{s}_\otimes^{(t)} = (\tilde{s}^{(t)}, \ldots, \tilde{s}^{(t)}) \in \bar{S}_\otimes$ be a product dual vector at step t, the decomposition update in Equation (52a) can be written as

$$\begin{aligned} \tilde{q}_\otimes^{(t+1)} &= \operatorname*{argmin}_{q'_\otimes \in \Delta_\otimes} \mathcal{S}_{\Delta,N}\Big(\tilde{s}_\otimes^{(t)}, q'_\otimes\Big), \\ &= \operatorname*{argmin}_{q'_\otimes \in \Delta_\otimes} \frac{1}{N}\sum_{i=1}^N \langle \varsigma_i | q'_i\rangle + \kappa\big(\Omega_\Delta(q'_i) + \Omega^*_\Delta(\tilde{s}^{(t)}) - \langle \tilde{s}^{(t)} | q'_i\rangle\big), \\ &= \operatorname*{argmin}_{q'_\otimes \in \Delta_\otimes} \sum_{i=1}^N \Omega_\Delta(q'_i) - \langle \tilde{s}^{(t)} - \frac{1}{\kappa}\varsigma_i | q'_i\rangle, \\ &= \Big(\nabla\Omega^*_\Delta\big(\tilde{s}^{(t)} - \frac{1}{\kappa}\varsigma_i\big)\Big)_{i\in[N]}. \end{aligned}$$

The minimizer corresponding to Equation (52b) can be written as

$$\begin{aligned} &\tilde{s}_\otimes^{(t+\frac{1}{2})} \in \operatorname*{argmin}_{s_\otimes \in \bar{S}_\otimes} \mathcal{S}_{\Delta,N}\Big(s_\otimes, \tilde{q}_\otimes^{(t+1)}\Big), \\ &\tilde{s}_\otimes^{(t+\frac{1}{2})} \in \operatorname*{argmin}_{s_\otimes \in \bar{S}_\otimes} \frac{1}{N}\sum_{i=1}^N \langle \varsigma_i | \tilde{q}_i^{(t+1)}\rangle + \kappa\big(\Omega_\Delta(\tilde{q}_i^{(t+1)}) + \Omega^*_\Delta(s_i) - \langle s_i | \tilde{q}_i^{(t+1)}\rangle\big), \\ &\tilde{s}_\otimes^{(t+\frac{1}{2})} = (\tilde{s}^{(t+\frac{1}{2})}, \ldots, \tilde{s}^{(t+\frac{1}{2})}), \quad \text{where} \quad \tilde{s}^{(t+\frac{1}{2})} \in \operatorname*{argmin}_{s \in \mathbb{R}^{|\mathcal{Y}|}} \Omega^*_\Delta(s) - \langle s | \frac{1}{N}\sum_{i=1}^N \tilde{q}_i^{(t+1)}\rangle, \\ &\tilde{s}_\otimes^{(t+\frac{1}{2})} = (\tilde{s}^{(t+\frac{1}{2})}, \ldots, \tilde{s}^{(t+\frac{1}{2})}), \quad \text{where} \quad \tilde{s}^{(t+\frac{1}{2})} \in \partial\Omega_\Delta\big(\frac{1}{N}\sum_{i=1}^N \tilde{q}_i^{(t+1)}\big). \end{aligned}$$

Finally, we reformulate the primal-dual alternating minimization updates of Equation (52)

$$\tilde{q}_i^{(t+1)} = \nabla\Omega_\Delta^*\big(\tilde{s}^{(t)} - \frac{1}{\kappa}\varsigma_i\big), \quad \forall i \in [N], \tag{57a}$$

$$\tilde{s}^{(t+\frac{1}{2})} \in \partial\Omega_\Delta\big(\frac{1}{N}\sum_{i=1}^N \tilde{q}_i^{(t+1)}\big), \tag{57b}$$

$$\tilde{s}^{(t+1)} = \alpha\tilde{s}^{(t+\frac{1}{2})} + (1-\alpha)\tilde{s}^{(t)}. \tag{57c}$$

Equality of the primal iterates. We recall that we denote by $H_\Delta = \mathrm{aff}(\Delta^{\mathcal{Y}})$ the affine hull of the probability simplex, and by V_Δ the vector subspace of $\mathbb{R}^{|\mathcal{Y}|}$ which is the direction of H_Δ. We have the direct sum $\mathbb{R}^{|\mathcal{Y}|} = V_\Delta \oplus V_\Delta^\perp$. Given the iterates of mirror descent in Equation (56), and the ones of our primal-dual alternating minimization scheme stated in Equation (57), we show by induction that there exists a sequence of vectors $z^{(t)} \in V_\Delta^\perp$ such that

$$\tilde{s}^{(t)} = \bar{s}^{(t)} + z^{(t)}, \tag{58a}$$

$$\tilde{q}_i^{(t+1)} = q_i^{(t+1)}, \quad \forall i \in [N]. \tag{58b}$$

For the initialization, we consider for both sequences

$$\tilde{s}^{(0)} = \bar{s}^{(0)} = 0 \quad \text{and} \quad \tilde{q}_i^{(1)} = q_i^{(1)} = \nabla\Omega_\Delta^*\big(-\frac{1}{\kappa}\varsigma_i\big), \quad \forall i \in [N].$$

Then, we assume that Equation (58) is satisfied up to step $t \geq 0$. We are going to show it is also satisfied for step $t+1$.

First, using Equation (57), the hypothesis at step t (leading to $\tilde{q}_i^{(t+1)} = q_i^{(t+1)}$), and the link between the subdifferential of Ω_Δ and the gradient of Ψ, there exists a vector $z^{(t+\frac{1}{2})} \in V_\Delta^\perp$ such that

$$\begin{aligned}
\tilde{s}^{(t+\frac{1}{2})} &\in \partial\Omega_\Delta\big(\frac{1}{N}\sum_{i=1}^N \tilde{q}_i^{(t+1)}\big), \\
&\in \partial\Omega_\Delta\big(\frac{1}{N}\sum_{i=1}^N q_i^{(t+1)}\big), \\
&= \nabla\Psi\big(\frac{1}{N}\sum_{i=1}^N q_i^{(t+1)}\big) + z^{(t+\frac{1}{2})}.
\end{aligned}$$

From this and the hypothesis at step t leading to $\tilde{s}^{(t)} = \bar{s}^{(t)} + z^{(t)}$, we have

$$\begin{aligned}
\tilde{s}^{(t+1)} &= \alpha\tilde{s}^{(t+\frac{1}{2})} + (1-\alpha)\tilde{s}^{(t)}, \\
&= \underbrace{\alpha\nabla\Psi\big(\frac{1}{N}\sum_{i=1}^N q_i^{(t+1)}\big) + (1-\alpha)\bar{s}^{(t)}}_{\bar{s}^{(t+1)}} + \underbrace{\alpha z^{(t+\frac{1}{2})} + (1-\alpha)z^{(t)}}_{z^{(t+1)} \in V_\Delta^\perp}.
\end{aligned}$$

We have therefore shown the property for the dual iterates. We now focus on the primal updates.

Using the mirror descent Equation (56), for $i \in [N]$, we have

$$
\begin{aligned}
q_i^{(t+2)} &= \nabla\Omega_\Delta^*\Big((1-\alpha)\nabla\Psi(\underbrace{q_i^{(t+1)}}_{\tilde{q}_i^{(t+1)}}) + \alpha\big(-\frac{1}{\kappa}\varsigma_i + \nabla\Psi(\frac{1}{N}\sum_{j=1}^N q_j^{(t+1)})\big)\Big),\\
&= \nabla\Omega_\Delta^*\Big((1-\alpha)\nabla\Psi\Big(\nabla\Omega_\Delta^*(\tilde{s}^{(t)} - \frac{1}{\kappa}\varsigma_i)\Big) + \alpha\big(-\frac{1}{\kappa}\varsigma_i + \nabla\Psi(\frac{1}{N}\sum_{j=1}^N q_j^{(t+1)})\big)\Big),\\
&= \nabla\Omega_\Delta^*\Big((1-\alpha)(\tilde{s}^{(t)} - \frac{1}{\kappa}\varsigma_i + \underbrace{\tilde{z}^{(t)}}_{\in V_\Delta^\perp}) + \alpha\big(-\frac{1}{\kappa}\varsigma_i + \nabla\Psi(\frac{1}{N}\sum_{j=1}^N q_j^{(t+1)})\big)\Big),\\
&= \nabla\Omega_\Delta^*\Big(\underbrace{\alpha\nabla\Psi(\frac{1}{N}\sum_{j=1}^N q_j^{(t+1)}) + (1-\alpha)\bar{s}^{(t)}}_{\bar{s}^{(t+1)}} - \frac{1}{\kappa}\varsigma_i\Big),\\
&= \nabla\Omega_\Delta^*(\tilde{s}^{(t+1)} - \frac{1}{\kappa}\varsigma_i) = \tilde{q}_i^{(t+2)}.
\end{aligned}
$$

In the computations above, we start with the assumption at step t and the definition of the mirror descent update. We then use the definition of $\tilde{q}_i^{(t+1)}$, and the result on the composition of the gradient of Ψ and the gradient of Ω_Δ^* given in Proposition 3. Then, from the third to the fourth line we use the fact that for any $s \in \mathbb{R}^{|\mathcal{Y}|}$ and any vector $z \in V_\Delta^\perp$, we have

$$
\nabla\Omega_\Delta^*(s+z) = \operatorname*{argmax}_{q\in\Delta^{\mathcal{Y}}}\langle s+z|q\rangle - \Omega_\Delta(q) = \operatorname*{argmax}_{q\in\Delta^{\mathcal{Y}}}\langle s|q\rangle - \Omega_\Delta(q) = \nabla\Omega_\Delta^*(s),
$$

together with the assumption at step t. Last, we use the equality of the dual iterates at step $t+1$ up to a vector in $V_\Delta^\perp$ shown above. We eventually recognize the definition of $\tilde{q}_i^{(t+2)}$.

Conclusion. Therefore, we have shown by induction that the primal iterates of mirror descent stated in Equation (53) and the ones of our primal-dual algorithm in Equation (52) coincide.

□

4.5 Convergence of the algorithm

The convergence results of this section rely on the work of Léger et al. 2023 on alternating minimization algorithms as introduced in Section 2.3. More precisely, they show a linear convergence under a global condition on ϕ called the five-point property, which was first introduced by Csiszár et al. 1984. This line of work is of particular interest in our case, as the five-point property does not require any smoothness property that would be conflicting with the Legendre-type assumptions used to define the regularizers for Fenchel-Young losses in Section 3. We place ourselves in a similar setting to Section 4.4 (non-contextual stochastic optimization where we optimize directly in the distribution space). Let us first recall the five-point property and convergence result of Léger et al. 2023.

Five-point property (Léger et al. 2023, Definition 2.1) Let $\phi : Y \times Z \to \mathbb{R}^d$ be a two-variable function bounded from below on $X \times Y$. We say that ϕ follows the five-point property if for all $y \in Y$, z, $z_0 \in Z$ we have

$$
\phi(y, z_1) + \phi(y_0, z_0) \le \phi(y, z) + \phi(y, z_0), \tag{59}
$$

where $y_0 = \operatorname{argmin}_{y\in Y}\phi(y, z_0)$ and $z_1 = \operatorname{argmin}_{z\in Z}\phi(y_0, z)$.

Theorem 6. *(Léger et al. 2023, Theorem 2.3) Suppose that ϕ satisfies the five-point property (59) and consider the alternating minimization algorithm. Suppose that ϕ satisfies the* five points property*, and the minimizers exist and are uniquely attained at each step of the algorithm, then the following statements hold:*

1. $\forall n \geq 0$, $\phi(y_{n+1}, z_{n+1}) \leq \phi(y_n, z_{n+1}) \leq \phi(y_n, z_n)$.
2. *For any $y \in Y$, $z \in Z$ and any $n \geq 1$,*

$$\phi(y_n, z_n) \leq \phi(y, z) + \frac{\phi(y, z_0) - \phi(y_0, z_0)}{n}.$$

In particular, $\phi(y_n, z_n) - \min_{y \in Y, z \in Z} \phi(y, z) = \mathcal{O}(\frac{1}{n})$

First, note that all the results of Léger et al. 2023 rely on the minimum being uniquely attained at each iteration of the alternating minimization scheme. While the minima always exist in our case, the unicity is not guaranteed for the dual update (41b). As it was already noted by the authors, their results remain valid in this case, as their proofs only consider the sequence of values of the function evaluated at the successive iterates, which is uniquely defined, and not the value of the iterates themselves. This allows us to introduce the main theorem of this section.

Theorem 7. *Let us consider the case $\alpha = 1$ in the algorithm (52) (no damping). Suppose that the Jensen gap of Ψ is a convex function. Then, the iterates of XX converge in value toward the global minimum of the surrogate problem $\bar{\mathcal{S}}_{\Delta,N}$ over $(\mathbb{R}^{|\mathcal{Y}|}, \Delta_\otimes)$ with a rate $\mathcal{O}(\frac{1}{n})$.*

The main condition required for convergence is the convexity of the Jensen gap of Ψ, which is not obvious. We prove in Section 5 that this assumption holds in the case of a separable regularization, which encompasses the negative entropy or square-norm regularizers. The convexity of the Jensen gap when regularizing by structured perturbation remains an open question that is left for future research.

Proof. The proof consists in two steps. First, we express the five-point property with the notations of the problem, and obtain an equivalent form expressed with the marginal minimizer $\bar{\mathcal{S}}_{\Delta,N}$ instead of the original function $\mathcal{S}_{\Delta,N}$. Then, we show that this equivalent form holds as a consequence of the convexity of the Jensen gap of ψ, which is sufficient to prove the convergence using Theorem 6.

First, let us write the five-point property (59) with the notations of the problem, that should hold for all $q_\otimes \in \Delta$ and $s_\otimes$, $s^0_\otimes \in \bar{\mathcal{S}}_\otimes$,

$$\mathcal{S}_{\Delta,N}(s^1_\otimes, q_\otimes) + \mathcal{S}_{\Delta,N}(s^0_\otimes, q^1_\otimes) \leq \mathcal{S}_{\Delta,N}(s_\otimes, q_\otimes) + \mathcal{S}_{\Delta,N}(s^0_\otimes, q_\otimes),$$

with $q^1_\otimes = \operatorname{argmin}_{q_\otimes \in \Delta_\otimes} \mathcal{S}_{\Delta,N}(s^0_\otimes, q_\otimes)$ and $s^1_\otimes \in \operatorname{argmin}_{s_\otimes \in \mathcal{S}_\otimes} \mathcal{S}_{\Delta,N}(s_\otimes, q^1_\otimes)$. Note that the index numbering is different then in Equation (59) to adapt to our case, but the two expressions are strictly equivalent.

We observe that $s_\otimes$ appears in a single term of the equation that should hold for all $s_\otimes \in \bar{\mathcal{S}}_\otimes$. Hence, we take the minimum in $s_\otimes$ and obtain the equivalent form

$$\bar{\mathcal{S}}_{\Delta,N}(q_\otimes) + \mathcal{S}_{\Delta,N}(s^0_\otimes, q_\otimes) \geq \mathcal{S}_{\Delta,N}(s^1_\otimes, q_\otimes) + \mathcal{S}_{\Delta,N}(s^0_\otimes, q^1_\otimes).$$

We move the term $\mathcal{S}_{\Delta,N}(s^0_\otimes, q_\otimes)$ to the right hand side, and we substract $\bar{\mathcal{S}}_{\Delta,N}(q^1_\otimes) = \mathcal{S}_{\Delta,N}(s^1_\otimes, q^1_\otimes)$ from both sides to obtain

$$\bar{\mathcal{S}}_{\Delta,N}(q_\otimes) - \bar{\mathcal{S}}_{\Delta,N}(q^1_\otimes) \geq \mathcal{S}_{\Delta,N}(s^1_\otimes, q_\otimes) + \mathcal{S}_{\Delta,N}(s^0_\otimes, q^1_\otimes) - \mathcal{S}_{\Delta,N}(s^1_\otimes, q^1_\otimes) - \mathcal{S}_{\Delta,N}(s^0_\otimes, q_\otimes).$$

On the right hand side, we observe that the function $\mathcal{S}_{\Delta,N}$ is evaluated in four different points, twice in positive and twice in negative. Each of the element $s^0_\otimes$, $s^1_\otimes$, $q^1_\otimes$ and $q_\otimes$ appear exactly twice in the evaluations, once in positive and once in negative. Hence, the terms that only depend on one of the variables will cancel themselves, and the sum will only consist of the terms that depend on both variables. The five-point property then becomes

$$\bar{\mathcal{S}}_{\Delta,N}(q_\otimes) - \bar{\mathcal{S}}_{\Delta,N}(q^1_\otimes) \geq \frac{\kappa}{N}\sum_{i=1}^{N}\big(-\langle s^1_i|q_i\rangle - \langle s^0_i|q^1_i\rangle + \langle s^1_i|q^1_i\rangle + \langle s^0_i|q_i\rangle\big), \tag{60}$$

$$\bar{\mathcal{S}}_{\Delta,N}(q_\otimes) - \bar{\mathcal{S}}_{\Delta,N}(q^1_\otimes) \geq \frac{\kappa}{N}\big(\langle s^1_\otimes|q^1_\otimes\rangle + \langle s^0_\otimes|q_\otimes\rangle - \langle s^1_\otimes|q_\otimes\rangle - \langle s^0_\otimes|q^1_\otimes\rangle\big). \tag{61}$$

We will now show that the modified five-point property (61) holds in our case. According to Lemma 4, $\bar{\mathcal{S}}_{\Delta,N}$ coincides with $\hat{\mathcal{S}}_{\Delta,N}$ over $\Delta_\otimes$, which is convex by hypothesis. Hence, for all $q_\otimes, q^1_\otimes \in \operatorname{rel\,int}(\Delta_\otimes)$ we have

$$\bar{\mathcal{S}}_{\Delta,N}(q_\otimes) \geq \bar{\mathcal{S}}_{\Delta,N}(q^1_\otimes) + \langle \nabla\hat{\mathcal{S}}_{\Delta,N}(q^1_\otimes)\,|\, q_\otimes - q^1_\otimes\rangle. \tag{62}$$

We can compute the gradient as follows

$$\nabla\hat{\mathcal{S}}_{\Delta,N}(q^1_\otimes) = \left(\frac{1}{N}\Big(\varsigma_i + \kappa\big(\nabla\Psi\big(q^1_i\big) - \nabla\Psi\big(\frac{1}{N}\sum_{j=1}^{N} q^1_j\big)\Big)\right)_{i\in[N]} \tag{63}$$

Since $q^1_i = \nabla\Omega^*_\Delta(s^0 - \frac{1}{\kappa}\varsigma_i)$ according to Proposition 9, we know using Proposition 3 that there exists a vector $z_i \in V^\perp$ such that

$$\nabla\psi(q^1_i) = s^0 - \frac{1}{\kappa}\varsigma_i + z_i.$$

Similarly, we can compute

$$\begin{aligned}
s^1_\otimes &= \operatorname*{argmin}_{s_\otimes\in\bar{\mathcal{S}}_\otimes} \mathcal{S}_{\Delta,N}(s_\otimes, q^1_\otimes) \\
&= \operatorname*{argmin}_{s_\otimes\in\bar{\mathcal{S}}_\otimes} \frac{1}{N}\sum_{i=1}^{N}\langle\varsigma_i\,|\,q^1_i\rangle + \kappa\mathcal{L}_{\Omega_\Delta}(s_i, q^1_i), \\
&= \operatorname*{argmin}_{s_\otimes\in\bar{\mathcal{S}}_\otimes} \frac{1}{N}\sum_{i=1}^{N}\big(\Omega^*_\Delta(s_i) - \langle s_i\,|\,q^1_i\rangle, \\
&= \left(\operatorname*{argmin}_{s\in\mathbb{R}^\Delta} \Omega^*_\Delta(s) - \langle s\,|\,\frac{1}{N}\sum_{j=1}^{N} q^1_i\rangle\right)_{i\in[N]}, \\
&= \left(\nabla\Omega_\Delta\big(\frac{1}{N}\sum_{j=1}^{N} q^1_i\big)\right)_{i\in[N]},
\end{aligned}$$

which gives us the existence of a vector $z \in V^\perp$ such that

$$\nabla\psi\big(\frac{1}{N}\sum_{j=1}^{N} q^1_j\big) = s^1 + z.$$

By replacing the terms of Equation (63) by their expression, we get

$$\nabla \hat{\mathcal{S}}_{\Delta,N}(q^1_\otimes) = \Big(\frac{1}{N}\big(\varsigma_i + \kappa\big(s^0 - \frac{1}{\kappa}\varsigma_i + z_i - s^1 - z\big)\big)\Big)_{i\in[N]} = \Big(\frac{\kappa}{N}(s^0 - s^1 + z_i - z)\Big)_{i\in[N]}.$$

We can now express Equation (62) as the following

$$\bar{\mathcal{S}}_{\Delta,N}(q_\otimes) \geq \bar{\mathcal{S}}_{\Delta,N}(q^1_\otimes) + \frac{\kappa}{N}\langle s^0_\otimes - s^1_\otimes \,|\, q_\otimes - q^1_\otimes \rangle,$$

which is equivalent to the modified five-point property (61). We can then apply Theorem 6, concluding the proof. □

4.6 Bound on the non-optimality

As in Section 4.4, we place ourselves in the setting of (non-contextual) stochastic optimization. We therefore have no context, but a collection of noise samples $(\xi_i)_{i\in[N]}$. As before, we drop the context from the sets and functions notations. We consider another partial minimum of the surrogate risk defined in Equation (38b)

$$\underline{\mathcal{S}_{\Delta,N}}(\theta) := \min_{q_\otimes \in \Delta_\otimes} \mathcal{S}_{\Delta,N}\big((Y^\top\theta)_{i\in[N]}, q_\otimes\big), \tag{64}$$

where $\theta \in \mathbb{R}^d$ is a vector in small dimension, and $(Y^\top\theta)_{i\in[N]}$ is a collection of N identical vectors, all equal to $Y^\top\theta \in \mathbb{R}^{|\mathcal{Y}|}$, in large dimension. We derive the following bounds.

Theorem 8. *Let $\theta \in \mathbb{R}^d$ be a vector, and $\mathcal{R}_N(\theta) := \mathcal{R}_{\Delta,N}\big((Y^\top\theta)_{i\in[N]}\big)$ be the empirical risk in the stochastic optimization setting. Provided that $\nabla\Omega^*_\Delta$ is $\frac{1}{L}$-Lipschitz-continuous with respect to $||\cdot||$, the absolute difference between the empirical risk and surrogate is bounded as*

$$|\underline{\mathcal{S}_{\Delta,N}}(\theta) - \mathcal{R}_N(\theta)| \leq \frac{3}{2NL\kappa}\sum_{i=1}^N ||\varsigma_i||^2. \tag{65}$$

From this inequality, we deduce a bound on the non-optimality of the solution to the surrogate problem. Let $\theta_\mathcal{S} \in \operatorname{argmin}_\theta \underline{\mathcal{S}_{\Delta,N}}(\theta)$ and $\theta_\mathcal{R} \in \operatorname{argmin}_\theta \mathcal{R}_N(\theta)$, provided that $\nabla\Omega^*_\Delta$ is $\frac{1}{L}$-Lipschitz-continuous with respect to $||\cdot||$,

$$\mathcal{R}_N(\theta_\mathcal{S}) - \mathcal{R}_N(\theta_\mathcal{R}) \leq \frac{3}{L\kappa N}\sum_{i=1}^N ||\varsigma_i||^2. \tag{66}$$

Theorem 8 shows that the joint minimum of the surrogate function leads to a parameter $\theta_\mathcal{S}$ with bounded non-optimality with respect to the empirical risk, provided that $\nabla\Omega^*_\Delta$ is Lipschitz-continuous. In the perturbation setting, we know from Berthet et al. (2020) that $F_{\varepsilon,\mathcal{C}}$ has a Lipschitz-continuous gradient. We plan to study the extension to the sparse perturbation $F_{\varepsilon,\Delta}$ in future work.

Proof. We first give an explicit expression for the partial minimum $\underline{\mathcal{S}_{\Delta,N}}$. Then, we bound its absolute difference with the empirical risk $\mathcal{R}_N$. Last, we deduce the bound in Equation (66).

Let $\theta \in \mathbb{R}^d$,

$$\begin{aligned}
\underline{\mathcal{S}_{\Delta,N}}(\theta) &= \min_{q_\otimes \in \Delta_\otimes} \mathcal{S}_{\Delta,N}\big((Y^\top \theta)_{i\in[N]}, q_\otimes\big), \\
&= \min_{q_\otimes \in \Delta_\otimes} \frac{1}{N}\sum_{i=1}^N \langle \varsigma_i | q_i\rangle + \kappa[\Omega_\Delta(q_i) + \Omega_\Delta^*(Y^\top\theta) - \langle Y^\top \theta | q_i\rangle], \\
&= \frac{1}{N}\sum_{i=1}^N \min_{q_i \in \Delta^{\mathcal{Y}}} \langle \varsigma_i | q_i\rangle + \kappa[\Omega_\Delta(q_i) + \Omega_\Delta^*(Y^\top\theta) - \langle Y^\top \theta | q_i\rangle], \\
&= \frac{1}{N}\sum_{i=1}^N \langle \varsigma_i | \nabla\Omega_\Delta^*\big(Y^\top\theta - \frac{1}{\kappa}\varsigma_i\big)\rangle + \kappa \mathcal{L}_{\Omega_\Delta}\Big(Y^\top\theta; \nabla\Omega_\Delta^*\big(Y^\top\theta - \frac{1}{\kappa}\varsigma_i\big)\Big).
\end{aligned}$$

In the computations above, we use the definition of $\underline{\mathcal{S}_{\Delta,N}}$, we recognize N independent minimization problems, and then we use Proposition 1. 4. Let now $\theta \in \mathbb{R}^d$ and $i \in [N]$, we recast the Fenchel-Young loss as

$$\begin{aligned}
&\mathcal{L}_{\Omega_\Delta}\Big(Y^\top\theta; \nabla\Omega_\Delta^*\big(Y^\top\theta - \frac{1}{\kappa}\varsigma_i\big)\Big), \\
&= \Omega_\Delta^*\big(Y^\top\theta\big) + \Omega_\Delta\Big(\nabla\Omega_\Delta^*\big(Y^\top\theta - \frac{1}{\kappa}\varsigma_i\big)\Big) - \langle Y^\top\theta | \nabla\Omega_\Delta^*\big(Y^\top\theta - \frac{1}{\kappa}\varsigma_i\big)\rangle, \\
&= \Omega_\Delta^*\big(Y^\top\theta\big) - \Omega_\Delta^*\big(Y^\top\theta - \frac{1}{\kappa}\varsigma_i\big) - \langle \frac{1}{\kappa}\varsigma_i | \nabla\Omega_\Delta^*\big(Y^\top\theta - \frac{1}{\kappa}\varsigma_i\big)\rangle, \\
&= \int_0^1 \langle \frac{1}{\kappa}\varsigma_i | \nabla\Omega_\Delta^*\big(Y^\top\theta - \frac{1}{\kappa}\varsigma_i + \frac{u}{\kappa}\varsigma_i\big)\rangle du - \langle \frac{1}{\kappa}\varsigma_i | \nabla\Omega_\Delta^*\big(Y^\top\theta - \frac{1}{\kappa}\varsigma_i\big)\rangle, \\
&= \int_0^1 \Big[\langle \frac{1}{\kappa}\varsigma_i | \nabla\Omega_\Delta^*\big(Y^\top\theta - \frac{1}{\kappa}\varsigma_i + \frac{u}{\kappa}\varsigma_i\big)\rangle - \langle \frac{1}{\kappa}\varsigma_i | \nabla\Omega_\Delta^*\big(Y^\top\theta - \frac{1}{\kappa}\varsigma_i\big)\rangle\Big] du,
\end{aligned}$$

where we have used the equality case of the Fenchel-Young inequality. We now get the bound

$$\begin{aligned}
&\mathcal{L}_{\Omega_\Delta}\Big(Y^\top\theta; \nabla\Omega_\Delta^*\big(Y^\top\theta - \frac{1}{\kappa}\varsigma_i\big)\Big), \\
&\leq \frac{||\varsigma_i||}{\kappa}\int_0^1 \Big|\Big|\nabla\Omega_\Delta^*\big(Y^\top\theta - \frac{1}{\kappa}\varsigma_i + \frac{u}{\kappa}\varsigma_i\big) - \nabla\Omega_\Delta^*\big(Y^\top\theta - \frac{1}{\kappa}\varsigma_i\big)\Big|\Big| du, \\
&\leq \frac{||\varsigma_i||}{\kappa}\int_0^1 \frac{1}{L}\frac{u}{\kappa}||\varsigma_i|| du = \frac{||\varsigma_i||^2}{2L\kappa^2}.
\end{aligned}$$

Above we use in turn Cauchy-Schwarz inequality, and the $\frac{1}{L}$-Lipschitz-continuity of the gradient $\nabla\Omega_\Delta^*$. Let now $\theta \in \mathbb{R}^d$, we derive a bound on the following absolute difference

$$\begin{aligned}
|\underline{\mathcal{S}_{\Delta,N}}(\theta) - \mathcal{R}_N(\theta)| &\leq \frac{1}{N}\sum_{i=1}^N \big|\langle \varsigma_i | \nabla\Omega_\Delta^*\big(Y^\top\theta - \frac{1}{\kappa}\varsigma_i\big) - \nabla\Omega_\Delta^*\big(Y^\top\theta\big)\rangle\big|, \\
&\quad + \frac{\kappa}{N}\sum_{i=1}^N \mathcal{L}_{\Omega_\Delta}\Big(Y^\top\theta; \nabla\Omega_\Delta^*\big(Y^\top\theta - \frac{1}{\kappa}\varsigma_i\big)\Big), \\
&\leq \frac{1}{N}\sum_{i=1}^N ||\varsigma_i||\frac{1}{L\kappa}||\varsigma_i|| + \kappa\frac{||\varsigma_i||^2}{2L\kappa^2}, \\
&\leq \frac{3}{2NL\kappa}\sum_{i=1}^N ||\varsigma_i||^2.
\end{aligned}$$

The computations above are based on the triangular inequality, Cauchy-Schwarz inequality, the $\frac{1}{L}$-Lipschitz-continuity of the gradient $\nabla\Omega^*_\Delta$, and the bound on the Fenchel-Young loss derived above. Last, let $\theta_{\mathcal{S}} \in \operatorname{argmin}_\theta \underline{\mathcal{S}_{\Delta,N}}(\theta)$ and $\theta_{\mathcal{R}} \in \operatorname{argmin}_\theta \mathcal{R}_N(\theta)$,

$$\begin{aligned}\mathcal{R}_N(\theta_{\mathcal{S}}) - \mathcal{R}_N(\theta_{\mathcal{R}}) &= \underbrace{\mathcal{R}_N(\theta_{\mathcal{S}}) - \underline{\mathcal{S}_{\Delta,N}}(\theta_{\mathcal{S}})}_{\leq \frac{3}{2NL\kappa}\sum_{i=1}^N ||\varsigma_i||^2} + \underbrace{\underline{\mathcal{S}_{\Delta,N}}(\theta_{\mathcal{S}}) - \underline{\mathcal{S}_{\Delta,N}}(\theta_{\mathcal{R}})}_{\leq 0} + \underbrace{\underline{\mathcal{S}_{\Delta,N}}(\theta_{\mathcal{R}}) - \mathcal{R}_N(\theta_{\mathcal{R}})}_{\leq \frac{3}{2NL\kappa}\sum_{i=1}^N ||\varsigma_i||^2},\\ &\leq \frac{3}{NL\kappa}\sum_{i=1}^N ||\varsigma_i||^2.\end{aligned}$$

We have used twice the inequality above, and the definition of $\theta_{\mathcal{S}} \in \operatorname{argmin}_\theta \underline{\mathcal{S}_{\Delta,N}}(\theta)$, which concludes the proof. $\square$

5 Convexity of the Jensen gap of a separable function

In this section, we show that the Jensen gap is convex for a separable regularization Ψ, provided some regularity assumptions detailed below. This result applies, for instance, to the squared and logistic losses that are commonly used in the Fenchel-Young learning literature (Blondel et al. 2020). We plan to study the perturbation case in the future. Before proving the main result of this section, let us first introduce a preliminary lemma.

Lemma 9. *Let X be a convex subset of $\mathbb{R}$ and f a convex, positive, and twice differentiable function over X. Then the following inequality holds for all $x \in X$:*

$$f''(x)f(x) \geq 2(f'(x))^2.$$

Proof. Let us define $g : x \mapsto -\frac{1}{x}$, with domain $\operatorname{dom}(g) = \mathbb{R}_{++}$. Since g is concave and increasing, $(-f)$ is concave, the function $g \circ (-f) : x \in X \mapsto \frac{1}{f(x)}$ is concave. Furthermore, $g \circ (-f)$ is clearly twice differentiable over X as a composition of two twice differentiable functions with derivatives:

$$(g \circ (-f))'(x) = -\frac{f'(x)}{(f(x))^2}$$

$$(g \circ (-f))''(x) = -\frac{f''(x)}{(f(x))^2} + 2\frac{(f'(x))^2}{(f(x))^3}$$

The concavity of $(g \circ (-f))$ implies that, for all $x \in X$:

$$(g \circ (-f))''(x) = -\frac{f''(x)}{(f(x))^2} + 2\frac{(f'(x))^2}{(f(x))^3} \leq 0$$

Since f is strictly positive we can multiply the inequality by $(f(x))^3$ and obtain

$$f''(x)f(x) \geq 2(f'(x))^2,$$

which concludes the proof. $\square$

Proposition 11. *(Convexity of the Jensen gap of a separable regularization) Suppose that Ψ is separable by coordinate, i.e. $\Psi(y) = \sum_{j=1}^d \Psi_j(y_{(j)})$. Additionally, suppose that the functions Ψ_j are strictly convex, twice differentiable, and their second derivatives Ψ''_j are convex and positive. Then the function $F(y) = \sum_{i=1}^N \Psi(y_i) - n\Psi(\bar{y})$, where $\bar{y} = \frac{1}{N}\sum_{i=1}^N y_i$ is convex.*

Proof. First, observe that, since Ψ is separable by coordinate, we can write

$$F(y) = \sum_{j=1}^{d} \left(\sum_{i=1}^{N} \Psi_j(y_{i(j)}) - n\Psi_j(\bar{y}_{(j)}) \right),$$

where $y_{i(j)}$ is the j^{th} coordinate of vector y_i. F is a sum of d terms, each depending only on the j^{th} coordinates of vectors $(y_i)_{1\le i\le N}$, hence we can restrict the proof to the unidimensional case $d = 1$ and the general convexity result follows.

We compute the gradients of F:

$$\nabla_{y_i} F = \Psi'(y_i) - \Psi'(\bar{y})$$

$$\nabla_{y_i}\nabla_{y_k} F = \begin{cases} \Psi''(y_i) - \frac{1}{N}\Psi''(\bar{y}) & \text{if } i = k \\ -\frac{1}{N}\Psi''(\bar{y}) & \text{if } i \neq k \end{cases}$$

The Hessian matrix of F is defined as

$$H_F(y) = I_N(\Psi''(y_1), \dots, \Psi''(y_N))^\top - \frac{1}{N}\Psi''(\bar{y}) \cdot J_N$$

where $I_N \in \mathbb{R}^{N\times N}$ is the identity matrix and $J_N \in \mathbb{R}^{N\times N}$ the square matrix whose entries are all ones. In this case, F is convex if, for all $x \in \mathbb{R}^N$ and $y \in \mathrm{dom}(\Psi)^N$, the following inequality holds.

$$x^\top H_F(y)x = \sum_{i=1}^{N} x_i^2\Psi''(y_i) - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2 \Psi''(\bar{y}) \ge 0$$

We observe that this expression is exactly the Jensen gap for the function h defined as:

$$h(x, y) : \ (x, y) \mapsto x^2\Psi''(y)$$

We will now prove that h is convex by computing its Hessian matrix.

$$J_h(x, y) = \Big(2x\Psi''(y),\, x^2\Psi^{(3)}(y)\Big)$$

$$H_h(x, y) = \begin{bmatrix} 2\Psi''(y) & 2x\Psi^{(3)}(y) \\ 2x\Psi^{(3)}(y) & x^2\Psi^{(4)}(y) \end{bmatrix}$$

According to Silvester's criterion, $H_h(x, y)$ is semi-definite positive if both $2\Psi''(y) \ge 0$ and $det\big(H_h(x, y)\big) \ge 0$. The first condition is verified since Ψ is convex, and the determinant of $H_h(x, y)$ can be computed as follows:

$$det\big(H_h(x, y)\big) = 2x^2\Psi''(y)\Psi^{(4)}(y) - 4x^2(\Psi^{(3)}(y))^2 = 2x^2\left(\Psi''(y)\Psi^{(4)}(y) - 2(\Psi^{(3)}(y))^2\right)$$

By the application of Lemma 9, we obtain that $det(H_h(x, y)) \ge 0$. Hence h is convex, and its Jensen gap is non-negative, which concludes the proof. □

6 Computational experiments

Experiments design. We would like to highlight the following points. First, using a simple toy problem in Section 6.1 where every computation is straightforward, our aim is to demonstrate the effect of the perturbation scale ε (corresponding to the regularization scale κ). We expect this hyperparameter to be useful at two levels: for training the neural network with stochastic gradient descent during the coordination step, and for "exploring" beyond the purely anticipative solutions during the decomposition step.

With a more advanced application in Section 6.2, we aim to study the performance of the resulting policy trained with our primal-dual algorithm. We evaluate inference results through primal-dual steps, and focus on the gap of the corresponding solutions with respect to unseen scenarios. We hope to achieve performance close to that of a policy trained with supervised learning, assuming we have access to a dataset of target solutions to imitate. However, we highlight that this dataset is typically not accessible in the context of this paper, and only used for benchmark purposes.

6.1 Toy problem

We design a very simple problem, where the most frequent anticipative solution is the worst for the stochastic optimization problem.

Problem statement. For the sake of simplicity, we consider a constant context x, which is equivalent to having no context. The one-dimensional solution space is $\mathcal{Y} = \{0, 1\}$, and the exogenous noise follows a uniform distribution over a three-states space $\boldsymbol{\xi} \sim \mathcal{U}(\{\xi_1, \xi_2, \xi_3\})$. We recall that the only information we can get about the noise $\boldsymbol{\xi}$ is through sampling. We do not have access to its distribution. In this tabular setting we can explicitly define the cost as done in Table 1.

	Scenario ξ_1	Scenario ξ_2	Scenario ξ_3
Solution 0	4	-1	-2
Solution 1	0	0	0

Table 1: Tabular definition of the cost for the toy problem.

The problem we would like to solve is the following:

$$\min_{y \in \{0,1\}} \mathbb{E}_{\boldsymbol{\xi} \sim \mathcal{U}(\xi_1, \xi_2, \xi_3)} \big[c(y, \boldsymbol{\xi}) \big]. \tag{67}$$

Stochastic and anticipative solutions. It is immediate to derive both the anticipative solutions per scenario, and the solution of the stochastic problem (67). The scenario ξ_1 gives $y = 1$ as optimal solution whereas scenarios ξ_2 and ξ_3 lead to the solution $y = 0$. Given the uniform distribution, the optimal solution of Equation (67) is $y = 1$. We see that the most represented solution among the anticipative ones is not the optimal solution of our stochastic problem.

Learning the optimal decision. We recall that we have access to a training set denoted as $\mathcal{D}_{\text{train}} = (\xi_i)_{i \in [n_{\text{train}}]}$. Our algorithms enable us to derive a policy by learning the parameters w of a statistical model φ_w. In our case, since we have no context, we directly learn a parameter θ to derive a solution of the stochastic problem (67).

Primal-dual algorithm. We implement the primal-dual algorithm, in a perturbation framework. Our hyperparameters are summarized in Table 2. We highlight that in our case, the cost function c is linear.

Name	Definition	Value
lr	Learning rate for the SGD	10^{-1}
nb_epochs	Number of epochs for the SGD	10
nb_samples	Number of perturbation samples	10^3
nb_iterations	Number of outer iterations	20
nb_seeds	Number of seeds to average over experiments	30
ε	Scale of the perturbation (primal and dual updates)	varies
α	Momentum coefficient for the dual update	0.5

Table 2: Hyper-parameters for the experiments on the toy problem.

We save the value of θ through the outer iterations of the primal-dual algorithm, and we compute its average value. This latter can then be used to derive a solution of the stochastic problem (67). We do this for different values of the perturbation (or regularization) scale ε, and average over nb_seeds = 30 seeds. Our aim is to show that regularizing helps. We highlight the results of our experiments with Figure 2. On the left plot we show ε ranging from 1 to 150. On the right we zoom at the critical range from 2 to 4. The curves are the proportions (over seeds) of averaged values of θ (over outer iteration) leading to the optimal solution.

First, we see that when the scale of perturbation ε is too small, typically smaller than 2.7 in this toy problem example, we learn a θ value that leads to the majority anticipative solution, which is not the optimal solution of problem (67). It is natural because the smaller is ε, the closer the primal variables are to anticipative decisions in the first outer iteration. The learning update – or dual update – then pushes θ to lead to the majority decision. In the next iterations, the primal variables are even more in accordance and in the wrong direction. Indeed, θ adds some weight on the wrong direction, and the regularization (perturbation) term in the primal update is too small to deviate. We thus get final and average values of θ leading to the majority anticipative decision which is suboptimal.

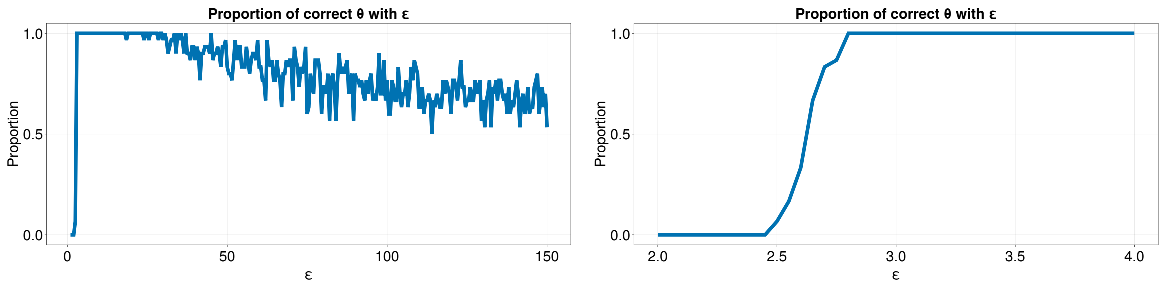

Figure 2: Proportion of θ values giving the optimal solution of problem (67) when ε varies.

Result 1. *When the scale of the perturbation ε is too small, we learn θ to imitate the majority anticipative solutions and possibly get poor performance.*

Then, when ε exceeds a certain threshold, the proportion over seeds of θ leading to the optimal solution increases, and reaches 1. It brings to light that regularizing in the primal updates is crucial. Nonetheless, the left plot illustrates that when ε reaches large values, the performance worsens. Indeed, very large perturbation or regularization scales lead to uninformative objective function in the primal problems. The resulting primal updates are random, in our case either 0 or 1 with same probability. We indeed observe an asymptotic proportion of 0.5 in the left plot.

Result 2. *When the scale of the perturbation ε is too large, it dominates the cost per scenario and the parameter θ, leading to random values of θ and poor performance.*

6.2 Two-stage minimum weight spanning tree

We now consider a richer contextual stochastic optimization problem, and compare several policies parameterized by neural networks.

Problem statement. Let $G = (V, E)$ be an undirected graph, and let $\boldsymbol{\xi} \in \Xi$ be an exogenous noise. The goal is to build a spanning tree on G over two stages at minimum cost, the second-stage building costs – depending on $\boldsymbol{\xi}$ – being unknown when first-stage decisions are taken. Nonetheless, we have access to some context random variable $\mathbf{x} \in \mathcal{X}$ correlated to $\boldsymbol{\xi}$. For each edge $e \in E$ and scenario $\xi \in \Xi$, we denote by c_e the scenario-independent first stage cost of building e, and by $d_e(\xi)$ the scenario-dependent second stage cost. Our contextual stochastic two-stage minimum weight spanning tree problem is Equation (68).

$$\begin{aligned}
\min_{\pi \in \mathcal{H}} \quad & \mathbb{E}_{\mathbf{x}}\Bigg[\sum_{e\in E} c_e \pi(\mathbf{x})_e + \mathbb{E}_{\boldsymbol{\xi}}\big[Q\big(\pi(\mathbf{x}); \boldsymbol{\xi}\big)\big]\Bigg], && && \text{(68a)}\\
\text{s.t.} \quad & \sum_{e\in E(Y)} \pi(x)_e \le |Y| - 1, && \forall x \in \mathcal{X}, \forall Y, \emptyset \subsetneq Y \subseteq V, && \text{(68b)}\\
& \pi(x)_e \in \{0,1\}, && \forall x \in \mathcal{X}, \forall e \in E. && \text{(68c)}
\end{aligned}$$

Notice that constraints (68b)-(68c) lead to a solution space $\mathcal{Y}(x)$ corresponding to the forests on G. The objective function in this case is given by $c^0(x; y) = \sum_{e\in E} c_e y_e + \mathbb{E}_{\boldsymbol{\xi}}\big[Q\big(y; \boldsymbol{\xi}\big)\big]$. The second stage is encapsulated in the function Q, as defined in Equation (69).

$$\begin{aligned}
Q\big(\pi(x); \xi\big) := \min_{z} \quad & \sum_{e\in E} d_e(\xi) z_e, && && \text{(69a)}\\
\text{s.t.} \quad & \sum_{e\in E} \pi(x)_e + z_e = |V| - 1, && && \text{(69b)}\\
& \sum_{e\in E(Y)} \pi(x)_e + z_e \le |Y| - 1, && \forall Y, \emptyset \subsetneq Y \subseteq V, && \text{(69c)}\\
& z_e \in \{0,1\}, && \forall e \in E. && \text{(69d)}
\end{aligned}$$

In problem (68), we look for a policy π that maps a context realization $x \in \mathcal{X}$ to a good first-stage forest. In this case, $\pi(x)_e = 1$ if and only if edge e is selected at first stage to build our spanning tree given a context value x. Similarly, variable z in problem (69) is such that $z_e = 1$ if and only if edge e is selected at second stage to build a spanning tree. We force integer decision variables with Equations (68c) and (69d), and constraints (69b) and (69c) make sure that we define a tree with the variable $\pi(x) + z$.

Remark 4. *In practice, we can even make the structure of the graph depend on the variable $\mathbf{x}$, but we intentionally ease notations above.*

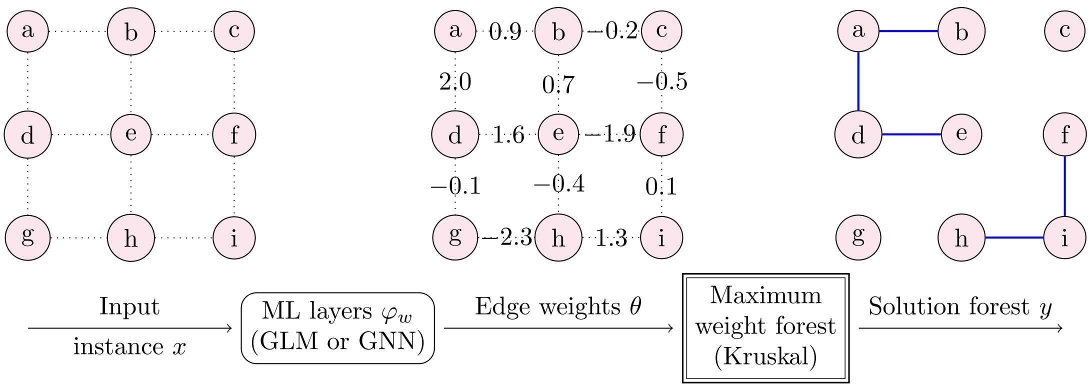

Figure 3: Two-stage minimum spanning neural network.

For this example, we define a neural network illustrated by Figure 3. Each edge e of an instance (context) x is encoded by a feature vector $\phi(x, e)$. The feature matrix is given as input to ML layers φ_w – a GLM or a graph neural network (GNN) – with learnable parameters w, which predicts edge weights θ_e. We use the predicted edge weights θ as the objective of a maximum weight forest problem layer (70). A maximum weight forest can be efficiently found using Kruskal's algorithm.

$$\pi_w(x) := \operatorname*{argmax}_y \sum_{e\in E} \varphi_w(x)_e y_e, \tag{70a}$$

$$\text{s.t.} \sum_{e\in E(Y)} y_e \leq |Y| - 1, \qquad \forall Y, \emptyset \subsetneq Y \subseteq V, \tag{70b}$$

$$y_e \in \{0, 1\}, \qquad \forall e \in E. \tag{70c}$$

Learning and evaluation. We recall that we have access to a training data set including context and noise realizations $\mathcal{D}_{\text{train}} = (x_i, \xi_i)_{i\in[n_{\text{train}}]}$ to learn the parameters w of a statistical model φ_w, used to parameterize a CO oracle f. Once the weights w are learned, we can evaluate the resulting policy $\pi_w : x \mapsto f \circ \varphi_w(x)$ for our problem (68). To do so, we have access to test $\mathcal{D}_{\text{test}} = (x_i, \xi_i)_{i\in[n_{\text{test}}]}$ and validation $\mathcal{D}_{\text{val}} = (x_i, \xi_i)_{i\in[n_{\text{val}}]}$ datasets, and we can approximate the two expectations in the objective function of Equation (68) by Monte-Carlo averages over the datasets.

Benchmarks. We derive two benchmark policies for this problem. The first one denoted as π_{median} is very simple, and not based on any learning process. It consists in solving a deterministic problem given a context realization x as follows:

$$\pi_{\text{median}}(x) := \operatorname*{argmin}_y \min_z \sum_{e\in E} c_e y_e + \hat{d}_e(x) z_e, \tag{71a}$$

$$\text{s.t.} \sum_{e\in E} y_e + z_e = |V| - 1, \tag{71b}$$

$$\sum_{e\in E(Y)} y_e + z_e \leq |Y| - 1, \qquad \forall Y, \emptyset \subsetneq Y \subseteq V, \tag{71c}$$

$$y_e \in \{0, 1\}, \qquad \forall e \in E, \tag{71d}$$

$$z_e \in \{0, 1\}, \qquad \forall e \in E. \tag{71e}$$

In the deterministic problem (71), we replace the unknown (random) cost vector of the second stage by an estimator of its median, which is estimated by sampling and included in the context x.

Our second benchmark policy is more sophisticated, and based on an additional dataset which is typically not available in the context of this paper. We follow the approach of Dalle et al. (2022) where for each context x_i of our training dataset we have several realizations of the endogenous noise $\boldsymbol{\xi}$, which enables us to derive Lagrangian heuristic-based labels y_i^L. It leads to a richer imitation learning dataset $\mathcal{D}_{\text{train}}^L = (x_i, y_i^L)_{i\in[n_{\text{train}}]}$. We leverage this latter to imitate the solutions of the Lagrangian heuristic, leading to some parameters w^L. We denote by π_{w^L} the induced policy, and we typically expect this latter to behave very well. Indeed, the Lagrangian heuristic we implement is known to lead to very good solutions, requiring heavy computations though. Besides, Fenchel-Young imitation learning in this setting has very good generalization performance. We refer to Dalle et al. (2022, Section 6.5) to have more details on this benchmark policy.

Primal-dual algorithm. We evaluate the quality of the policy learned with our primal-dual algorithm. Our hyperparameters are defined in Table 3. The instances we consider are defined on grid-graphs (see Figure 3) with grid_size $= 20 \times 20$ nodes. In our training set, we have train_size instances, each with 20 scenarios. At each (outer) iteration, we randomly sub-sample nb_scenarios $= 10$ scenarios per instance. It leads to reduced computational time, and good performance as we show below.

Name	**Definition**	**Value**
grid_size	Size of the squared grid-graph instances	20×20
train_size	Number of train samples	50
val_size	Number of validation samples	50
test_size	Number of test samples	50
lr_init	Initial learning rate for Adam optimizer (dual update)	10^{-5}
nb_epochs	Number of epochs for the SGD (dual update)	30
nb_scenarios	Number of scenarios sub-sampled per instance	10
nb_samples	Number of perturbation samples (primal and dual updates)	20
nb_iterations	Number of outer iterations	50
ε	Scale of the perturbation (primal and dual updates)	10^{-4}
α	Momentum coefficient for the dual update	0.1

Table 3: Constants and hyperparameters for the primal-dual algorithm on the maximum weight two-stage spanning tree problem.

We plot validation and test estimated average gaps over iterations in Figure 4, with a focus on four policies. The dotted black line corresponds to the simple benchmark policy π_{median}, which does not evolve over iterations. We observe a poor performance for this simple approach, with roughly 12% average validation and test gaps. The green dashed line corresponds to the policy derived by imitating the Lagrangian heuristic, denoted as π_{w^L}. As expected it reaches very good performance, with average gaps smaller than 2%. Then, we can evaluate two different policies learned with our primal-dual algorithm. The first one – in dotted-dashed blue line – uses the current outer iteration weights $w^{(t)}$ to parameterize the problem (70). We denote it by $\pi_{w^{(t)}}$. The second one – in solid red line – uses the average weights over the past outer iterations $\bar{w}^{(t)} = \frac{1}{t}\sum_{t'\leq t} w^{(t')}$ to parameterize the same problem. We name this policy $\pi_{\bar{w}}$ in the legend of Figure 4. We observe that using the current weights with policy $\pi_{w^{(t)}}$ leads to small oscillations over outer iterations. It can be understood based on the alternating nature of the primal-dual algorithm. The sequence that converges in classic mirror descent framework is the average primal variable. In our primal-dual approach, we average the weights of the neural networks over outer iterations, and we indeed observe a convergence behaviour with the policy $\pi_{\bar{w}}$. Maybe more surprisingly, the limit policy reaches the performance of the Lagrangian heuristic-based policy φ_{w^L}. This is very interesting in practice, because the Lagrangian heuristic is computationally heavy, and typically can not be applied to very large instances.

Result 3. *The average weights policy $\pi_{\bar{w}}$ converges, and reaches the performance of the computationally demanding Lagrangian heuristic-based benchmark.*

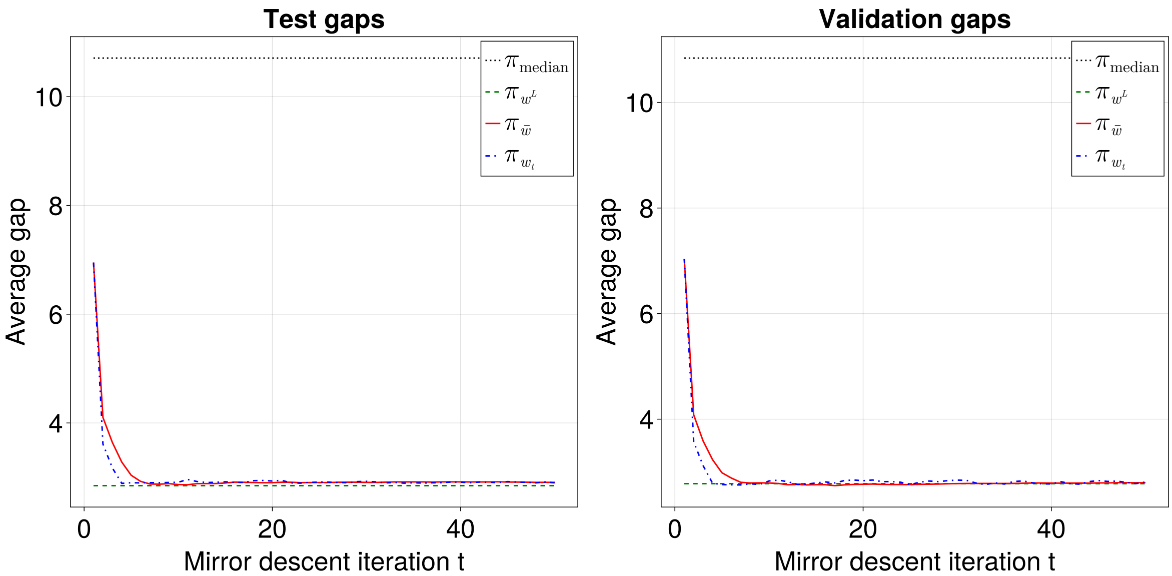

Figure 4: Validation and test average gaps of policies over the iterations of our primal-dual algorithm, for the two-stage minimum weight spanning tree.

7 Conclusion and perspectives

In this paper, we explore policies for contextual stochastic combinatorial optimization, which are based on neural networks with combinatorial optimization (CO) layers. The natural paradigm for training these networks is empirical risk minimization. However, the score function estimator has a high variance, which can hinder the efficient determination of the neural network weights.

To address this, we design a surrogate learning problem based on Fenchel-Young losses and introduce a novel primal-dual alternating minimization scheme to solve it. We study the structure of these primal-dual updates using convex analysis tools and demonstrate their practical tractability. Specifically, primal updates decompose per scenario, similar to classic stochastic programming approaches. Dual updates correspond to a supervised learning problem with Fenchel-Young loss, which we solve using stochastic gradient descent. In both cases, we rely on sampling and leverage an oracle for a simpler deterministic problem, allowing us to scale to relatively large instances.

8 Appendix

8.1 Proof of Lemma 2

Lemma 2. Let $\mathcal{Y} \subset \mathbb{R}^d$ be a finite set. We make the assumption that no element of $\mathcal{Y}$ is a strict convex combination of other elements of $\mathcal{Y}$. In other words, $\mathcal{Y}$ is the set of vertices of the polytope $\mathcal{C} = \text{conv}(\mathcal{Y})$. Let $\Delta^{\mathcal{Y}}$ be the probability simplex whose vertices are indexed by $\mathcal{Y}$, and H_Δ its affine hull $H_\Delta = \text{aff}(\Delta^{\mathcal{Y}})$. We denote by V_Δ the vector subspace of $\mathbb{R}^{|\mathcal{Y}|}$ which is the direction of H_Δ, and we have the orthogonal sum $\mathbb{R}^{|\mathcal{Y}|} = V_\Delta \oplus \text{span}(\mathbf{1})$. For any vector $s \in \mathbb{R}^{|\mathcal{Y}|}$, we denote by $s(y) \in \mathbb{R}$ the component of s indexed by $y \in \mathcal{Y}$. We also consider $\varepsilon \in \mathbb{R}_{++}$ a positive real number, and $\mathbf{Z}$, a random variable with standard multivariate Gaussian distribution over $\mathbb{R}^d$.

Then, for two vectors $(s_1, s_2) \in (V_\Delta)^2$, $s_1 \neq s_2$,

$$\mathbb{P}_{\mathbf{Z}}\left(\operatorname*{argmax}_{y\in\mathcal{Y}}\{s_1(y) + \varepsilon \mathbf{Z}^\top y\} \cap \operatorname*{argmax}_{y\in\mathcal{Y}}\{s_2(y) + \varepsilon \mathbf{Z}^\top y\} = \emptyset \right) > 0.$$

Proof. Let $(s_1, s_2) \in (V_\Delta)^2$, $s_1 \neq s_2$ be two vectors. We are going to show that

$$\operatorname*{argmax}_{y\in\mathcal{Y}}\{s_1(y) + \varepsilon Z^\top y\} \cap \operatorname*{argmax}_{y\in\mathcal{Y}}\{s_2(y) + \varepsilon Z^\top y\} = \emptyset,$$

for Z almost everywhere with respect to the Lebesgue measure in an open ball in $\mathbb{R}^d$. Since $\mathbf{Z}$ follows a non-degenerate Gaussian measure, the result yields.

Suppose first that

$$\operatorname*{argmax}_{y\in\mathcal{Y}} s_1(y) \cap \operatorname*{argmax}_{y\in\mathcal{Y}} s_2(y) = \emptyset.$$

Then, a ball centered on $0_{\mathbb{R}^d}$ with sufficiently small radius gives the result.

Let now y^* be a common maximizer of s_1 and s_2. Since s_1 and s_2 are elements of $V_\Delta = \text{span}(\mathbf{1})^\perp$, and they are distinct, they are not equal up to a constant. We can thus fix a $\bar{y} \in \mathcal{Y}$ such that

$$s_1(\bar{y}) - s_1(y^*) \neq s_2(\bar{y}) - s_2(y^*).$$

In particular, it implies that $\bar{y}$ does not belong to the intersection of the argmax,

$$\bar{y} \notin \operatorname*{argmax}_{y\in\mathcal{Y}} s_1(y) \cap \operatorname*{argmax}_{y\in\mathcal{Y}} s_2(y).$$

Let $\bar{Z}$ be in the relative interior of the normal cone of $\text{conv}(\mathcal{Y})$ at $\bar{y}$, such that the function g defined as

$$g : y \mapsto \varepsilon \bar{Z}^\top y,$$

is injective over $\mathcal{Y}$. It is possible since the normal cone is full dimensional, and the union of the hyperplanes where two dot products are equal is not full dimensional. Then, for $\lambda \in \mathbb{R}_+$ sufficiently large,

$$\operatorname*{argmax}_{y\in\mathcal{Y}} s_1(y) + \varepsilon\lambda \bar{Z}^\top y = \operatorname*{argmax}_{y\in\mathcal{Y}} s_2(y) + \varepsilon\lambda \bar{Z}^\top y = \{\bar{y}\}.$$

For $i \in \{1, 2\}$, let us define F_i as the single variable function

$$F_i : \lambda \in \mathbb{R} \mapsto \max_{y\in\mathcal{Y}} \left(s_i(y) + \varepsilon\lambda \bar{Z}^\top y\right) - s_i(y^*).$$

Note that if there exists $\lambda^* \in \mathbb{R}$ such that $\operatorname{argmax}_{y\in\mathcal{Y}} \left(s_i(y) + \lambda^* \varepsilon \bar{Z}^\top y\right)$ for $i \in \{1,2\}$ are disjoint singletons, then considering a small enough open ball around the vector $\lambda^* \bar{Z}$ gives the result. We now show that such a λ^* exists.

Since $\mathcal{Y}$ is finite, for $i \in \{1,2\}$, F_i is a maximum of a finite number of affine functions in λ, it is therefore a piecewise affine function. Since g is injective, for $i \in \{1,2\}$, there exists a collection of $k_i + 1$ real numbers, $k_i \in \mathbb{N}$, $k_i > 1$, denoted as $0 = \lambda_0^i < \ldots < \lambda_{k_i}^i$ and a unique collection $y^* = y_1^i, \ldots, y_{k_i}^i = \bar{y}$ of two-by-two distinct vectors such that

$$F_i(\lambda) = s_i(y_j^i) + \varepsilon\lambda\bar{Z}^\top y_j^i - s_i(y^*), \quad \forall \lambda \in [\lambda_{j-1}^i, \lambda_j^i].$$

Furthermore, the fact that g is injective implies that y_j^i is the unique maximizer over $\mathcal{Y}$ of $y \mapsto s_i(y) + \varepsilon\lambda\bar{Z}^\top y$ for $\lambda \in]\lambda_{j-1}^i, \lambda_j^i[$. If the collections for $i \in \{1,2\}$ are not identical, the proof is finished. By contradiction, we assume that the two collections are identical and denote by $0 = \lambda_0 < \ldots < \lambda_k$ and $y^* = y_1, \ldots, y_k = \bar{y}$ the common collections.

Let $\tilde{j}$ be the smallest $j' \in [k]$ such that $s_1(y_{j'}) - s_1(y^*) \neq s_2(y_{j'}) - s_2(y^*)$. It exists since the inequality holds for $\bar{y}$. Without loss of generality, we can assume from now on that $s_1(y_{\tilde{j}}) - s_1(y^*) > s_2(y_{\tilde{j}}) - s_2(y^*)$. For $i \in \{1,2\}$, we define $\tilde{\lambda}_i$ as:

$$\tilde{\lambda}_i := \min\{\lambda \mid y_{\tilde{j}} \in \operatorname*{argmax}_{y\in\mathcal{Y}} \left(s_i(y) + \varepsilon\lambda\bar{Z}^\top y\right)\}.$$

We eventually get a contradiction by proving $\tilde{\lambda}_1 < \tilde{\lambda}_2$ with the following inequality.

$$\begin{aligned}
& s_2(y_{\tilde{j}}) - s_2(y^*) + \tilde{\lambda}_1 g(y_j), && && \text{(72a)}\\
& < s_1(y_{\tilde{j}}) - s_1(y^*) + \tilde{\lambda}_1 g(y_j), && \text{(hypothesis right above)}, && \text{(72b)}\\
& = s_1(y_{\tilde{j}-1}) - s_1(y^*) + \tilde{\lambda}_1 g(y_{\tilde{j}-1}), && \text{(Both } y_{\tilde{j}} \text{ and } y_{\tilde{j}-1} \text{ are optimal at the junction)}, && \text{(72c)}\\
& = s_2(y_{\tilde{j}-1}) - s_2(y^*) + \tilde{\lambda}_1 g(y_{\tilde{j}-1}), && \text{(by definition of } \tilde{j}). && \text{(72d)}
\end{aligned}$$

Hence, for $\lambda = \tilde{\lambda}_1 + \eta$ with $\eta > 0$ sufficiently small, $y_{\tilde{j}}$ is the unique argmax of $s_1(y) + \lambda g(y)$ and $y_{\tilde{j}-1} \neq y_{\tilde{j}}$ is the unique argmax of $s_2(y) + \lambda g(y)$. □

References

Abernethy, J., C. Lee, and A. Tewari (2016). "Perturbation Techniques in Online Learning and Optimization". In: *Perturbations, Optimization, and Statistics*. Ed. by T. Hazan, G. Papandreou, and D. Tarlow. The MIT Press, pp. 233–264. ISBN: 978-0-262-33793-9. DOI: 10.7551/mitpress/10761.003.0009.

Attouch, H., J. Bolte, and B. F. Svaiter (2013). "Convergence of Descent Methods for Semi-Algebraic and Tame Problems: Proximal Algorithms, Forward–Backward Splitting, and Regularized Gauss–Seidel Methods". In: *Mathematical Programming* 137.1, pp. 91–129. ISSN: 1436-4646. DOI: 10.1007/s10107-011-0484-9.

Auslender, A. (1971/1972). "Méthodes numériques pour la décomposition et la minimisation de fonctions non differentiates." In: *Numerische Mathematik* 18, pp. 213–223. ISSN: 0029-599X; 0945-3245/e.

Azizan, N., S. Lale, and B. Hassibi (2022). "Stochastic Mirror Descent on Overparameterized Nonlinear Models". In: *IEEE Transactions on Neural Networks and Learning Systems* 33.12, pp. 7717–7727. ISSN: 2162-237X, 2162-2388. DOI: 10.1109/TNNLS.2021.3087480.

Bauschke, H. H. and P. L. Combettes (2017). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Cham: Springer International Publishing. ISBN: 978-3-319-48310-8 978-3-319-48311-5. DOI: 10.1007/978-3-319-48311-5.

Beck, A. and M. Teboulle (2003). “Mirror Descent and Nonlinear Projected Subgradient Methods for Convex Optimization”. In: *Operations Research Letters* 31.3, pp. 167–175. ISSN: 0167-6377. DOI: 10.1016/S0167-6377(02)00231-6.

Beck, A. and L. Tetruashvili (2013). “On the Convergence of Block Coordinate Descent Type Methods”. In: *SIAM Journal on Optimization* 23.4, pp. 2037–2060. ISSN: 1052-6234. DOI: 10.1137/120887679.

Berthet, Q., M. Blondel, O. Teboul, M. Cuturi, J.-P. Vert, and F. Bach (2020). “Learning with Differentiable Pertubed Optimizers”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 9508–9519.

Bertsekas, D. P., ed. (2009). *Convex Optimization Theory*. Belmont, Mass: Athena Scientific. ISBN: 978-1-886529-31-1.

Blondel, M., A. F. T. Martins, and V. Niculae (2020). “Learning with Fenchel-Young Losses”. In: *Journal of Machine Learning Research* 21.35, pp. 1–69. ISSN: 1533-7928.

Bolte, J., S. Sabach, and M. Teboulle (2014). “Proximal Alternating Linearized Minimization for Nonconvex and Nonsmooth Problems”. In: *Mathematical Programming* 146.1, pp. 459–494. ISSN: 1436-4646. DOI: 10.1007/s10107-013-0701-9.

Bubeck, S. (2015). “Convex Optimization: Algorithms and Complexity”. In: *Foundations and Trends® in Machine Learning* 8.3-4, pp. 231–357. ISSN: 1935-8237, 1935-8245. DOI: 10.1561/2200000050.

Csiszár, I. and G. Tusnády (1984). “Information Geometry and Alternating Minimization Procedures”. In: *Statistics and Decisions, Dedewicz* 1, pp. 205–237.

D’Orazio, R., N. Loizou, I. H. Laradji, and I. Mitliagkas (2023). “Stochastic Mirror Descent: Convergence Analysis and Adaptive Variants via the Mirror Stochastic Polyak Stepsize”. In: *Transactions on Machine Learning Research*. ISSN: 2835-8856.

Dalle, G., L. Baty, L. Bouvier, and A. Parmentier (2022). *Learning with Combinatorial Optimization Layers: A Probabilistic Approach*. arXiv: 2207.13513 [cs, math, stat].

Dang, C. D. and G. Lan (2015). “Stochastic Block Mirror Descent Methods for Nonsmooth and Stochastic Optimization”. In: *SIAM Journal on Optimization* 25.2, pp. 856–881. ISSN: 1052-6234, 1095-7189. DOI: 10.1137/130936361.

Duchi, J. C., A. Agarwal, and M. J. Wainwright (2012). “Dual Averaging for Distributed Optimization: Convergence Analysis and Network Scaling”. In: *IEEE Transactions on Automatic Control* 57.3, pp. 592–606. ISSN: 0018-9286, 1558-2523. DOI: 10.1109/TAC.2011.2161027.

Frank, M. and P. Wolfe (1956). “An Algorithm for Quadratic Programming”. In: *Naval Research Logistics Quarterly* 3.1-2, pp. 95–110. ISSN: 0028-1441, 1931-9193. DOI: 10.1002/nav.3800030109.

Grünwald, P. D. and A. P. Dawid (2004). “Game Theory, Maximum Entropy, Minimum Discrepancy and Robust Bayesian Decision Theory”. In: *The Annals of Statistics* 32.4, pp. 1367–1433. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/009053604000000553.

Juditsky, A. B., A. V. Nazin, A. B. Tsybakov, and N. Vayatis (2005). “Recursive Aggregation of Estimators by the Mirror Descent Algorithm with Averaging”. In: *Problems of Information Transmission* 41.4, pp. 368–384. ISSN: 1608-3253. DOI: 10.1007/s11122-006-0005-2.

Juditsky, A., J. Kwon, and É. Moulines (2023). “Unifying Mirror Descent and Dual Averaging”. In: *Mathematical Programming* 199.1, pp. 793–830. ISSN: 1436-4646. DOI: 10.1007/s10107-022-01850-3.

Léger, F. and P.-C. Aubin-Frankowski (2023). *Gradient Descent with a General Cost*. arXiv: 2305.04917 [math].

Li, J., G. Li, Z. Wu, and C. Wu (2018). “Stochastic Mirror Descent Method for Distributed Multi-Agent Optimization”. In: *Optimization Letters* 12.6, pp. 1179–1197. ISSN: 1862-4472, 1862-4480. DOI: 10.1007/s11590-016-1071-z.

Nemirovsky, A., D. Yudin, and E. Dawson (1983). *Wiley-Interscience Series in Discrete Mathematics*.

Nesterov, Y. (2009). "Primal-Dual Subgradient Methods for Convex Problems". In: *Mathematical Programming* 120.1, pp. 221–259. ISSN: 0025-5610, 1436-4646. DOI: 10.1007/s10107-007-0149-x.

Powell, M. J. D. (1973). "On Search Directions for Minimization Algorithms". In: *Mathematical Programming* 4.1, pp. 193–201. ISSN: 1436-4646. DOI: 10.1007/BF01584660.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton: Princeton University Press. ISBN: 978-1-4008-7317-3. DOI: doi:10.1515/9781400873173.

Sadana, U., A. Chenreddy, E. Delage, A. Forel, E. Frejinger, and T. Vidal (2024). "A Survey of Contextual Optimization Methods for Decision-Making under Uncertainty". In: *European Journal of Operational Research*. ISSN: 0377-2217. DOI: 10.1016/j.ejor.2024.03.020.

Wright, S. J. (2015). "Coordinate Descent Algorithms". In: *Mathematical Programming* 151.1, pp. 3–34. ISSN: 1436-4646. DOI: 10.1007/s10107-015-0892-3.

Yuan, D., Y. Hong, D. W. C. Ho, and G. Jiang (2018). "Optimal Distributed Stochastic Mirror Descent for Strongly Convex Optimization". In: *Automatica* 90, pp. 196–203. ISSN: 0005-1098. DOI: 10.1016/j.automatica.2017.12.053.

Zhou, Z., P. Mertikopoulos, N. Bambos, S. P. Boyd, and P. W. Glynn (2020). "On the Convergence of Mirror Descent beyond Stochastic Convex Programming". In: *SIAM Journal on Optimization* 30.1, pp. 687–716. ISSN: 1052-6234, 1095-7189. DOI: 10.1137/17M1134925.