Primal-dual algorithm for contextual stochastic combinatorial optimization

Louis Bouvier[1*], Thibault Prunet[2], Axel Parmentier[2], Vincent Leclère[2]

[1*]Co-innovation lab, Ecole Nationale des Ponts et Chaussées, avenue Blaise Pascal, Champs sur Marne, 77455, France.
[2]CERMICS, Ecole Nationale des Ponts et Chaussées, avenue Blaise Pascal, Champs sur Marne, 77455, France.

*Corresponding author(s). E-mail(s): louis.bouvier@enpc.fr;
Contributing authors: thibault.prunet@enpc.fr; axel.parmentier@enpc.fr; vincent.leclere@enpc.fr;

Abstract

This paper introduces a novel approach to contextual stochastic optimization, combining operations research and machine learning to address decision-making under uncertainty. Traditional methods fall short in leveraging contextual information, prompting the need for new algorithms. We use neural networks with combinatorial optimization layers to encode policies. Our goal is to minimize the empirical risk, estimated from past data on uncertain parameters and contexts. To that end, we present a surrogate learning problem and a generic primal-dual algorithm applicable to various combinatorial stochastic optimization settings. Our approach extends classic Fenchel-Young loss results and introduces a new regularization method using sparse perturbations on the distribution simplex. This allows for tractable updates in the original space and handles diverse objective functions.

We demonstrate the linear convergence of our algorithm under certain conditions and provide a bound on the non-optimality of the resulting policy regarding empirical risk. Experiments on a contextual stochastic minimum weight spanning tree problem show that our algorithm is efficient and scalable, outperforming a sophisticated Lagrangian-based heuristic with reduced computational effort.

Keywords: contextual stochastic combinatorial optimization, empirical risk minimization, neural networks with combinatorial optimization layers, alternating minimization, Fenchel-Young loss

1 Introduction

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, which means 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. Assuming that the policy π has to belong to some hypothesis class $\mathcal{H}$, our *contextual stochastic optimization* problem [1] 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 $\mathbf{x}$. In many situations, the noise ξ is observed once the decision has been taken, and the training set comes from historical data, we thus place ourselves in a *learning* setting.

Assumption 1. *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 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 2. *For every possible context* x*, the set of admissible decisions* $\mathcal{Y}(x) \subset \mathbb{R}^{d(x)}$ *is finite. Further, we assume that* $\mathcal{Y}(x)$ *is the set of exposed vertices of its convex envelope* $\mathcal{C}(x) = \operatorname{conv}(\mathcal{Y}(x))$*, i.e., there are no* $y \in \mathcal{Y}(x)$ *is a strict convex combination of other points in* $\mathcal{Y}(x)$*.*

Hence, for any $\bar{y}$ in $\mathcal{Y}(x)$, there exists a $\theta \in \mathbb{R}^{d(x)}$ such that $\bar{y}$ is the unique argmax of $\max_{y\in\mathcal{Y}(x)} \langle \theta | y \rangle$, which enables to build policies based on linear optimizers.

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 Equation (2). Decomposition-coordination methods such as progressive hedging [2] 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 leverage 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.*

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

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 (CO) layer to build such policies. We build upon recent contributions [3–6] 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 2.1), where $\mathcal{C}(x) = \operatorname{conv} \mathcal{Y}(x)$. Here, $\Omega_{\mathcal{C}(x)} : \operatorname{dom}(\Omega_{\mathcal{C}(x)}) \to \mathbb{R}$ is a proper convex lower-semicontinuous regularization function, with $\mathcal{C}(x) \subseteq \operatorname{cl}\big(\operatorname{dom}(\Omega_{\mathcal{C}(x)})\big)$. 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, typically a neural network parameterized by $w \in \mathcal{W} \subseteq \mathbb{R}^{n_w}$ to predict θ from the context x. 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.

Fenchel-Young losses.

Given the regularized problem (5), the Fenchel-Young loss [3] generated by Ω measures the difference between the solution y of Equation (5) and a target $\bar{y} \in \mathcal{Y}(x)$, as the non-optimality of the target for this problem. Such a loss is typically convex in θ,

nonnegative, and equal to 0 if and only if $p_\Omega(\cdot|\theta)$ is a Dirac in $\bar{y}$. Fenchel duality enables to characterize it as the left-hand side of the Fenchel Young inequality

$$\mathcal{L}_\Omega(\theta; y) := \Omega^*(\theta) + \Omega(y) - \langle \theta | y \rangle. \tag{7}$$

During the last few years, they have become the main approach for supervised training of policies of the form (6) as they lead to a tractable and convex learning problem. Under some hypotheses, they happen to coincide with Bregman divergences, and are a key element in our risk minimization algorithm.

Alternating minimization algorithm.

We rely on the following assumption – which is common in (contextual) stochastic optimization – to derive a tractable alternating minimization algorithm.

Assumption 3. *We suppose having a reasonably efficient algorithm to solve the deterministic single scenario problem* (3).

However, instead of using it online, that is once the context x is known, 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*

$$\mathcal{S}\big(w, y_\otimes\big) = \frac{1}{N}\sum_{i=1}^{N} c(x_i, y_i, \xi_i) + \kappa \mathcal{L}(\theta_i, y_i), \quad \text{with} \quad \begin{cases} \theta_i = \varphi_w(x_i), \\ y_\otimes = (y_i)_{i\in[N]} \in \mathcal{Y}_\otimes, \end{cases} \tag{8}$$

where $[N]$ denotes the set $\{1, \dots, N\}$, $\kappa > 0$ is a positive constant, and $\mathcal{L}$ is a Fenchel-Young loss and

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

Our learning algorithm is an alternating minimization algorithm of the form

$$y_\otimes^{(t+1)} = \operatorname*{argmin}_{y_\otimes} \mathcal{S}(w^{(t)}, y_\otimes), \qquad \text{(Decomposition)}, \tag{9a}$$

$$w^{(t+1)} \in \operatorname*{argmin}_{w\in\mathcal{W}} \mathcal{S}(w, y_\otimes^{(t+1)}), \qquad \text{(Coordination)}, \tag{9b}$$

where $y_\otimes^{(t+1)} = (y_i^{(t+1)})_{i\in[N]}$. In (9a), 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 2. Suffices to say at this point that Step (9a) decomposes per scenario and requires solving deterministic single scenario problems of the form (3), and that the coordination step (9b) amounts to a supervised learning problem with a Fenchel-Young loss, for which efficient algorithms exist.

Related works.

The mirror descent algorithm, introduced by Nemirovsky et al. [7], and analyzed by Beck and Teboulle [8] and Bubeck [9] to name just a few, aims at solving the following problem:

$$\min_{y \in \mathcal{C}} c(y), \tag{10}$$

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 subgradients 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 subgradients, that exploits the geometry of Problem (10). Mirror maps are deeply connected to Legendre-type functions as we discuss in Appendix A. We discuss the connection between mirror descent and our algorithm in Appendix B.

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

$$\min_{y \in \mathcal{C}} \mathbb{E}_{\boldsymbol{\xi}}\big[c(y, \boldsymbol{\xi})\big], \tag{11}$$

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. [10], Zhou et al. [11] for the details, and to Dang and Lan [12] for its block variant. It has many applications, notably in deep learning [13], 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 subgradient of the objective function based on $\xi^{(t)}$.

Our approach relies on the work of Léger and Aubin-Frankowski [14] on alternating minimization algorithms, where a two-variable function $\phi : (\mathcal{Y}, \mathcal{Z}) \mapsto \phi(y, z)$ is iteratively minimized along one of its coordinates while the other is fixed. Starting from an initial point $(y^{(0)}, z^{(0)})$, the algorithm follows the updates

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

The convergence of such algorithms is not a new topic [15], but the wide range of applications in machine learning and signal processing lead to a revived interest in the recent years [16]. Generally, the alternating minimization scheme is studied in a structured context where $\mathcal{Y}$ and $\mathcal{Z}$ are Euclidean and ϕ is convex. In this setting, Beck and Tetruashvili [17] were the first to prove a sublinear rate of convergence when ϕ is assumed L-smooth. In the present case, our function ϕ is neither convex or L-smooth. A more general setting, without any structural assumption on $\mathcal{Y}$, $\mathcal{Z}$ nor ϕ is studied by Léger and Aubin-Frankowski [14]. They proved the convergence of the alternating minimization algorithm when ϕ satisfies the five-point property, first introduced by Csiszár and Tusnády [18], that is a non-local inequality involving ϕ evaluated at different points. We leverage this literature in Section 4 .

Contributions and outline.

Our main contribution is to introduce an alternating minimization algorithm for contextual stochastic combinatorial optimization, 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 where assumptions 1-3 hold. 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 that, in this restricted setting and provided some regularity assumptions, our algorithm converges linearly in value to the optimum of the surrogate. 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 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. [4], 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 [3] 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.

The remainder of the paper is organized as follows. Section 2 provides, without proof, our main results. More technical sections follow, with Section 3 extending regularization and introducing a sparse perturbation on the distribution space to suit our setting, and Section 4 providing some convergence results for the algorithm. Finally, Section 5 details some numerical experiments.

General notations.

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 $s_2 \in S_2$.

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 $\mathrm{dom}(\Psi)$ 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,

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

and by $\partial\Psi$ its subdifferential

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

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.

Moment polytope.

Let $\mathcal{Y}$ be a finite combinatorial set in $\mathbb{R}^d$, and $\mathcal{C} = \mathrm{conv}(\mathcal{Y})$ be its convex hull. As stated above, we assume 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 $H = \mathrm{aff}(\mathcal{Y})$ 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 $H_\Delta = \mathrm{aff}(\Delta^{\mathcal{Y}})$. We denote by V_Δ the vector subspace (hyperplane) in $\mathbb{R}^{\mathcal{Y}}$ that 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 $V_\Delta^\perp = \mathrm{span}(\mathbf{1})$. 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.

2 Main results

We return to the problem of learning structured policies framed in the introduction. In Section 2.1, we start by defining policies π_w, that is to say conditional distributions over combinatorial spaces. 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 2.1 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 $\mathcal{S}_N$ in Section 2.2, and an alternating minimization algorithm to solve

$$\min_{\substack{w\in\mathcal{W},\\ q_\otimes\in\Delta_\otimes}} \mathcal{S}_N\Big(\pi_w, q_\otimes\Big).$$

In Section 2.3, we highlight that the alternating minimization algorithm leads to tractable approximate updates based on decomposition, sampling and stochastic gradient descent. Then, in Section 2.4, we emphasize that provided some technical assumptions, the alternating minimization algorithm converges to the optimum value of the surrogate problem. Last, in a restricted setting and provided some technical assumptions, we can control the difference between the empirical risk of the surrogate optimum and the best empirical risk, as depicted in Section 2.5.

2.1 Policies over combinatorial spaces and empirical risk minimization

We start by a preliminary definition, which is helpful throughout this paper to study our policies.

Legendre-type functions Rockafellar [19, Section 26]

A function $\Psi : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is *Legendre-type* if it is strictly convex on $\mathrm{int}(\mathrm{dom}(\Psi))$ and *essentially smooth*, i.e., (i) $\mathrm{int}(\mathrm{dom}(\Psi))$ is non-empty; (ii) Ψ is differentiable through $\mathrm{int}(\mathrm{dom}(\Psi))$; (iii) $\lim_{i\to\infty} ||\nabla\Psi(\mu_i)|| = +\infty$ 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 $\mathrm{int}(\mathrm{dom}(\Psi))$.

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 $\mathcal{D}^* := \mathrm{int}\big(\mathrm{dom}(\Psi^*)\big)$. Theorem 26.5 of Rockafellar [19] states that, 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 $\nabla\Psi^* = (\nabla\Psi)^{-1}$. As a consequence, the gradient of Legendre-type functions can be used as one-to-one mapping from the primal to the dual space.

Mapping scores to distributions.

Recall that $\Delta^{\mathcal{Y}}$ is the distribution simplex over $\mathcal{Y}$. Let $\Omega_{\Delta^{\mathcal{Y}}}$ be a proper l.s.c. convex function with domain $\Delta^{\mathcal{Y}}$ whose restriction to H_Δ (the affine hull of $\Delta^{\mathcal{Y}}$) is Legendre type. Then

$$\nabla\Omega^*_{\Delta^{\mathcal{Y}}} : s \in \mathbb{R}^{\mathcal{Y}} \mapsto \operatorname*{argmax}_{q\in\Delta^{\mathcal{Y}}} \{s^\top q - \Omega_{\Delta^{\mathcal{Y}}}(q)\}$$

maps any score vector $s \in \mathbb{R}^{\mathcal{Y}}$ to a (unique) probability distribution over $\mathcal{Y}$. We refer to proposition 7 for Fenchel duality results that underpin this definition.

Example 1 (Entropy)**.** *The most classic regularization $\Omega_{\Delta^{\mathcal{Y}}}$ is arguably the negentropy $\Omega_{\Delta^{\mathcal{Y}}} = \sum_{y\in\mathcal{Y}} q_y \log(q_y) + \mathbb{I}_{\Delta^{\mathcal{Y}}}$. In that case, $\nabla\Omega^*_{\Delta^{\mathcal{Y}}}(s)$ maps the score s to its softmax,*

which is the exponential family on $\mathcal{Y}$ parametrized by s

$$\nabla\Omega^*_{\Delta^{\mathcal{Y}}}(s) = \Big(\frac{\exp(s_y)}{\sum_{y'\in\mathcal{Y}}\exp(s_{y'})}\Big)_{y\in\mathcal{Y}}.$$

When $\mathcal{Y}$ is large, probability computations can become challenging and require approximations such as variational inference or Markov Chain Monte-Carlo (MCMC) methods [20].

Example 2 (Sparse perturbation)**.** *In Section 3.3, we extend the work of Berthet et al. [4] and introduce a new regularization function $\Omega_{\Delta^{\mathcal{Y}}}$ that we define as the Fenchel conjugate $\Omega_{\Delta^{\mathcal{Y}}} := F^*_{\varepsilon,\Delta}$ of the function $F_{\varepsilon,\Delta}$ defined over $\mathbb{R}^{\mathcal{Y}}$ as $F_{\varepsilon,\Delta}(s) = \mathbb{E}[\max_{y\in\mathcal{Y}} s_y + \varepsilon\mathbf{Z}^\top y] = \mathbb{E}[\max_{q\in\Delta^{\mathcal{Y}}}(s+\varepsilon Y^\top\mathbf{Z})^\top q]$ where $\mathbf{Z}$ is a random variable, typically a standard Gaussian. This regularization enjoys convenient properties that we describe in Section 3.3. In particular,*

$$\nabla\Omega^*_{\Delta^{\mathcal{Y}}}(s) = \nabla F_{\varepsilon,\Delta}(s) = \mathbb{E}[\operatorname*{argmax}_{q\in\Delta^{\mathcal{Y}}}(s+\varepsilon Y^\top\mathbf{Z})^\top q],$$

enabling to compute stochastic gradients using Monte-Carlo approaches by sampling $\mathbf{Z}$, which is particularly convenient for our algorithms in Section 2.3.

Policies over combinatorial sets.

A policy maps a context value $x \in \mathcal{X}$ to a distribution over the corresponding combinatorial set $q \in \Delta^{\mathcal{Y}(x)}$. To define such policies, we map a context x to the score space, and then map the score to the distribution space as stated above:

$$\pi_w(\cdot|x) = \operatorname*{argmax}_{q\in\Delta^{\mathcal{Y}(x)}}\{\underbrace{\langle Y(x)^\top \overbrace{\varphi_w(x)}^{\theta\in\mathbb{R}^{d(x)}}}_{s\in\mathbb{R}^{|\mathcal{Y}(x)|}}|q\rangle - \Omega_{\Delta^{\mathcal{Y}(x)}}(q)\} = \nabla\Omega^*_{\Delta^{\mathcal{Y}(x)}}\big(Y(x)^\top\varphi_w(x)\big),$$

where we recall that $Y(x) = (y)_{y\in\mathcal{Y}(x)}$ is the wide matrix of solutions in $\mathcal{Y}(x)$, and $\varphi_w : x \in \mathcal{Y} \mapsto \theta \in \mathbb{R}^{d(x)}$ is a machine learning predictor. As illustrated on the left-hand side of Figure 1, an instance x is mapped to a direction vector $\theta = \varphi_w(x)$, then to the score space $s = Y(x)^\top\theta$, and finally to a distribution $q = \nabla\Omega^*_{\Delta^{\mathcal{Y}}(x)}(s)$. The rest of the figure is further detailed in Section 3.2. Such a policy depends on the weights w and the choice of the regularization function $\Omega_{\Delta^{\mathcal{Y}(x)}}$. We do not explicit this second dependency to alleviate notations, since the a single regularization $\Omega_{\Delta^{\mathcal{Y}(x)}}$ is chosen once and for all.

The empirical risk minimization problem.

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. Our goal is to find w values that lead to low empirical risk $R_N(\pi_w)$.

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

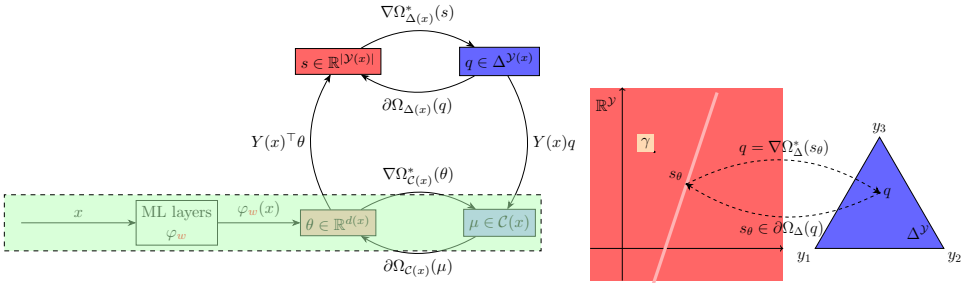

Fig. 1: Policies over combinatorial spaces leveraging Fenchel duality.

With the Ω_Δ previously introduced, the empirical risk is differentiable with respect to w, and stochastic gradients can be computed using score function estimators [21]. We could therefore directly minimize the empirical risk using stochastic gradient descent (SGD). This SGD performs poorly as the score function estimator suffers from a high variance, and R_N is highly non-convex as it is a smoothed piecewise constant function [22]. We therefore follow a different approach based on (less noisy) pathwise estimators for gradients and a convexified problem.

To that purpose, we reformulate the empirical risk minimization as a linear problem on the distribution space. Let $\gamma(x,\xi) = \big(c(x,y,\xi)\big)_{y\in\mathcal{Y}(x)} \in \mathbb{R}^{|\mathcal{Y}(x)|}$ be the score vector induced by the cost function $c(x,\cdot,\xi)$ on the (finite) combinatorial space $\mathcal{Y}(x)$. The right-hand side of Figure 1 highlights that γ (in orange) does not belong to the image of $Y(x)^\top$ (pink line) in general. Given a distribution $q \in \Delta^{\mathcal{Y}}$ on $\mathcal{Y}$, let us define $R_\Delta(q;x,\xi)$ as the expected cost under q. We can recast it as

$$R_\Delta(q;x,\xi) = \mathbb{E}_{\mathbf{y}\sim q}\Big[c(x,\mathbf{y},\xi)\Big] = \langle\gamma(x,\xi)|q\rangle.$$

Using the shortened notation γ_i for $\gamma(x_i,\xi_i)$, we can rewrite the empirical risk as

$$\begin{aligned} R_N(\pi_w) &= \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbf{y}\sim\pi_w(\cdot|x_i)}\Big[c(x_i,\mathbf{y},\xi_i)\Big] \\ &= \frac{1}{N}\sum_{i=1}^N R_\Delta\Big(\underbrace{\nabla\Omega^*_{\Delta^{\mathcal{Y}(x_i)}}\big(\overbrace{Y(x_i)^\top\varphi_w(x_i)}^{s_i}\big)}_{\pi_w(\cdot|x_i)}, w_i, \xi_i\Big) \end{aligned}$$

We rewrite the risk $\mathcal{R}_{\Delta,N}$ as a function of the score $s_\otimes$ on the following product space

$$S_\otimes = \{s_\otimes = (s_i)_{i\in[N]} \,|\, \forall i \in [N], s_i \in \mathbb{R}^{|\mathcal{Y}(x_i)|}\}, \tag{13a}$$

$$\Delta_\otimes = \{q_\otimes = (q_i)_{i\in[N]} \,|\, \forall i \in [N], q_i \in \Delta^{\mathcal{Y}(x_i)}\}, \tag{13b}$$

$$\mathcal{R}_{\Delta,N}(s_\otimes) = \frac{1}{N}\sum_{i=1}^N R_\Delta(\nabla\Omega^*_{\Delta^{\mathcal{Y}(x_i)}}(s_i); x_i, \xi_i) \quad \text{where} \quad s_\otimes \in S_\otimes. \tag{13c}$$

The expression of $R_N(\pi_w)$ above enables to rewrite the empirical risk minimization problem (4) as

$$\min_{w\in\mathcal{W}} R_N(\pi_w) = \min_{w\in\mathcal{W}} \mathcal{R}_{\Delta,N}\Big(\big(Y(x_i)^\top \varphi_w(x_i)\big)_{i\in[N]}\Big). \tag{14}$$

2.2 Surrogate problem and alternating minimization algorithm

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

$$S_{\Omega_\Delta}(s,q;x,\xi) = \langle \gamma(x,\xi)|q\rangle + \kappa\mathcal{L}_{\Omega_{\Delta(x)}}(s;q), \quad \gamma(x,\xi) = \big(c(x,y,\xi)\big)_{y\in\mathcal{Y}(x)}, \tag{15a}$$

$$\mathcal{S}_{\Omega_\Delta,N}(s_\otimes,q_\otimes) = \frac{1}{N}\sum_{i=1}^N S_{\Omega_\Delta}(s_i,q_i;x_i,\xi_i), \quad s_\otimes=(s_i)_{i\in[N]},\ q_\otimes=(q_i)_{i\in[N]}. \tag{15b}$$

Notice that if $S_{\Omega_\Delta}(\cdot,\cdot;x,\xi)$ defined in Equation (15a) takes as inputs $s\in\mathbb{R}^{|\mathcal{Y}(x)|}$ and its dual variable $q=\nabla\Omega^*_{\Delta^{\mathcal{Y}(x)}}(s)$, by Fenchel duality, the Fenchel-Young loss reaches zero, and we recover the expected cost. In general, we have the Fenchel-Young inequality.

$$S_{\Omega_\Delta}(s,\nabla\Omega^*_{\Delta^{\mathcal{Y}(x)}}(s);x,\xi) = R_\Delta\big(\nabla\Omega^*_{\Delta^{\mathcal{Y}(x)}}(s);x,\xi\big), \tag{16a}$$

$$S_{\Omega_\Delta}(s,q;x,\xi) = R_\Delta(q;x,\xi) + \kappa\underbrace{\mathcal{L}_{\Omega_{\Delta(x)}}(s;q)}_{\geq 0}. \tag{16b}$$

This is why we use the term surrogate functions for Equation (15). Our *surrogate learning problem* can then be written as

$$\min_{\substack{w\in\mathcal{W},\\ q_\otimes\in\Delta_\otimes}} \underbrace{\mathcal{S}_{\Omega_\Delta,N}\Big(\big(Y(x_i)^\top\varphi_w(x_i)\big)_{i\in[N]},q_\otimes\Big)}_{=\frac{1}{N}\sum_{i=1}^N S_{\Omega_\Delta}\left(Y(x_i)^\top\varphi_w(x_i),q_i;x_i,\xi_i\right)}. \tag{17}$$

In order to derive solutions to Equation (17), we rely on the following *primal-dual alternating minimization scheme*, initialized by some weights values $\bar{w}^{(0)}$

$$q_\otimes^{(t+1)} = \operatorname*{argmin}_{q_\otimes\in\Delta_\otimes} \mathcal{S}_{\Omega_\Delta,N}\Big(\big(Y(x_i)^\top\varphi_{\bar{w}^{(t)}}(x_i)\big)_{i\in[N]},q_\otimes\Big), \quad \text{(decomposition)}, \tag{18a}$$

$$\bar{w}^{(t+1)} \in \operatorname*{argmin}_{w\in\mathcal{W}} \mathcal{S}_{\Omega_\Delta,N}\Big(\big(Y(x_i)^\top\varphi_w(x_i)\big)_{i\in[N]},q_\otimes^{(t+1)}\Big), \quad \text{(coordination)}. \tag{18b}$$

The primal update in Equation (18a) 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 (18b) is named coordination, because this time a single vector of weights is used, coupling the N terms of the sum in $\mathcal{S}_{\Omega_\Delta,N}$, and coordinating the primal updates. As we discuss below, it corresponds to a simple supervised learning problem with a Fenchel-Young loss.

2.3 Tractable updates for the surrogate problem

We come back to the primal-dual algorithm given in Equation (18), and exploit the structure of the surrogate function $\mathcal{S}_{\Omega_\Delta,N}$ to study the tractability of each step.

Proposition 1. *Using the notations of Section 2.2, Equations* (18a)-(18b) *can be recast as*

$$q_i^{(t+1)} = \operatorname*{argmin}_{q_i \in \Delta^{\mathcal{Y}(x_i)}} S_{\Omega_\Delta}\big(Y(x_i)^\top \varphi_{\bar{w}^{(t)}}(x_i), q_i; x_i, \xi_i\big), \tag{19a}$$

$$= \nabla \Omega^*_{\Delta^{\mathcal{Y}(x_i)}}\big(Y(x_i)^\top \varphi_{\bar{w}^{(t)}}(x_i) - \frac{1}{\kappa}\gamma_i\big), \tag{19b}$$

$$\bar{w}^{(t+1)} \in \operatorname*{argmin}_{w \in \mathcal{W}} \frac{1}{N}\sum_{i=1}^N \mathcal{L}_{\Omega_{\mathcal{C}(x_i)}}\big(\varphi_w(x_i); Y(x_i) q_i^{(t+1)}\big), \tag{19c}$$

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*

$$\Omega_{\mathcal{C}(x)}(\mu) := \min_{q \in \Delta^{\mathcal{Y}(x)}} \{\Omega_{\Delta^{\mathcal{Y}(x)}}(q) \,|\, Y(x)q = \mu\}. \tag{20}$$

The dual update in Equation (19c) 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 $Y(x_i)q_i^{(t+1)}$ 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. 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 (19b) requires working with distributions. Ideally, we would like to do all the computations in the space $\mathcal{Y}(x)$ of small-dimension. We highlight below how we can do so for the negentropy and the sparse perturbation.

The case of negentropy.

The negentropy introduced in Example 1. leads to the exponential family

Proposition 2.

$$\mu_i^{t+1} = \mathbb{E}(\mathbf{y}|Y_i^\top \varphi_{\bar{w}^{(t)}} - \frac{1}{\kappa}\gamma_i) \tag{21a}$$

$$\bar{w}^{(t+1)} = \operatorname*{argmin}_w \sum_{i=1}^N A_{\mathcal{C}}(\varphi_w(x_i)) - \langle \varphi_w(x_i) | \mu_i^{(t+1)} \rangle \tag{21b}$$

While the expectation is in general not tractable, it can be computed for instance using a MCMC approach. To avoid having to design such MCMC, in our numerical experiments, we rely on the sparse perturbation regularization.

The case of sparse perturbation.

The sparse perturbation is introduced in Example 2, and further detailed in Section 3.3. We show how we can leverage its structure to derive tractable updates in the small dimension of the polytope $\mathcal{C}(x)$.

Proposition 3. *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 (35)-(36), the primal-dual updates in Equations (19b)-(19c) become*

$$\mu_i^{(t+1)} = \mathbb{E}_{\mathbf{Z}}\big[\operatorname*{argmin}_{y_i \in \mathcal{Y}(x_i)} c(x_i, y_i, \xi_i) - (\varphi_{\bar{w}^{(t)}}(x_i) + \varepsilon \mathbf{Z})^\top y_i\big], \tag{22a}$$

$$\bar{w}^{(t+1)} \in \operatorname*{argmin}_w \frac{1}{N} \sum_{i=1}^{N} F_{\varepsilon,\mathcal{C}(x_i)}\big(\varphi_w(x_i)\big) - \langle \varphi_w(x_i) | \mu_i^{(t+1)} \rangle, \tag{22b}$$

where, given the primal update in Equation (19b), 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 (22b) 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. [4]. The primal update in Equation (22a) 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 $\mathbf{Z}$. Instead of the distribution variables q_i, our primal-dual algorithm in Equation (22) now only relies on the moment variables μ_i. Although these variables belong to the convex hulls of the combinatorial sets $\mathcal{C}(x_i) = \operatorname{conv}(\mathcal{Y}(x_i))$, the updates in Equation (22) 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.

2.4 Convergence to surrogate problem optimum

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 focus on a special case of our primal-dual algorithm where the dual update minimizes directly on the score vector, instead of the weights w of a machine learning predictor $Y^\top \varphi_w$. Then, we will show that the following special case of Algorithm (18)

$$q_\otimes^{(t+1)} = \operatorname*{argmin}_{q_\otimes \in \Delta_\otimes} \mathcal{S}_{\Omega_\Delta, N}\Big(\bar{s}_\otimes^{(t)}, q_\otimes\Big), \qquad \text{(decomposition)}, \tag{23a}$$

$$\bar{s}_{\otimes}^{(t+1)} \in \operatorname*{argmin}_{s_{\otimes} \in \bar{S}_{\otimes}} \mathcal{S}_{\Omega_\Delta, N}\Big(s_{\otimes}, q_{\otimes}^{(t+1)}\Big), \qquad \text{(coordination)}, \tag{23b}$$

converges with a linear rate toward the global minimum of $\mathcal{S}_{\Omega_\Delta,N}$, provided some regularity assumption on the regularizer Ω_Δ. Detailed proofs of the results presented in this section are postponed to Section 4.

Before presenting the main result, let us introduce the partial minimizer of the surrogate function $\mathcal{S}_{\Omega_\Delta,N}$ introduced in Equation (15b) as

$$\bar{\mathcal{S}}_{\Omega_\Delta,N}(q_{\otimes}) := \min_{s_{\otimes} \in \bar{S}_{\otimes}} \mathcal{S}_{\Omega_\Delta,N}(s_{\otimes}, q_{\otimes}), \tag{24}$$

where $\bar{S}_{\otimes} = \{(s_i)_{i\in[N]} \in (\mathbb{R}^{\mathcal{Y}})^N \,|\, s_1 = \ldots = s_N\}$. This function is a key object to our convergence analysis. Surprisingly, it relates to the Jensen gap of the Ω_Δ according to Lemma 15.

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

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

We refer to this function as the Jensen gap *of Ω_Δ.*

Convergence of the algorithm

We establish convergence rates for (23) in the following theorem.

Theorem 5. *Suppose that the Jensen gap of Ω_Δ is a convex function. Then, the iterates of (23) converge in value toward the global minimum of the surrogate problem $\bar{\mathcal{S}}_{\Omega_\Delta,N}$ over $(\mathbb{R}^{\mathcal{Y}}, \Delta_{\otimes})$ with a rate $\mathcal{O}(\frac{1}{n})$.*

This result relies on the convexity of the Jensen gap of the regularizer Ω_Δ, which is far from straightforward. In Appendix C we settle the issue for a class of separable regularizers encompassing the ℓ_2 norm and negentropy, proving the converge of Algorithm (23) in these cases. The convexity of the Jensen gap when regularizing via random perturbation remains an open question.

In the general case of Algorithm (18), the coordination step is parametrized by a machine learning predictor $Y^\top \varphi_w$ with weights w. The convergence is, in this case, related to the convexity of a "cross"Jensen gap function

$$q_{\otimes} \mapsto \sum_{i=1}^{N} \Omega_\Delta(q_i) - N\Omega_{\mathcal{C}}\big(Y \frac{1}{N}\sum_{j=1}^{N} q_i\big) \tag{25}$$

that involves regularizers in both the moment and distribution spaces. The study of the cross Jensen gap (25) is left for future research.

Relationship to mirror descent

From a geometrical standpoint, our primal-dual algorithm can be interpreted as a mirror descent algorithm applied to a specific function. We show in Appendix B

that, when adding a damping step on the dual update, the primal updates $q_\otimes$ of Algorithm (41) coincide with the ones of mirror descent applied to $\bar{\mathcal{S}}_{\Omega_\Delta,N}$. This result provides an alternative proof of convergence, albeit with stronger assumptions. In particular, it relies on the gradients of $\bar{\mathcal{S}}_{\Omega_\Delta,N}$ being Lipschitz-continuous, which is, at first sight, conflicting with the hypothesizes we make on the regularizer Ω_Δ.

2.5 Empirical risk of the surrogate optimum

We have shown in the previous section that, under some regularity assumptions, the primal dual algorithm converges toward the minimum of the surrogate problem (18). The natural question that comes next concerns the quality of the returned solution for the original problem (4), and thus the pertinence of the surrogate itself. In this section, we exhibit a performance guarantee on the solution returned by the primal-dual algorithm for the empirical risk minimization problem.

We place ourselves in the setting of (non-contextual) stochastic optimization, we can then drop the context from the sets and functions notations. First, let us introduce the other partial minimizer of the surrogate risk as

$$\underline{\mathcal{S}_{\Omega_\Delta,N}}(\theta) := \min_{q_\otimes \in \Delta_\otimes} \mathcal{S}_{\Omega_\Delta,N}\big((Y^\top\theta)_{i\in[N]}, q_\otimes\big), \tag{26}$$

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 6. *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 Ω_Δ is L-strongly convex, which means $\nabla\Omega^*_\Delta$ is $\frac{1}{L}$-Lipschitz-continuous with respect to $||\cdot||$, the absolute difference between the empirical risk and surrogate is bounded as*

$$|\underline{\mathcal{S}_{\Omega_\Delta,N}}(\theta) - \mathcal{R}_N(\theta)| \leq \frac{3}{2NL\kappa}\sum_{i=1}^N ||\gamma_i||^2. \tag{27}$$

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

$$\mathcal{R}_N(\theta_\mathcal{S}) - \mathcal{R}_N(\theta_\mathcal{R}) \leq \frac{3}{L\kappa N}\sum_{i=1}^N ||\gamma_i||^2. \tag{28}$$

The proof of Theorem 6 is postponed to Appendix D. The main assumption of Theorem 6 is the strong convexity of Ω_Δ. This assumption holds, for instance, when regularizing with the ℓ_2-norm or the negentropy.

Remark 1. *The recent work of Aubin-Frankowski et al. [23] studies the extension of generalization bounds to the structured learning paradigm. Our paper falls under the scope of their study and, under mild assumptions, the application of [23, Theorem 3] allows us to bound the (true) risk of the surrogate optimum.*

3 Regularization and sparse perturbation on the distribution space

To study the convergence and tractability of our algorithm, we rely on a regularization on the distribution space $\Delta^{\mathcal{Y}}$. The theory of Fenchel-Young losses has been introduced by the structured learning community [3]. Most applications in this literature are based on linear optimization layers on the polytope $\mathcal{C} = \text{conv}(\mathcal{Y})$. Ranking or top-$k$ applications fall, for instance, within this scope. If some theory covers the case of non-full dimensional $\mathcal{C}$ [3], some important approaches for us have been studied only in the linear and full dimensional case [4]. 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.

Our main contribution in this direction is the introduction of regularization based on a sparse perturbation on that space. An essential and technical contribution for our approach is to make links between the moment $\mathcal{C}$ and the distribution $\Delta^{\mathcal{Y}}$ spaces. In more technical terms, one must consider the case when regularization functions are defined on non-full-dimensional spaces. The reader can thus safely skip Section 3.1 in a first read. Proposition 9 makes a link between the moment and distribution polytopes. Section 3.3 contains our main results on sparse perturbation.

Let $\Omega \in \Gamma_0(\mathbb{R}^d)$ be a proper l.s.c. convex function mapping $\mathbb{R}^d$ to $(-\infty, +\infty]$. In this section, we consider the *regularized prediction* problem defined as

$$\hat{y}_\Omega(\theta) \in \underset{\mu \in \text{dom}(\Omega)}{\text{argmax}} \langle \theta | \mu \rangle - \Omega(\mu), \tag{29}$$

and introduce some new geometric results related to it, which are useful for the study of our algorithm.

3.1 Regularized prediction on non-full-dimensional spaces

To the best of our knowledge, most of the theory of Fenchel-Young losses has been made either introducing a regularization function Ω where $\text{dom}(\Omega) = \mathcal{C}$ is full-dimensional, and Ω is Legendre-type, or using a decomposition $\Omega := \Psi + \mathbb{I}_{\mathcal{C}}$, where Ψ is Legendre-type. In many applications in operations research, we consider polytopes $\mathcal{C}$ that are not full dimensional (take the simplex as a case in point). When defining Ω directly on the polytope (using a perturbation as in Berthet et al. [4] 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 (see Appendix A for the precise definition). The following proposition extend classic results of Legendre-type function to function with non-full dimensional domain $\mathcal{C}$ whose restriction to the affine hull of $\mathcal{C}$ is Legendre-type.

Proposition 7. *Let $\mathcal{C} \subset \mathbb{R}^d$ be a non-empty 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 $\mathbb{R}^d$).*

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 have the following results:

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

$$\nabla\Omega^*(\partial\Omega(y)) = y, \quad \forall y \in \mathrm{rel\,int}(\mathcal{C}). \tag{30}$$

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 $V^\perp$:*

$$\Omega^*(\theta) = \Omega^*(\theta_V) + \langle \theta_{V^\perp} | y_0 \rangle. \tag{31}$$

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

$$\partial\Omega(y) = \partial(\Omega_{|H})(y) + V^\perp, \tag{32}$$

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

$$\partial\Omega(y) = \{\nabla\Omega_{|H}(y)\} + V^\perp. \tag{33}$$

We illustrate Proposition 7 in Figure 2, 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^*. The proof relies on classic convex duality results and provided in Appendix E for completeness.

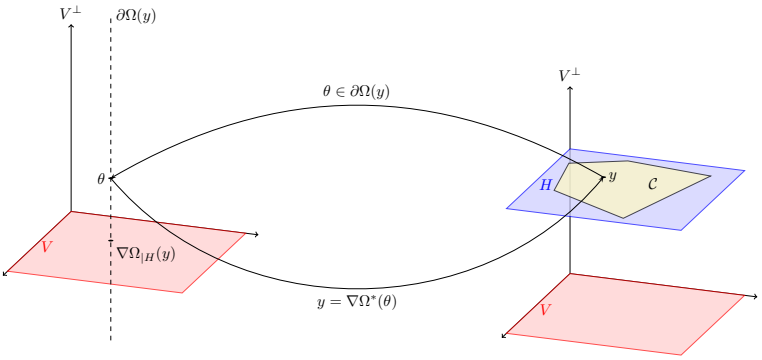

Fig. 2: Primal-dual maps for non-full-dimensional $\mathrm{dom}(\Omega)$.

We now show that any Ω satisfying the assumption of Proposition 7, can be written as $\Omega = \Psi + \mathbb{I}_{\mathcal{C}}$ for some Legendre-type function Ψ.

Proposition 8. *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 $\mathrm{dom}(\Omega) = \mathcal{C}$. We assume*

that the restriction of Ω *to* $H = \mathrm{aff}(\mathcal{C})$ *is Legendre-type (with respect to the metric of* H*). Then, there exists a Legendre-type function* Ψ*, with* $\mathcal{C} \subset \mathrm{cl}\big(\mathrm{int}(\mathrm{dom}(\Psi))\big)$*,* $\mathrm{int}(\mathrm{dom}(\Psi)) \cap \mathcal{C} \neq \emptyset$*, and such that*

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

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 Juditsky et al. [24, Proposition 2.11], where given a $\mathcal{C}$-compatible mirror map (see Appendix A) Ψ, a $\mathcal{C}$-regularizer Ω is defined as $\Omega := \Psi + \mathbb{I}_{\mathcal{C}}$. Indeed, we know with Proposition 7 that Ω defined in the preamble of Proposition 8 is (in particular) a $\mathcal{C}$-regularizer. We use Proposition 8 in Appendix B to study some links between our primal-dual algorithms and mirror descent. The proof is provided in Appendix E.

3.2 Regularizing in the distribution space

Until now in this section, the Fenchel-Young loss has only been introduced in the context of the regularization (see Equation (29)) 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. The proposition below explores regularization on the distribution polytope. We recall that notations are introduced at the end of Section 1.

Proposition 9. *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*

$$\Omega_{\mathcal{C}}(\mu) := \min\{\Omega_\Delta(q) \colon Yq = \mu\}. \tag{34}$$

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

1. $\langle s_\theta|q\rangle = \langle Y^\top\theta|q\rangle = \theta^\top Yq = \langle\theta|Yq\rangle = \theta^\top\mu_q$*.*
2. $\Omega^*_{\mathcal{C}}(\theta) = \Omega^*_\Delta(Y^\top\theta)$*, therefore* $\Omega^*_{\mathcal{C}}$ *has domain* $\mathrm{dom}(\Omega^*_{\mathcal{C}}) = \mathbb{R}^d$*, 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. [3, Section 7.1], but in the case of Ω_Δ being a generalized entropy [25].

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 7.1 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 7.2 to Ω^*_Δ with $V^\perp_\Delta = \operatorname{span}(\mathbf{1})$, 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.

□

3.3 Structured perturbation

We use the notations defined in Section 1 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 $\operatorname{dom}(\Omega) = \mathcal{C}$, and computing the regularized predictions $\hat{y}_\Omega(\theta)$ defined in Equation (29) can be challenging. It may rely on Frank-Wolfe [26] algorithm in practice. We follow another approach pioneered by Berthet et al. [4], defining instead $\Omega^*_\mathcal{C}$ and Ω^*_Δ directly. More precisely, let $\varepsilon \in \mathbb{R}_{++}$, we introduce:

$$F_{\varepsilon,\mathcal{C}}(\theta) = \mathbb{E}[\max_{y \in \mathcal{Y}} (\theta + \varepsilon \mathbf{Z})^\top y] = \mathbb{E}[\max_{y \in \mathcal{C}} (\theta + \varepsilon \mathbf{Z})^\top y], \tag{35}$$

$$F_{\varepsilon,\Delta}(s) = \mathbb{E}[\max_{y \in \mathcal{Y}} s(y) + \varepsilon \mathbf{Z}^\top y] = \mathbb{E}[\max_{q \in \Delta^\mathcal{Y}} (s + \varepsilon Y^\top \mathbf{Z})^\top q], \tag{36}$$

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 (35) is introduced by Berthet et al. [4], while Equation (36) 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. [4, Proposition 2.2] the following properties for $F_{\varepsilon,\mathcal{C}}$ to the case when $\mathcal{C}$ is not full-dimensional.

Proposition 10. *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:*
$$\nabla_\theta F_{\varepsilon,\mathcal{C}}(\theta) = \mathbb{E}[\operatorname*{argmax}_{y \in \mathcal{C}} (\theta + \varepsilon \mathbf{Z})^\top y]. \tag{37}$$
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. [4], $F_{\varepsilon,\mathcal{C}}$ 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 (35) of $F_{\varepsilon,\mathcal{C}}$ is well-defined. The expectation with respect to the Gaussian distribution remains finite, and thus $\operatorname{dom}(F_{\varepsilon,\mathcal{C}}) = \mathbb{R}^d$. 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 $\operatorname{dom}(F_{\varepsilon,\mathcal{C}}) = \mathbb{R}^d$ 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. [4]. 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. [4, Proposition 3.1], we can apply the technique of Abernethy et al. [27, 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 [19, 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,

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

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

Therefore, we have shown that $\mathrm{dom}(F^*_{\varepsilon,\mathcal{C}}) = \mathcal{C}$. 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 $\mathrm{dom}\left((F_{\varepsilon,\mathcal{C}})_{|V}\right) = V$. 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 **??** 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 $\mathrm{Im}(Y^\top)$, which is a subspace of dimension $d' \ll |\mathcal{Y}|$. The proofs of Berthet et al. [4], therefore, do not stand any more. However, perhaps surprisingly, many properties remain valid.

Theorem 11. *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 $V_\Delta^\perp = \mathrm{span}(\mathbf{1})$. 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:*

$$\nabla_s F_{\varepsilon,\Delta}(s) = \mathbb{E}[\operatorname*{argmax}_{q \in \Delta^{\mathcal{Y}}} (s + \varepsilon Y^\top \mathbf{Z})^\top q]. \tag{38}$$

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. [27] cannot be applied to show smoothness. Besides, the proof of strict convexity by Berthet et al. [4] 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 10 point 1. We now show the Lipschitz continuity, and the lower-semicontinuity follows.

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

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

$$\leq \mathbb{E}_{\mathbf{Z}}\big[\max_{q\in\Delta^{\mathcal{Y}}}\langle s_1 - s_2|q\rangle\big] = \max_{q\in\Delta^{\mathcal{Y}}}\langle s_1 - s_2|q\rangle,$$
$$\leq \max_{q\in\Delta^{\mathcal{Y}}}||s_1 - s_2||||q|| = ||s_1 - s_2||.$$

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 provided in Appendix E.

Lemma 12. *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 $(s_1, s_2) \in (V_\Delta)^2$, $s_1 \neq s_2$,

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

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

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

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 10 Point 2.

3. Since we are in the setting given in the preamble of Lemma 12, for $s \in \mathbb{R}^{\mathcal{Y}}$, the argmax in the definition of $F_{\varepsilon,\Delta}$ is reduced to a singleton almost surely:

$$P_{\mathbf{Z}}\big[\big|\operatorname*{argmax}_{y\in\mathcal{Y}}\{s(y) + \varepsilon\mathbf{Z}^\top y\}\big| > 1\big] = 0. \tag{39}$$

Indeed, because the standard multivariate Gaussian distribution has $\mathbb{R}^d$ as support, proving Equation (39) reduces to prove that for any pair of distinct

vectors $(y, y') \in \mathcal{Y}^2, y \neq y'$,

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

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 (39), Danskin's lemma [28, 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 10 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 13. *$\Omega_{\varepsilon,\mathcal{C}}(\mu) = \min_{q:\, Yq=\mu} \Omega_{\varepsilon,\Delta}(q)$, and hence all the properties of Proposition 9 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 and Combettes [29, Corollary 15.28] with $g = F_{\varepsilon,\Delta}$ and $L : x \mapsto Y^\top x$ leads to the result. □

4 Convergence of the alternating minimization scheme

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 focus on a special case of our primal-dual algorithm where the dual update minimizes directly on the score vector, instead of the weights w of a machine learning predictor $Y^\top \varphi_w$. Then, we will show that Algorithm (23), where the dual update minimizes directly on the score vector, converges with a linear rate toward the global minimum of $\mathcal{S}_{\Omega_\Delta,N}$, provided that the Jensen gap of the regularizer Ω_Δ is a convex function. While this result does not imply the convergence of the original algorithm, it provides useful insights on its behavior with respect of the parametrization of the dual vector. The convergence of the algorithm we use in practice would indeed be a very challenging task, as we use various approximation (e.g., Monte-Carlo sampling, see Section 2.3). Furthermore, we typically parametrize the exploration of the space of score via a neural network, whose lack of convenient mathematical properties hinders the development of a generic convergence proof. Let us first recall the five-point property and convergence result of Léger and Aubin-Frankowski [14].

Five-point property [14, 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

$$\phi(y, z_1) + \phi(y_0, z_0) \leq \phi(y, z) + \phi(y, z_0), \tag{40}$$

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 14. *[14, Theorem 2.3] Suppose that ϕ satisfies the five-point property (40) 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 and Aubin-Frankowski [14] 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 (23b). 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.

Before stating the main result of this section, let us first introduce some definitions and lemmas. The partial minimization of our surrogate function $\mathcal{S}_{\Omega_\Delta, N}$ introduced in Equation (15b) is defined as

$$\bar{\mathcal{S}}_{\Omega_\Delta, N}(q_\otimes) := \min_{s_\otimes \in \bar{S}_\otimes} \mathcal{S}_{\Omega_\Delta, N}(s_\otimes, q_\otimes), \tag{41}$$

where $\bar{S}_\otimes = \{(s_i)_{i \in [N]} \in (\mathbb{R}^{\mathcal{Y}})^N \mid s_1 = \ldots = s_N\}$.

Lemma 15. *Let $q_\otimes = (q_i)_{i \in [N]} \in \Delta_\otimes$, the function $\bar{\mathcal{S}}_{\Omega_\Delta, N}$ defined above is equal to*

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

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

$$\bar{\mathcal{S}}_{\Psi_\Delta, N}(q_\otimes) := \frac{1}{N} \sum_{i=1}^N \langle \gamma_i | q_i \rangle + \frac{\kappa}{N} \Big[\sum_{i=1}^N \Psi_\Delta(q_i) - N \Psi_\Delta(\frac{1}{N} \sum_{i=1}^N q_i) \Big]. \tag{42}$$

Note that it is convex if the function $q_\otimes \mapsto \frac{1}{N}\sum_{i=1}^N \Psi_\Delta(q_i) - \Psi_\Delta(\frac{1}{N}\sum_{i=1}^N q_i)$ *is convex. This latter corresponds to the Jensen gap of* Ψ_Δ*. We postpone the convexity study to Appendix C.*

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

$$\bar{\mathcal{S}}_{\Omega_\Delta,N}(q_\otimes) = \min_{s_\otimes \in \bar{S}_\otimes} \mathcal{S}_{\Omega_\Delta,N}(s_\otimes, q_\otimes), \tag{43a}$$

$$= \min_{s\in\mathbb{R}^{\mathcal{Y}}} \frac{1}{N}\sum_{i=1}^N \langle\gamma_i|q_i\rangle + \kappa\big(\Omega_\Delta(q_i) + \Omega_\Delta^*(s) - \langle s|q_i\rangle\big), \tag{43b}$$

$$= \frac{1}{N}\sum_{i=1}^N \langle\gamma_i|q_i\rangle + \frac{\kappa}{N}\Big[\sum_{i=1}^N \Omega_\Delta(q_i) - N\Omega_\Delta(\frac{1}{N}\sum_{i=1}^N q_i)\Big]. \tag{43c}$$

In the computations above, we use the definition (15) of the function $\mathcal{S}_{\Omega_\Delta,N}$, and the fact that Ω_Δ is the Fenchel conjugate of Ω_Δ^* to derive the last line. □

Main result

This allows us to introduce the main theorem of this section.

Theorem 16. *Suppose that the Jensen gap of* Ψ_Δ *is a convex function, with* $\Omega_\Delta = \Psi_\Delta + \mathbb{I}_\Delta$*. Then, the iterates of (23) converge in value toward the global minimum of the surrogate problem* $\bar{\mathcal{S}}_{\Omega_\Delta,N}$ *over* $(\mathbb{R}^{\mathcal{Y}}, \Delta_\otimes)$ *with a rate* $\mathcal{O}(\frac{1}{t})$*.*

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 $\bar{\mathcal{S}}_{\Omega_\Delta,N}$ instead of the original function $\mathcal{S}_{\Omega_\Delta,N}$. Then, we show that this equivalent form holds as a consequence of the convexity of the Jensen gap of Ψ_Δ, which is sufficient to prove the convergence using Theorem 14.

Let $s_\otimes^0 \in \bar{S}_\otimes$, where $\bar{S}_\otimes = \{(s_i)_{i\in[N]} \in (\mathbb{R}^{\mathcal{Y}})^N \,|\, s_1 = \ldots = s_N\}$. First, let us write the five-point property (40) with the notations of the problem, that should hold for all $q_\otimes \in \Delta_\otimes$, and $s_\otimes \in \bar{S}_\otimes$,

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

with $q_\otimes^1 = \operatorname{argmin}_{q_\otimes\in\Delta_\otimes} \mathcal{S}_{\Omega_\Delta,N}(s_\otimes^0, q_\otimes)$ and $s_\otimes^1 \in \operatorname{argmin}_{s_\otimes\in S_\otimes} \mathcal{S}_{\Omega_\Delta,N}(s_\otimes, q_\otimes^1)$. Note that the index numbering is different then in Equation (40) 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 $s_\otimes \in \bar{S}_\otimes$. Hence, we take the minimum in $s_\otimes$ and obtain the equivalent form

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

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

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

On the right-hand side, we observe that the function $\mathcal{S}_{\Omega_\Delta,N}$ is evaluated in four different points, twice in positive and twice in negative. Each of the element $s_\otimes^0$, $s_\otimes^1$, $q_\otimes^1$ 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

$$\bar{\mathcal{S}}_{\Omega_\Delta,N}(q_\otimes) - \bar{\mathcal{S}}_{\Omega_\Delta,N}(q_\otimes^1) \geq \frac{\kappa}{N}\sum_{i=1}^N \big(- \langle s_i^1|q_i\rangle - \langle s_i^0|q_i^1\rangle + \langle s_i^1|q_i^1\rangle + \langle s_i^0|q_i\rangle\big), \tag{44}$$

$$\bar{\mathcal{S}}_{\Omega_\Delta,N}(q_\otimes) - \bar{\mathcal{S}}_{\Omega_\Delta,N}(q_\otimes^1) \geq \frac{\kappa}{N}\big(\langle s_\otimes^1|q_\otimes^1\rangle + \langle s_\otimes^0|q_\otimes\rangle - \langle s_\otimes^1|q_\otimes\rangle - \langle s_\otimes^0|q_\otimes^1\rangle\big). \tag{45}$$

We will now show that the modified five-point property (45) holds in our case. According to Lemma 15, $\bar{\mathcal{S}}_{\Omega_\Delta,N}$ coincides with $\bar{\mathcal{S}}_{\Psi_\Delta,N}$ over $\Delta_\otimes$, which is convex by hypothesis. Hence, for all $(q_\otimes, q_\otimes^1) \in \operatorname{rel\,int}(\Delta_\otimes)^2$ we have

$$\bar{\mathcal{S}}_{\Omega_\Delta,N}(q_\otimes) \geq \bar{\mathcal{S}}_{\Omega_\Delta,N}(q_\otimes^1) + \langle \nabla\bar{\mathcal{S}}_{\Psi_\Delta,N}(q_\otimes^1) \,|\, q_\otimes - q_\otimes^1\rangle. \tag{46}$$

We can compute the gradient as follows

$$\nabla\bar{\mathcal{S}}_{\Psi_\Delta,N}(q_\otimes^1) = \left(\frac{1}{N}\Big(\gamma_i + \kappa\big(\nabla\Psi_\Delta\big(q_i^1\big) - \nabla\Psi_\Delta\big(\frac{1}{N}\sum_{j=1}^N q_j^1\big)\Big)\right)_{i\in[N]} \tag{47}$$

Since $q_i^1 = \nabla\Omega_\Delta^*(\bar{s}^0 - \frac{1}{\kappa}\gamma_i)$ according to Proposition 1, we know using Proposition 8 that there exists a vector $z_i \in V^\perp$ such that

$$\nabla\Psi_\Delta(q_i^1) = s^0 - \frac{1}{\kappa}\gamma_i + z_i.$$

Similarly, we can compute

$$\begin{aligned} s_\otimes^1 &= \operatorname*{argmin}_{s_\otimes\in\bar{S}_\otimes} \mathcal{S}_{\Omega_\Delta,N}(s_\otimes, q_\otimes^1), \\ &= \operatorname*{argmin}_{s_\otimes\in\bar{S}_\otimes} \frac{1}{N}\sum_{i=1}^N \langle\gamma_i \,|\, q_i^1\rangle + \kappa\mathcal{L}_{\Omega_\Delta}(s_i, q_i^1), \\ &= \operatorname*{argmin}_{s_\otimes\in\bar{S}_\otimes} \frac{1}{N}\sum_{i=1}^N \Omega_\Delta^*(s_i) - \langle s_i \,|\, q_i^1\rangle, \end{aligned}$$

$$= \left(\operatorname*{argmin}_{s \in \mathbb{R}^{\mathcal{Y}}} \Omega^*_\Delta(s) - \langle s \,|\, \frac{1}{N} \sum_{j=1}^{N} q^1_j \rangle \right)_{i \in [N]},$$

$$= (s^1)_{i \in [N]}, \quad \text{where} \quad s^1 \in \partial \Omega_\Delta(\frac{1}{N} \sum_{j=1}^{N} q^1_j).$$

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

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

By replacing the terms of Equation (47) by their expression, we get

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

We can now express Equation (46) as the following

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

which is equivalent to the modified five-point property (45). We can then apply Theorem 14, concluding the proof. □

5 Computational experiments

Experiments design.

We would like to highlight the following points. First, using a simple toy problem in Section 5.1 where every computation is straightforward, our aim is to demonstrate the effect of the perturbation scale ε (corresponding to the regularization scale κ). Then, we study the performance of the policy trained with our primal-dual algorithm in Section 5.2. We benchmark against a policy trained via supervised learning, where the dataset contains target solution computed with an expensive Lagrangian heuristic.

5.1 Toy problem

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

Problem statement.

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 $\boldsymbol{\xi} \sim \mathcal{U}(\{\xi_1, \xi_2, \xi_3\})$. 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:

$$\min_{y \in \{0,1\}} \mathbb{E}_{\boldsymbol{\xi} \sim \mathcal{U}(\xi_1, \xi_2, \xi_3)} \big[c(y, \boldsymbol{\xi}) \big]. \tag{48}$$

Stochastic and anticipative solutions.

It is immediate to derive both the anticipative solutions per scenario, and the solution of the stochastic problem (48). 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 (48) is $y = 1$. We see that the most represented solution among the anticipative ones is not the optimal solution of our stochastic problem.

Primal-dual algorithm.

We have access to a training set denoted as $\mathcal{D}_{\text{train}} = (\xi_i)_{i \in [n_{\text{train}}]}$. Since we have no context, we directly learn a parameter θ instead of parameterizing a statistical model φ_w. 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 (48). We do this for different values of the perturbation (or regularization) scale ε, and average over nb_seeds $= 30$ seeds. Figure 3 presents the results, where the curves display the proportions of averaged values of θ leading to the optimal solution. 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.

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 (48). It is natural because the smaller is ε, the closer the primal variables are to anticipative decisions in the first outer iteration. The dual updates then push θ to lead to the majority decision, and the regularization term is too small to deviate from anticipative solutions in the primal updates.

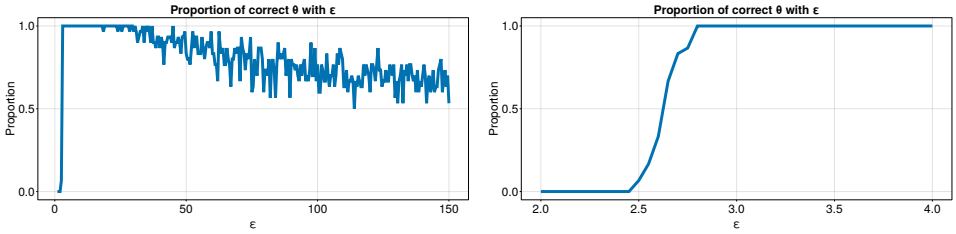

Fig. 3: Proportion of θ values giving the optimal solution of problem (48) when ε varies.

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.*

When ε exceeds a certain threshold, the proportion over seeds of θ leading to the optimal solution increases, and reaches 1, highlighting the crucial role of regularization. However, too large values of ε degrade the performances, as the primal update objectives become dominated by the regularization, leading to random and uninformative solutions. We indeed observe an asymptotic proportion of 0.5 in the left plot of Figure 3.

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.*

5.2 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 (49).

$$\min_{\pi \in \mathcal{H}} \quad \mathbb{E}_{\mathbf{x}}\left[\sum_{e \in E} c_e \pi(\mathbf{x})_e + \mathbb{E}_{\boldsymbol{\xi}}\big[Q\big(\pi(\mathbf{x}); \boldsymbol{\xi}\big)\big]\right], \tag{49a}$$

$$\text{s.t.} \quad \sum_{e \in E(Y)} \pi(x)_e \leq |Y| - 1, \qquad \forall x \in \mathcal{X}, \forall Y, \emptyset \subsetneq Y \subseteq V, \tag{49b}$$

$$\pi(x)_e \in \{0, 1\}, \qquad \forall x \in \mathcal{X}, \forall e \in E. \tag{49c}$$

Notice that constraints (49b)-(49c) 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 (50).

$$Q\big(\pi(x); \xi\big) := \min_{z} \quad \sum_{e \in E} d_e(\xi) z_e, \tag{50a}$$

$$\text{s.t.} \quad \sum_{e \in E} \pi(x)_e + z_e = |V| - 1, \tag{50b}$$

$$\sum_{e \in E(Y)} \pi(x)_e + z_e \leq |Y| - 1, \qquad \forall Y, \emptyset \subsetneq Y \subseteq V, \tag{50c}$$

$$z_e \in \{0, 1\}, \qquad \forall e \in E. \tag{50d}$$

In problem (49), 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 (50) 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 (49c) and (50d), and constraints (50b) and (50c) make sure that we define a tree with the variable $\pi(x) + z$.

Remark 2. *In practice, we can even make the structure of the graph depend on the variable* $\mathbf{x}$*, but we intentionally ease notations above.*

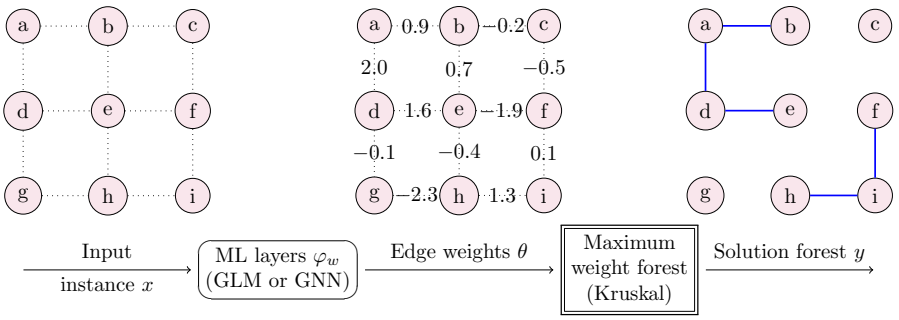

Fig. 4: Two-stage minimum spanning neural network.

For this example, we define a neural network illustrated by Figure 4. 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 (51). A maximum weight forest can be efficiently found using Kruskal's algorithm.

$$\pi_w(x) := \underset{y}{\operatorname{argmax}} \quad \sum_{e \in E} \varphi_w(x)_e y_e, \tag{51a}$$

$$\text{s.t.} \sum_{e \in E(Y)} y_e \leq |Y| - 1, \qquad \forall Y, \emptyset \subsetneq Y \subseteq V, \tag{51b}$$

$$y_e \in \{0, 1\}, \qquad \forall e \in E. \tag{51c}$$

Learning and evaluation.

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. Once the weights w are learned, we can evaluate the resulting policy $\pi_w : x \mapsto f \circ \varphi_w(x)$ for our problem (49) via . 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 (49) 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 (52) where, given a context realization x, we replace the unknown cost vector of the second stage by an estimator of its median.

$$\pi_{\text{median}}(x) := \underset{y}{\operatorname{argmin}} \min_{z} \quad \sum_{e \in E} c_e y_e + \hat{d}_e(x) z_e, \tag{52a}$$

$$\text{s.t.} \quad \sum_{e \in E} y_e + z_e = |V| - 1, \tag{52b}$$

$$\sum_{e \in E(Y)} y_e + z_e \leq |Y| - 1, \qquad \forall Y, \emptyset \subsetneq Y \subseteq V, \tag{52c}$$

$$y_e \in \{0, 1\}, \qquad \forall e \in E, \tag{52d}$$

$$z_e \in \{0, 1\}, \qquad \forall e \in E. \tag{52e}$$

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. [5], Section 6.5, where for each context x_i of our training dataset we have several realizations of the endogenous noise $\boldsymbol{\xi}$. This allows us to derive Lagrangian heuristic-based labels y_i^L to imitate. We leverage the richer training set $\mathcal{D}^L_{\text{train}} = (x_i, y_i^L)_{i \in [n_{\text{train}}]}$ to imitate the solutions of the Lagrangian heuristic, leading to some parameter w^L. We expect the resulting policy π_{w^L} to perform very well, as the Lagrangian heuristic is known to produce very good solutions.

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 4) 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 5. The dotted black line corresponds to the simple 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 π_{w^L} derived by imitating the Lagrangian heuristic. As expected it reaches very good performance, with average gaps smaller than 2%. Then, we 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 (51). 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 5. We observe that using the current weights with policy $\pi_{w^{(t)}}$ leads to small oscillations over outer iterations, while the convergence of policy $\pi_{\bar{w}}$ seems smoother. It can be understood based on the alternating nature of the primal-dual algorithm. Maybe more surprisingly, the limit policy reaches the performance of the Lagrangian heuristic-based policy π_{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.*

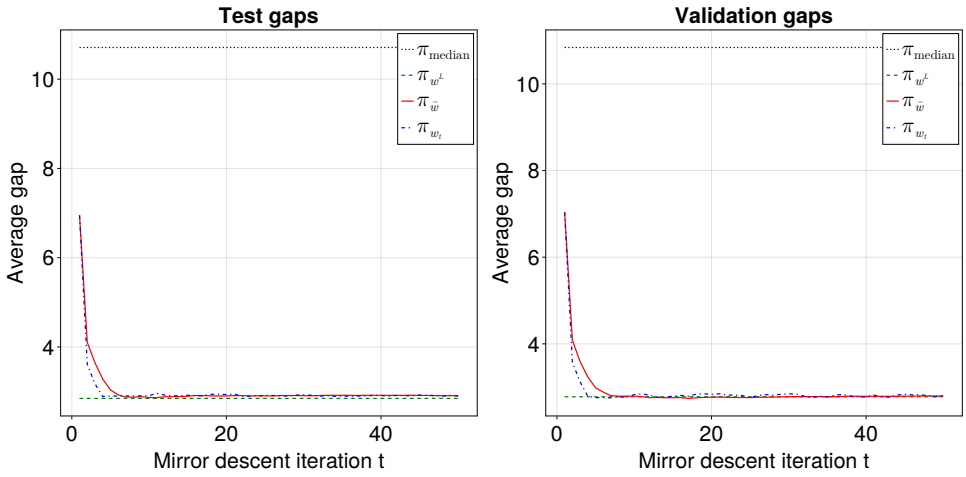

Fig. 5: Validation and test average gaps of policies over the iterations of our primal-dual algorithm, for the two-stage minimum weight spanning tree.

6 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.

A Regularizers, Legendre-type functions and mirror maps

We detail here the definitions of mirror-maps and regularizers, and study their connections with Legendre-type functions.

Mirror maps Juditsky et al. [24, 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 $\text{int}(\text{dom}(\Psi))$,
3. the gradient of Ψ takes all possible values, *i.e.*, $\nabla\Psi\big(\text{int}(\text{dom}(\Psi))\big) = \mathbb{R}^d$.
4. $\mathcal{C} \subset \text{cl}\big(\text{int}(\text{dom}(\Psi))\big)$,
5. $\text{int}(\text{dom}(\Psi)) \cap \mathcal{C} \neq \emptyset$.

Remark 3. *Let $\mathcal{C} \subset \mathbb{R}^d$ be closed convex set, and Ψ be a Legendre-type function such that $\mathcal{C} \subset \text{cl}\big(\text{int}(\text{dom}(\Psi))\big)$, $\text{int}(\text{dom}(\Psi)) \cap \mathcal{C} \neq \emptyset$, and $\text{dom}(\Psi^*) = \mathbb{R}^d$. Then Ψ is a $\mathcal{C}$-compatible mirror map.*

Remark 4. *Sometimes Ψ is called a Bregman potential, and the term of mirror map is used for its gradient $\nabla\Psi$.*

Regularizers Juditsky et al. [24, 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 $\text{cl}(\text{dom}(\Omega)) = \mathcal{C}$. If in addition $\text{dom}(\Omega^*) = \mathbb{R}^d$, then Ω is said to be a *$\mathcal{C}$-regularizer*.

Remark 5. *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.*

Mirror descent 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 introduced above, and considered recently by Juditsky et al. [24] 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 [30, 31].

B Relationship with mirror descent

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 (18) 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 $q_\otimes \in \Delta_\otimes$

$$\Omega_{\Delta_\otimes}(q_\otimes) := \sum_{i=1}^{N} \Omega_\Delta(q_i). \tag{B1}$$

Using Proposition 8, we define a $\Delta^{\mathcal{Y}}$-compatible mirror map Ψ_Δ, such that $\Omega_\Delta = \Psi_\Delta + \mathbb{I}_{\Delta^{\mathcal{Y}}}$. Similarly, for $q_\otimes \in \Delta_\otimes$ in the product distribution space, we define

$$\Psi_\otimes(q_\otimes) := \sum_{i=1}^{N} \Psi_\Delta(q_i). \tag{B2}$$

The following lemma is useful to derive mirror descent below.

Lemma 17. *The function $\Psi_\otimes$ defined in Equation (B2) is a $\Delta_\otimes$-compatible mirror map. Besides, its conjugate $\Psi_\otimes^*$, defined over $(\mathbb{R}^\mathcal{Y})^N$ is such that, for $s_\otimes = (s_i)_{i\in[N]} \in (\mathbb{R}^\mathcal{Y})^N$,*

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

Proof. Extending the properties of the $\Delta^\mathcal{Y}$-compatible mirror map Ψ_Δ 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 $q_\otimes \in \operatorname{rel\,int}(\Delta_\otimes)$, the gradient of $\Psi_\otimes$ is the product of gradients of Ψ_Δ

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

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

$$\Psi_\otimes^*(s_\otimes) = \sup_{q_\otimes\in\operatorname{dom}(\Psi_\otimes)} \{\langle s_\otimes|q_\otimes\rangle - \Psi_\otimes(q_\otimes)\}, \tag{B3a}$$

$$= \sum_{i=1}^N \sup_{q_i\in\operatorname{dom}(\Psi_\Delta)} \Big\{\langle s_i|q_i\rangle - \Psi_\Delta(q_i)\Big\}, \tag{B3b}$$

$$= \sum_{i=1}^N \Psi_\Delta^*(s_i). \tag{B3c}$$

□

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

Theorem 18. *Let $\alpha \in (0,1)$ and $\eta = N\frac{\alpha}{\kappa}$. Then, the following special case of algorithm (18)*

$$q_\otimes^{(t+1)} = \operatorname*{argmin}_{q_\otimes\in\Delta_\otimes} \mathcal{S}_{\Omega_\Delta,N}\Big(\bar{s}_\otimes^{(t)}, q_\otimes\Big), \qquad \textit{(decomposition)}, \tag{B4a}$$

$$\bar{s}_\otimes^{(t+\frac{1}{2})} \in \operatorname*{argmin}_{s_\otimes\in\bar{S}_\otimes} \mathcal{S}_{\Omega_\Delta,N}\Big(s_\otimes, q_\otimes^{(t+1)}\Big), \qquad \textit{(coordination)}, \tag{B4b}$$

$$\bar{s}_\otimes^{(t+1)} = \alpha\bar{s}_\otimes^{(t+\frac{1}{2})} + (1-\alpha)\bar{s}_\otimes^{(t)}, \qquad \textit{(damping)}, \tag{B4c}$$

has the same primal iterates $q_\otimes^{(t)}$ as a mirror descent algorithm applied to the function $\bar{\mathcal{S}}_{\Psi_\Delta,N}$ defined in Section 2.4, using the $\Delta_\otimes$-compatible mirror map $\Psi_\otimes$ and step size η

$$s_\otimes^{(t)} = \nabla\Psi_\otimes(q_\otimes^{(t)}), \tag{B5a}$$

$$q_{\otimes}^{(t+1)} \in \underset{q_{\otimes} \in \operatorname{rel\,int}(\Delta_{\otimes})}{\operatorname{argmin}} B_{\Psi_{\otimes}}\big(q_{\otimes} || \nabla \Psi_{\otimes}^{*}(s_{\otimes}^{(t)} - \eta g^{(t)})\big), \quad \textit{where} \quad g^{(t)} = \nabla \bar{\mathcal{S}}_{\Psi_{\Delta},N}(q_{\otimes}^{(t)}), \tag{B5b}$$

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 $\bar{\mathcal{S}}_{\Psi_{\Delta},N}$ is convex and L-Lipschitz (has bounded gradients) with respect to the same norm $||\cdot||$, the average primal iterates $q_{\otimes}^{(t)}$ in Equation (B4) converge in value to the minimum of $\bar{\mathcal{S}}_{\Psi_{\Delta},N}$ over $\Delta_{\otimes}$, which is also the minimum of the function $\bar{\mathcal{S}}_{\Omega_{\Delta},N}$ by definition.

This theorem shows that, from a geometrical point of view, the primal-dual algorithm can be interpreted as a mirror descent applied to $\bar{\mathcal{S}}_{\Psi_{\Delta},N}$. This result provides an alternate proof of convergence, albeit with stronger assumptions than Theorem 5. In particular, the Lipschitz-continuity of the gradients of $\bar{\mathcal{S}}_{\Psi_{\Delta},N}$ seems, at first sight, challenging to obtain considering the assumptions we made on Ψ_{Δ} in Section **??**.

Proof. The proof is in three steps. First, we recast the mirror descent updates given in Equation (B5). Second, we explicit the primal-dual alternating minimization steps in Equation (B4). Last, we show by induction that the primal updates of the two algorithms coincide.

Mirror descent updates.

The gradient of $\bar{\mathcal{S}}_{\Psi_{\Delta},N}$ at $q_{\otimes} \in \operatorname{rel\,int}(\Delta_{\otimes})$ is

$$\nabla \bar{\mathcal{S}}_{\Psi_{\Delta},N}(q_{\otimes}) = \left(\frac{1}{N} \Big[\gamma_i + \kappa \big[\nabla \Psi_{\Delta}(q_i) - \nabla \Psi_{\Delta}\big(\frac{1}{N} \sum_{j=1}^{N} q_j \big) \big] \Big] \right)_{i \in [N]}. \tag{B6}$$

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

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

Using the definition of $\Psi_{\otimes}$ in Equation (B2),

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

Using in addition the expression of the gradient of $\bar{\mathcal{S}}_{\Psi_{\Delta},N}$ given in Equation (B6),

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

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

$$s_{\otimes}^{(t)} - \eta g^{(t)} = \Big((1-\alpha)\nabla\Psi_\Delta(q_i^{(t)}) + \alpha\Big(-\frac{1}{\kappa}\gamma_i + \nabla\Psi_\Delta\big(\frac{1}{N}\sum_{j=1}^N q_j^{(t)}\big)\Big)\Big)_{i\in[N]}. \tag{B7}$$

We see with Equation (B7) above that we can consider each component $i \in [N]$ of the product variable separately. Using Lemma 17, this property remains true when applying both the gradient of $\Psi_\otimes^*$ and the Bregman divergence generated by $\Psi_\otimes$. Remark now that since $\Omega_\Delta = \Psi_\Delta + \mathbb{I}_{\Delta^\mathcal{Y}}$, we have for $i \in [N]$, by [3, Proposition 3.1]

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

We now define the following iterates, for an additional notation

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

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

$$q_i^{(t+1)} = \nabla\Omega_\Delta^*\Big((1-\alpha)\nabla\Psi_\Delta(q_i^{(t)}) + \alpha\Big(-\frac{1}{\kappa}\gamma_i + \nabla\Psi_\Delta\big(\frac{1}{N}\sum_{j=1}^N q_j^{(t)}\big)\Big)\Big), \quad \forall i \in [N], \tag{B8a}$$

$$\bar{s}^{(t+1)} = \alpha\nabla\Psi_\Delta\big(\frac{1}{N}\sum_{i=1}^N q_i^{(t+1)}\big) + (1-\alpha)\bar{s}^{(t)}. \tag{B8b}$$

Primal-dual alternating minimization updates.

Now, we are going to detail the primal-dual updates of Equation (B4). Let $\tilde{s}_\otimes^{(t)} = (\tilde{s}^{(t)}, \dots, \tilde{s}^{(t)}) \in \bar{S}_\otimes$ be a product dual vector at step t, the decomposition update in Equation (B4a) can be written as

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

$$= \Big(\nabla\Omega_\Delta^*\big(\tilde{s}^{(t)} - \frac{1}{\kappa}\gamma_i\big)\Big)_{i\in[N]}.$$

The minimizer corresponding to Equation (B4b) can be written as

$$\tilde{s}_\otimes^{(t+\frac{1}{2})} \in \operatorname*{argmin}_{s_\otimes \in \bar{S}_\otimes} \mathcal{S}_{\Omega_\Delta, N}\Big(s_\otimes, \tilde{q}_\otimes^{(t+1)}\Big),$$

$$\tilde{s}_\otimes^{(t+\frac{1}{2})} \in \operatorname*{argmin}_{s_\otimes \in \bar{S}_\otimes} \frac{1}{N}\sum_{i=1}^N \langle \gamma_i | \tilde{q}_i^{(t+1)}\rangle + \kappa\big(\Omega_\Delta(\tilde{q}_i^{(t+1)}) + \Omega_\Delta^*(s_i) - \langle s_i | \tilde{q}_i^{(t+1)}\rangle\big),$$

$$\tilde{s}_\otimes^{(t+\frac{1}{2})} = (\tilde{s}^{(t+\frac{1}{2})}, \dots, \tilde{s}^{(t+\frac{1}{2})}), \quad \text{where} \quad \tilde{s}^{(t+\frac{1}{2})} \in \operatorname*{argmin}_{s\in\mathbb{R}^\mathcal{Y}} \Omega_\Delta^*(s) - \langle s | \frac{1}{N}\sum_{i=1}^N \tilde{q}_i^{(t+1)}\rangle,$$

$$\tilde{s}_\otimes^{(t+\frac{1}{2})} = (\tilde{s}^{(t+\frac{1}{2})}, \dots, \tilde{s}^{(t+\frac{1}{2})}), \quad \text{where} \quad \tilde{s}^{(t+\frac{1}{2})} \in \partial\Omega_\Delta\big(\frac{1}{N}\sum_{i=1}^N \tilde{q}_i^{(t+1)}\big).$$

Finally, we reformulate the primal-dual alternating minimization updates of Equation (B4)

$$\tilde{q}_i^{(t+1)} = \nabla\Omega_\Delta^*\big(\tilde{s}^{(t)} - \frac{1}{\kappa}\gamma_i\big), \quad \forall i \in [N], \tag{B9a}$$

$$\tilde{s}^{(t+\frac{1}{2})} \in \partial\Omega_\Delta\big(\frac{1}{N}\sum_{i=1}^N \tilde{q}_i^{(t+1)}\big), \tag{B9b}$$

$$\tilde{s}^{(t+1)} = \alpha\tilde{s}^{(t+\frac{1}{2})} + (1-\alpha)\tilde{s}^{(t)}. \tag{B9c}$$

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 (B8), and the ones of our primal-dual alternating minimization scheme stated in Equation (B9), we show by induction that there exists a sequence of vectors $z^{(t)} \in V_\Delta^\perp$ such that

$$\tilde{s}^{(t)} = \bar{s}^{(t)} + z^{(t)}, \tag{B10a}$$

$$\tilde{q}_i^{(t+1)} = q_i^{(t+1)}, \quad \forall i \in [N]. \tag{B10b}$$

For the initialization, we consider for both sequences

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

Then, we assume that Equation (B10) is satisfied up to step $t \geq 0$. We are going to show it is also satisfied for step $t+1$.

First, using Equation (B9), 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_Δ, there exists a

vector $z^{(t+\frac{1}{2})} \in V_\Delta^\perp$ such that

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

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

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

We have therefore shown the property for the dual iterates. We now focus on the primal updates. Using the mirror descent Equation (B8), for $i \in [N]$, we have

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

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_Δ and the gradient of Ω_Δ^* given in Proposition 8. 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 $\tilde{q}_i^{(t+2)}$. Therefore, we have shown by induction that the primal iterates of mirror descent stated in Equation (B5) and the ones of our primal-dual algorithm in Equation (B4) coincide, whihc concludes the proof.

□

C 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 [3]. 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 19. *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. □

Proposition 20. *(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 $(y_i)_{1\leq i\leq N}$, 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

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

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.

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

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 $2\Psi''(y) \geq 0$ and $det\big(H_h(x, y)\big) \geq 0$. 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 19, we obtain that $det(H_h(x, y)) \geq 0$. Hence h is convex, and its Jensen gap is non-negative, which concludes the proof. □

D Proof of the bound on the surrogate error

In this appendix, we provide a formal proof of Theorem 6 on the surrogate error. We place ourselves in the setting of (non-contextual) stochastic optimization. We therefore have no context, but a collection of noise samples $(\xi_i)_{i\in[N]}$. Let us first recall the theorem.

Theorem 21. *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 Ω_Δ is L-strongly convex, which means $\nabla\Omega_\Delta^*$ is $\frac{1}{L}$-Lipschitz-continuous with respect to $||\cdot||$, the absolute difference between the empirical risk and surrogate is bounded as*

$$|\underline{\mathcal{S}_{\Omega_\Delta,N}}(\theta) - \mathcal{R}_N(\theta)| \leq \frac{3}{2NL\kappa}\sum_{i=1}^N ||\gamma_i||^2. \tag{D11}$$

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

$$\mathcal{R}_N(\theta_\mathcal{S}) - \mathcal{R}_N(\theta_\mathcal{R}) \leq \frac{3}{L\kappa N}\sum_{i=1}^N ||\gamma_i||^2. \tag{D12}$$

Proof. We first give an explicit expression for the partial minimum $\underline{\mathcal{S}_{\Omega_\Delta,N}}$. Then, we bound its absolute difference with the empirical risk $\mathcal{R}_N$. Last, we deduce the bound in Equation (D12).

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

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

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

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

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

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

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

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

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

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 $\theta_\mathcal{S} \in \operatorname{argmin}_\theta \underline{\mathcal{S}_{\Omega_\Delta,N}}(\theta)$ and $\theta_\mathcal{R} \in \operatorname{argmin}_\theta \mathcal{R}_N(\theta)$,

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

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

E Technical proofs on regularization functions

Proof of Proposition 7. 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 [19, 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 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 (32).

   Equation (33) comes from the fact that $\Omega_{|H}$ is Legendre-type, thus differentiable, and for $y \in \operatorname{rel\,int}(\operatorname{dom}(\Omega))$, $\partial(\Omega_{|H})(y) = \{\nabla(\Omega_{|H})(y)\}$.

□

Proof of Proposition 8. 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:

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

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

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

2. By composition with linear projections and sum, Ψ is differentiable over $\mathrm{int}(\mathrm{dom}(\Psi))$. 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 $(y, h) \in \mathrm{int}(\mathrm{dom}(\Psi)) \times \mathbb{R}^d$ be two vectors such that $y + h \in \mathrm{dom}(\Psi)$,

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

   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 $\mathrm{int}(\mathrm{dom}(\Psi))$ is

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

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

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

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

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

Since Π_V is continuous,

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

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 $\theta_{V^\perp}=\theta-\theta_V$,

$$\Psi^*(\theta) = \sup_{y\in\mathbb{R}^d}\{\langle\theta|y\rangle - \Psi(y)\}, \tag{E13a}$$

$$= \sup_{y\in\mathbb{R}^d}\Big\{\langle\theta_V|\Pi_V(y)\rangle - \Omega(\Pi_V(y)) + \langle\theta_{V^\perp}|y-\Pi_V(y)\rangle - \frac{1}{2}||y-\Pi_V(y)||_2^2\Big\}, \tag{E13b}$$

$$= \sup_{\substack{y_V \in V, \\ y_{V^\perp} \in V^\perp}} \Big\{ \langle \theta_V | y_V \rangle - \Omega(y_V) + \langle \theta_{V^\perp} | y_{V^\perp} \rangle - \frac{1}{2} ||y_{V^\perp}||_2^2 \Big\}, \tag{E13c}$$

$$= \Omega^*(\theta_V) + \frac{1}{2} ||\theta_{V^\perp}||_2^2. \tag{E13d}$$

Now, since Ω^* has full domain using Proposition 7, 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 7, Ω^* is indeed differentiable over $\mathbb{R}^d$. Its gradient corresponds to the regularized prediction defined by Equation (29), 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 $\nabla\Psi\big(\nabla\Omega^*(\theta)\big) = \theta + z$. □

Proof of Lemma 12. 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 $V_\Delta = \mathrm{span}(\mathbf{1})^\perp$, 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 $\operatorname{conv}(\mathcal{Y})$ 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.

$$s_2(y_{\tilde{j}}) - s_2(y^*) + \tilde{\lambda}_1 g(y_j), \tag{E14a}$$

$$< s_1(y_{\tilde{j}}) - s_1(y^*) + \tilde{\lambda}_1 g(y_j), \quad \text{(hypothesis right above)}, \tag{E14b}$$

$$= s_1(y_{\tilde{j}-1}) - s_1(y^*) + \tilde{\lambda}_1 g(y_{\tilde{j}-1}), \quad \text{(Both } y_{\tilde{j}} \text{ and } y_{\tilde{j}-1} \text{ are optimal at the junction)}, \tag{E14c}$$

$$= s_2(y_{\tilde{j}-1}) - s_2(y^*) + \tilde{\lambda}_1 g(y_{\tilde{j}-1}), \quad \text{(by definition of } \tilde{j}). \tag{E14d}$$

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 $s_2(y) + \lambda g(y)$. □

Proof of Proposition 1. To derive Equation (19b) remark the following. First, $\mathcal{S}_{\Omega_\Delta,N}$ is defined as the sum of S_{Ω_Δ}, and since $\bar{w}^{(t)}$ is fixed, we obtain N independent problems. Now, using the expression of S_{Ω_Δ}, we see that $q_i^{(t+1)}$ belongs to the minimizers of

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

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

For the dual update in Equation (19c), using the expression of S_{Ω_Δ}, and omitting the term that does not depend on w, we can first recast Equation (18b) as

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

Now, we leverage the computations of Section 3.2. 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 (20). We then use Proposition 9, more precisely Points 1-2, to get Equation (19c). □

Proof of Proposition 3. For the primal update in Equation (22a), we can reformulate Equation (19b) in this perturbation setting

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

In the computations above, we use Proposition 11 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 (22b), we simply use the expression of the Fenchel-Young loss in the perturbation setting, and omit the terms that do not depend on w. □

References

[1] Sadana, U., Chenreddy, A., Delage, E., Forel, A., Frejinger, E., Vidal, T.: A survey of contextual optimization methods for decision-making under uncertainty. European Journal of Operational Research (2024) https://doi.org/10.1016/j.ejor.2024.03.020

[2] Rockafellar, R.T., Wets, R.J.-.: Scenarios and policy aggregation in optimization under uncertainty. Mathematics of operations research **16**(1), 119–147 (1991)

[3] Blondel, M., Martins, A.F.T., Niculae, V.: Learning with Fenchel-Young losses. Journal of Machine Learning Research **21**(35), 1–69 (2020)

[4] Berthet, Q., Blondel, M., Teboul, O., Cuturi, M., Vert, J.-P., Bach, F.: Learning with Differentiable Pertubed Optimizers. In: Advances in Neural Information Processing Systems, vol. 33, pp. 9508–9519. Curran Associates, Inc., ??? (2020)

[5] Dalle, G., Baty, L., Bouvier, L., Parmentier, A.: Learning with Combinatorial Optimization Layers: A Probabilistic Approach. arXiv (2022)

[6] Parmentier, A.: Learning to Approximate Industrial Problems by Operations Research Classic Problems. Operations Research **70**(1), 606–623 (2022) https://doi.org/10.1287/opre.2020.2094

[7] Nemirovsky, A.S., Yudin, D.B., Dawson, E.R.: Wiley-Interscience Series in Discrete Mathematics. John Wiley and Sons, Inc New York (1983)

[8] Beck, A., Teboulle, M.: Mirror descent and nonlinear projected subgradient methods for convex optimization. Operations Research Letters **31**(3), 167–175 (2003) https://doi.org/10.1016/S0167-6377(02)00231-6

[9] Bubeck, S.: Convex Optimization: Algorithms and Complexity. Foundations and Trends® in Machine Learning **8**(3-4), 231–357 (2015) https://doi.org/10.1561/2200000050

[10] D'Orazio, R., Loizou, N., Laradji, I.H., Mitliagkas, I.: Stochastic Mirror Descent: Convergence Analysis and Adaptive Variants via the Mirror Stochastic Polyak Stepsize. Transactions on Machine Learning Research (2023)

[11] Zhou, Z., Mertikopoulos, P., Bambos, N., Boyd, S.P., Glynn, P.W.: On the Convergence of Mirror Descent beyond Stochastic Convex Programming. SIAM Journal on Optimization **30**(1), 687–716 (2020) https://doi.org/10.1137/17M1134925

[12] Dang, C.D., Lan, G.: Stochastic Block Mirror Descent Methods for Nonsmooth and Stochastic Optimization. SIAM Journal on Optimization **25**(2), 856–881 (2015) https://doi.org/10.1137/130936361

[13] Azizan, N., Lale, S., Hassibi, B.: Stochastic Mirror Descent on Overparameterized Nonlinear Models. IEEE Transactions on Neural Networks and Learning Systems **33**(12), 7717–7727 (2022) https://doi.org/10.1109/TNNLS.2021.3087480

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

[15] Auslender, A.: Méthodes numériques pour la décomposition et la minimisation de fonctions non differentiates. Numerische Mathematik **18**, 213–223 (1971/1972)

[16] Wright, S.J.: Coordinate descent algorithms. Mathematical Programming **151**(1), 3–34 (2015) https://doi.org/10.1007/s10107-015-0892-3

[17] Beck, A., Tetruashvili, L.: On the Convergence of Block Coordinate Descent Type Methods. SIAM Journal on Optimization **23**(4), 2037–2060 (2013) https://doi.org/10.1137/120887679

[18] Csiszár, I., Tusnády, G.: Information geometry and alternating minimization procedures. Statistics and Decisions, Dedewicz **1**, 205–237 (1984)

[19] Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970). https://doi.org/10.1515/9781400873173

[20] Wainwright, M.J., Jordan, M.I., *et al.*: Graphical models, exponential families, and variational inference. Foundations and Trends® in Machine Learning **1**(1–2), 1–305 (2008)

[21] Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning **8**(3), 229–256 (1992) https://doi.org/10.1007/BF00992696

[22] Parmentier, A.: Learning structured approximations of combinatorial optimization problems. arXiv preprint arXiv:2107.04323 (2021)

[23] Aubin-Frankowski, P.-C., Castro, Y.D., Parmentier, A., Rudi, A.: Generalization Bounds of Surrogate Policies for Combinatorial Optimization Problems. arXiv. arXiv:2407.17200 [stat] (2024). https://doi.org/10.48550/arXiv.2407.17200 . http://arxiv.org/abs/2407.17200 Accessed 2025-03-03

[24] Juditsky, A., Kwon, J., Moulines, É.: Unifying mirror descent and dual averaging. Mathematical Programming **199**(1), 793–830 (2023) https://doi.org/10.1007/s10107-022-01850-3

[25] Grünwald, P.D., Dawid, A.P.: Game theory, maximum entropy, minimum

discrepancy and robust Bayesian decision theory. The Annals of Statistics **32**(4), 1367–1433 (2004) https://doi.org/10.1214/009053604000000553

[26] Frank, M., Wolfe, P.: An algorithm for quadratic programming. Naval Research Logistics Quarterly **3**(1-2), 95–110 (1956) https://doi.org/10.1002/nav.3800030109

[27] Abernethy, J., Lee, C., Tewari, A.: Perturbation Techniques in Online Learning and Optimization. In: Hazan, T., Papandreou, G., Tarlow, D. (eds.) Perturbations, Optimization, and Statistics, pp. 233–264. The MIT Press, ??? (2016). https://doi.org/10.7551/mitpress/10761.003.0009

[28] Bertsekas, D.P. (ed.): Convex Optimization Theory. Athena Scientific, Belmont, Mass (2009)

[29] Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. CMS Books in Mathematics. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-48311-5

[30] Juditsky, A.B., Nazin, A.V., Tsybakov, A.B., Vayatis, N.: Recursive Aggregation of Estimators by the Mirror Descent Algorithm with Averaging. Problems of Information Transmission **41**(4), 368–384 (2005) https://doi.org/10.1007/s11122-006-0005-2

[31] Nesterov, Y.: Primal-dual subgradient methods for convex problems. Mathematical Programming **120**(1), 221–259 (2009) https://doi.org/10.1007/s10107-007-0149-x