Generalization Bounds of Surrogate Policies for Combinatorial Optimization Problems

Pierre Cyril Aubin-Frankowski[1], Yohann De Castro[2,3], Axel Parmentier[4], and Alessandro Rudi[5]

[1] *TU Wien, Vienna, Austria.*
[2] *Institut Camille Jordan, École Centrale Lyon, CNRS UMR 5208, France.*
[3] *Institut Universitaire de France (IUF)*
[4] *CERMICS, École des Ponts, France.*
[5] *SIERRA, INRIA Paris, France.*

July 24, 2024

Abstract

A recent stream of structured learning approaches has improved the practical state of the art for a range of combinatorial optimization problems with complex objectives encountered in operations research. Such approaches train policies that chain a statistical model with a surrogate combinatorial optimization oracle to map any instance of the problem to a feasible solution. The key idea is to exploit the statistical distribution over instances instead of dealing with instances separately. However learning such policies by risk minimization is challenging because the empirical risk is piecewise constant in the parameters, and few theoretical guarantees have been provided so far. In this article, we investigate methods that smooth the risk by perturbing the policy, which eases optimization and improves generalization. Our main contribution is a generalization bound that controls the perturbation bias, the statistical learning error, and the optimization error. Our analysis relies on the introduction of a uniform weak property, which captures and quantifies the interplay of the statistical model and the surrogate combinatorial optimization oracle. This property holds under mild assumptions on the statistical model, the surrogate optimization, and the instance data distribution. We illustrate the result on a range of applications such as stochastic vehicle scheduling. In particular, such policies are relevant for contextual stochastic optimization and our results cover this case.

1 Introduction

1.1 Learning policies instead of minimizing

Operations Research (OR) aims to optimize industrial processes by allocating resources efficiently. Often, these resources are indivisible, such as a truck that either delivers a request or not. Indivisible ressources lead to combinatorial optimization problems, where larger instances offer more flexibility and economy of scale. Consider for instance a vehicle routing problem: if there are more requests to deliver in the same neighborhood, the distance travelled between two requests is smaller, and the marginal costs per request decreases. However, achieving economy of scale requires algorithms that can handle huge instances. Exploring large combinatorial spaces (with up to 10^{100} solutions) is extremely challenging and most of the research focus has been on scalable algorithms. This scalability has been achieved at the price of simplifying the objective function, a fact reflected in the OR practitioners' tools, where mixed integer linear programming (MILP) dominates. When it comes to making algorithms scale, focusing on MILP has been a success story, with solvers in 2015 being 540 billion times faster than in 1991 as recalled by Bertsimas et al. [2016], and instances with millions of variables and constraints being solved routinely in the industry.

Consider now the following combinatorial optimization problem where we have made the dependency on the instance $\boldsymbol{x}$ in $\mathcal{X}$ explicit

$$\min_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})} f^0(\boldsymbol{y}, \boldsymbol{x}) \tag{1.1}$$

where $\mathcal{Y}(\boldsymbol{x})$ is the finite set of feasible solutions for instance $\boldsymbol{x} \in \mathcal{X}$. The objective f^0 may model any complex phenomena. We assume that f^0 comes with an oracle, that can be mildly expansive to compute, i.e., can take a few seconds to evaluate and for instance be the result of a simulation. We focus on settings where the problem defined by f^0 is challenging in two ways: i) its modeling is too complex to be cast as a mathematical programming formulation that can be solved with off the shelf algorithms; ii) its computational cost prohibits the use of a local descent heuristic. Since combinatorial algorithms must scale to be useful in OR, practitioners generally simplify (1.1) into

$$\min_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})} \tilde{f^0}(\boldsymbol{y}, \boldsymbol{x}) \tag{1.2}$$

where $\tilde{f^0}$ a possibly-over-simplified deterministic objective amenable to classical algorithms. Such algorithms are implicitly supposed to work on any instance of the problem, and distributional aspects over the instances, i.e., that instances that have to be solved in practice have a specific structure, have been overlooked in the literature. This results in worst case analysis to be the main setting considered by complexity theory.

On the other hand, a recent line of works detailed below has proposed with some successes a learning methodology that directly address the objective of (1.1) in its full complexity. Instead of solving each instance independently, we consider a family $\mathcal{H}$ of policies $h : \boldsymbol{x} \in \mathcal{X} \mapsto \boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})$. Our goal is to *learn* a mapping $h^\star$ that minimizes the risk

$$\min_{h \in \mathcal{H}} \mathcal{R}(h) := \mathbb{E}_X\big[f^0(h(X), X)\big], \tag{1.3}$$

where X is a random instance distributed according to some distribution $\mathbb{P}_X$. If the set $\mathcal{H}$ corresponds to all the possible measurable mappings $\boldsymbol{x} \in \mathcal{X} \to (\mathcal{Y}(\boldsymbol{x}))_{\boldsymbol{x} \in \mathcal{X}}$, then we get back to the classical approach where each instance is solved independently.

We instead restrict ourselves to a family of tractable policies parametrized by a vector $\boldsymbol{w}$ belonging to a set $\mathcal{W}$, and we learn the parameter $\boldsymbol{w}$ that minimizes the risk. For such an approach to work, the policy mappings of $\mathcal{H}$ need to be able to properly explore the large combinatorial set of solutions $\mathcal{Y}(\boldsymbol{x})$. One approach has proved successful on a variety of applications such as portfolio optimization [Elmachtoub and Grigas, 2021], path finding in images [Pogancic et al., 2020], vehicle scheduling [Parmentier, 2021a], machine scheduling [Parmentier and T'Kindt, 2021], dynamic vehicle routing [Baty et al., 2024] in the last few years. It consists in delegating the exploration of $\mathcal{Y}(\boldsymbol{x})$ to the linear optimization problem

$$\hat{y}(\boldsymbol{\theta}) \in \arg\max_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})} \langle \boldsymbol{y}, \boldsymbol{\theta} \rangle \,, \tag{1.4}$$

which we assume tractable, and where $\hat{\boldsymbol{y}} \,:\, \theta \in \mathcal{Y}(\boldsymbol{x}) \mapsto \hat{\boldsymbol{y}}(\boldsymbol{\theta}) \in \mathcal{Y}(\boldsymbol{x})$ is an oracle returning an optimal solution of (1.4). We omit the dependency of $\hat{\boldsymbol{y}}$ in $\boldsymbol{x}$ throughout the paper, for sake of readability. In Equation (1.4), vector $\boldsymbol{\theta}$ is a direction in $\mathbb{R}^{d(\boldsymbol{x})}$ and $\langle \cdot, \cdot \rangle$ denotes the standard inner product. A statistical model $\psi_{\boldsymbol{w}}$ is used to find a direction vector $\boldsymbol{\theta}$ that leads to promising solutions. In other words, we consider the following family of parameterized policy mappings

$$\mathcal{H} := \left\{ h_{\boldsymbol{w}} \,:\, \boldsymbol{x} \in \mathcal{X} \mapsto \hat{\boldsymbol{y}}(\psi_{\boldsymbol{w}}(\boldsymbol{x})) \in \mathcal{Y}(\boldsymbol{x}) \,,\ \boldsymbol{w} \in \mathcal{W} \right\}, \tag{1.5}$$

Note that the objective of (1.4) nevertheless relies on an embedding of $\mathcal{Y}(\boldsymbol{x})$ in $\mathbb{R}^{d(\boldsymbol{x})}$, whose dimension $d(\boldsymbol{x})$ may depend on the instance considered.

In a nutshell, instead of approximating the objective function f^0 in (1.1) by a simpler $\tilde{f^0}$ as in (1.2), we study an average error $\mathbb{E}_X\big[f^0(h(X), X)\big]$ as in (1.3). The latter involves a statistical model $\mathcal{H}$ defined in (1.5) and derived from the linear combinatorial optimization problem (1.4). The role of the model is to explore efficiently the solution set $\mathcal{Y}(\boldsymbol{x})$ and we do not seek to approximate f^0 in any way. In practice, as usual in

machine learning, it is the distribution $\mathbb{P}_X$ in (1.3) that is approximated through samples $(X_1, \dots, X_n)$, in an empirical risk minimization fashion. Our Theorem 3 provides error guarantees when n goes to $+\infty$ to show the convergence to (1.3).

The idea of transforming an optimization problem into a learning one is motivated by industrial practice. In many real-world scenarios, the problem instances that practitioners in the industry need to solve come from a specific but unknown distribution. For example, an airline optimizes the routes of its airplanes based on a schedule and some operational constraints that do not vary much from one day to another. The traditional approach does not take advantage of this distributional information. Learning is a sensible strategy as it allows the algorithm to adapt to the problem structure. Moreover, it enables to focus on optimizing the true objective f^0 instead of pushing the optimization of (1.2) beyond the statistical accuracy of $\tilde{f}$. Furthermore, the choice of $\mathcal{H}$ allows the combinatorial aspects to be handled only through the linear optimization oracle. This enables the use of very efficient algorithms, developed by the combinatorial optimization community for problems with linear objectives. Finally, most of the computational burden is pushed into the learning problem, which is solved offline: only the linear surrogate (1.4) needs to be solved online. This is helpful in industrial settings, where the full information on the instance is often known at the last minute, which strongly constrains the computing time and forces to use only objective functions simple enough to allow for fast optimization algorithms.

1.2 Learning problem

Finding a suitable policy in $\mathcal{H}$ requires designing a learning algorithm that selects a parameter $\boldsymbol{w}$ that leads to a good performance in practice. Two typologies of learning strategies have been proposed in the literature: supervised learning and regret minimization, a.k.a. imitation learning and learning by experience [Bengio et al., 2021]. Supervised learning approaches assume that the training set contains instances $X_1, \dots, X_n$ of (1.1) and their solutions $\bar{\boldsymbol{y}}_1, \dots, \bar{\boldsymbol{y}}_n$. In this setting, the objective f^0 is dismissed, and surrogate losses evaluate how far the model output $\hat{\boldsymbol{y}}(\psi_{\boldsymbol{w}}(\boldsymbol{x})) \in \mathcal{Y}(\boldsymbol{x})$ of the model is from the training set target $\bar{\boldsymbol{y}}_i$. Regret minimization approaches instead only require instances $X_1, \dots, X_n$ and being able to evaluate f^0, by focusing on minimizing the empirical regret

$$\min_{\boldsymbol{w} \in \mathcal{W}} \mathcal{R}_n(h_{\boldsymbol{w}}), \quad \text{where} \quad \mathcal{R}_n(h_{\boldsymbol{w}}) = \frac{1}{n} \sum_{i=1}^{n} f^0\big(h_{\boldsymbol{w}}(X_i), X_i\big). \tag{1.6}$$

Unfortunately, due to the combinatorial nature of (1.4), this objective is piecewise constant, which has two negative impacts: it makes the optimization problem intractable, and leads to poor generalization properties. To that purpose, we rely on a perturbation strategy developed in Berthet et al. [2020] to turn the linear optimization problem (1.4) into a family of tractable distributions $p_\lambda(\boldsymbol{y}|\boldsymbol{\theta})$, with $\lambda > 0$ a parameter controlling the smoothness of $\boldsymbol{\theta} \mapsto p_\lambda(\boldsymbol{y}|\boldsymbol{\theta})$. And we formulate the learning problem as

$$\min_{\boldsymbol{w} \in \mathcal{W}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \sum_{\boldsymbol{y} \in \mathcal{Y}(X_i)} p_\lambda(\boldsymbol{y}|\psi_{\boldsymbol{w}}(X_i))\, f^0\big(\boldsymbol{y}, X_i\big) \right\}. \tag{1.7}$$

1.3 Related works

Relevant applications A first natural application of our approach is black-box optimization for combinatorial problems, see for instance Vu et al. [2017]. When modeling complex phenomena, the cost function oracle $f^0(\boldsymbol{y}, \boldsymbol{x})$ typically involves running a simulation, which can be computationally expensive. When the set of feasible solutions $\mathcal{Y}(\boldsymbol{x})$ is large, it is intractable to directly optimize the objective function f^0 in (1.1). It is then natural to replace the black-box f^0 by a tractable surrogate in a parametric family. As a variant, f^0 may be sufficiently simple to be optimized directly, but the algorithm solving (1.1) is too slow or does not scale to large instances. In that setting, using a surrogate f can be a way to obtain a fast heuristic for large instances. Parmentier

and T'Kindt [2021] obtained the best fast heuristic for large instances of a classic single machine scheduling problem using such an approach. The need for such algorithms is notably important in the context of stochastic optimization [Shapiro et al., 2021], where the objective $f^0(\boldsymbol{y}, \boldsymbol{x}) = \mathbb{E}_\xi g(\boldsymbol{y}, \xi, \boldsymbol{x})$ involves an expectation with respect to an uncertain parameter. There exists plenty of algorithms to solve (1.1), but these are typically able to solve only instances of moderate size, while algorithms for deterministic versions can address large industrial instances. Such techniques have been applied to a stochastic vehicle scheduling problem [Parmentier, 2021a]. They can also be used to approximate of two stage stochastic problems by single stage deterministic surrogates [Dalle et al., 2022, Parmentier, 2021b]. In that context, $f^0(\boldsymbol{y}, \boldsymbol{x}) = \mathbb{E}_{\xi,z} g(\boldsymbol{y}, \boldsymbol{z}, \xi, \boldsymbol{x})$, where ξ is again an uncertainty, $\boldsymbol{y}$ is a decision taken before knowing ξ, and $\boldsymbol{z}$ is a decision taken after observing ξ. The surrogate $\psi_{\boldsymbol{w}}(\boldsymbol{x})^\top \boldsymbol{y}$ is then a deterministic approximation of the function $\min_z \mathbb{E}_{\xi,z} g(\boldsymbol{y}, \boldsymbol{z}, \xi, \boldsymbol{x})$. Moreover, these methods are particularly well suited for contextual stochastic optimization problems [Sadana et al., 2024], see Example 1 below. Finally, one context where such surrogates are very useful but which is beyond the scope of this paper is multistage stochastic optimization [Shapiro et al., 2021] for combinatorial problems, as illustrated by the winning approach for the 2022 EURO-NeurIPS challenge on dynamic vehicle routing [EUR, Baty et al., 2024].

To illustrate our theoretical guarantees, we present three examples, one generic and two more industrial.

Example 1. *(Contextual Stochastic Optimization) Consider a stochastic optimization problem where the cost over a solution is influenced by some noise ξ, [Sadana et al., 2024]. The decision maker does not observe ξ, but has access to some contextual data $\tilde{X}$ that is correlated to ξ. A solution is a policy h that maps any context $\tilde{\boldsymbol{x}}$ to a feasible decision $\boldsymbol{y} \in \mathcal{Y}(\tilde{\boldsymbol{x}})$ given $\tilde{\boldsymbol{x}}$. Given a joint distribution over $(\tilde{X}, \xi)$ and a hypothesis class $\mathcal{H}$ for h, the goal of contextual stochastic optimization is to find the policy that minimizes the expected risk*

$$\min_{h \in \mathcal{H}} \mathbb{E}\Big[f^c\big(h(\tilde{X}), \tilde{X}, \xi\big) \Big] \tag{1.8}$$

An optimal policy is provided by the conditional distribution

$$h^\star : \tilde{\boldsymbol{x}} \longmapsto \operatorname*{arg\,min}_{\boldsymbol{y} \in \mathcal{Y}(\tilde{\boldsymbol{x}})} \mathbb{E}\Big[c\big(\delta(\tilde{X}), \tilde{X}, \xi\big) \Big| \tilde{X} = \tilde{\boldsymbol{x}} \Big] \tag{1.9}$$

However, contextual stochastic optimization considers the setting where the joint distribution over $(\tilde{X}, \xi)$ is unknown, but we have access to a training set $\tilde{\boldsymbol{x}}_1, \boldsymbol{\xi}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{\xi}_n$. This makes the optimal policy (1.9) impractical. Sadana et al. [2024] survey this emergent field. Contextual stochastic optimization can be considered as a special case of our setting. It suffices to define X as $(\tilde{X}, \xi)$, and $f^0(\boldsymbol{y}, \boldsymbol{x})$ as $f^c(\boldsymbol{y}, \tilde{\boldsymbol{x}}, \boldsymbol{\xi})$ for $\boldsymbol{x} = (\tilde{\boldsymbol{x}}, \boldsymbol{\xi})$. We then restrict ourselves to statistical models $\psi_{\boldsymbol{w}}(\boldsymbol{x})$ of the form $\tilde{\psi}_{\boldsymbol{w}}(\tilde{\boldsymbol{x}})$ that exploit only the contextual information and not the noise.

Example 2. *(Stochastic Vehicle Scheduling [Parmentier, 2021a]) Consider a set of tasks with fixed begin and end times operated by a fleet of vehicles. When a task is late, the assigned vehicle may propagate the delay to the next tasks. The goal is to build the sequence of tasks operated by each vehicle in such a way that each task is operated by exactly one vehicle and the expected cost of delay $f^0(\boldsymbol{y}, \boldsymbol{x})$ is minimized. The instance $\boldsymbol{x}$ can typically be encoded as a digraph $D = (V, A)$ where V is the set of tasks, and there is an arc a in A between v and v' if these two tasks can be operated by the same vehicle in a sequence, and a vector $\boldsymbol{\phi}(a, \boldsymbol{x})$ of additional information for each arc a. Since a solution corresponds to a partition of the vertices of an acyclic digraph into paths, the set of feasible solutions $\mathcal{Y}(\boldsymbol{x})$ can be identified with the vertices of the flow polytope over D. As statistical model, one can take $\psi_{\boldsymbol{w}} : \boldsymbol{x} \mapsto \boldsymbol{\theta} = (\theta_a)_{a \in A}$ with $\boldsymbol{\theta}_a = \boldsymbol{w}^\top \boldsymbol{\phi}(a, \boldsymbol{x})$ for each a. Aircraft routing problems solved routinely by airlines are variants of this problem. The stochastic version of it aims for its part at controlling delay propagation. Solvers used in production typically solve the linear oracle. Switching from the deterministic and delay-agnostic approaches in production to our policy that avoids delay propagation does not impact the algorithm used online in production. It amounts to changing the parametrization of the online solver. Although, computing this new parametrization requires solving a computationally intensive learning problem, this can be done offline.*

Example 3. *(Single Machine Scheduling [Parmentier and T'Kindt, 2021]) In single machine scheduling problems, n jobs have to be performed on a given machine. A solution is a permutation of* $\{1, \dots, n\}$ *giving the order in which the jobs are processed. Several objectives can be used. In the reference above, jobs have release times, before which they cannot be started, and processing times, which indicate the time needed by the machine to operate the job. The vector* $\boldsymbol{y}$ *encodes a permutation, and the goal here is to minimize* $f^0(\boldsymbol{y}, \boldsymbol{x})$ *which is equal to the sum (the average) of the job completion times, taking into account the release time. Permutations* $\boldsymbol{y}$ *can be encoded by vectors* $\boldsymbol{y}_i$ *giving the position of job i in the permutation. Optimizing a linear objective* $\boldsymbol{\theta}^\top \boldsymbol{y}$ *amounts to sorting* $\boldsymbol{\theta}$*, and we can therefore use a sorting algorithm as linear minimization oracle.*

Learning algorithms and generalization bounds. In order to train policies that embed a combinatorial optimization oracle using a supervised learning approach, recent advances have focused on Fenchel-Young losses [Berthet et al., 2020, Blondel et al., 2020, Dalle et al., 2022]. In contrast, the regret minimization paradigm has been overlooked [Parmentier, 2021b]. As far as we know, the only well-established for regret is the "smart predict then optimize" [Elmachtoub and Grigas, 2021], which relies on a surrogate Structured Hinge Loss [Nowozin and Lampert, 2011, Chapter 6]. It applies only to the case of an f^0 that is linear but noisy. In this paper, we focus on the theoretical analysis of the regret minimization approach (1.6). One natural estimator consists in solving the empirical learning problem. To the best of our knowledge, generalization bounds for structured regret minimization have been studied in Parmentier [2021b], which bounds the empirical process error on the regularized risk when parameters $\boldsymbol{w}$ are restricted to some ℓ_∞ ball. A similar result has been shown in Qi et al. [2021] which also used Rademacher complexity to derive the statistical excess risk in contextual stochastic optimization.

Our problem could also be seen as a special case of data-driven algorithm design, in a setting introduced by Gupta and Roughgarden [2016]. We refer to Balcan [2021] for a summary of the literature on this topic. Given a training set of instances of a problem coming from an unknown distribution, a regret minimization approach is used to select the hyperparameters. Balcan et al. [2021] consider generalization guarantees for the optimum of piecewise decomposable risk such as our piecewise constant non-regularized risk (1.6). This framework could apply in our context when ψ_w is linear, but we do not have an algorithm to optimize such a piecewise constant risk. Since the regularized risk is smooth and therefore no more piecewise decomposable, their result cannot be applied in our context. The counterpart of their result is our Theorem 3 and relies on tools more appropriate to our smoothed context.

Such questions of generalization are also studied in the structured learning literature [Ciliberto et al., 2016, 2020], where the problem consist in finding the best function h_w minimizing $\mathbb{E}_{X,Y} f^0(h_w(X), Y)$ for some probability distribution over (X, Y), given examples sampled from the same probability. The results are very general, holding under mild assumptions on f^0, and have been specialized to discrete spaces, in Nowak-Vila et al. [2019]. In particular, the results in the latter paper allow deriving automatically functions h_w with a method that avoids the curse of dimensionality in the dimension of the discrete space. However, it is unclear how to generalize this technique to an arbitrary f^0 since the result depends on some internal properties of such function that are not easy to check in a general case.

1.4 Contributions

In this paper, we build upon the normal fan geometry (see Section 2) involved by the linear optimization program (1.4) and upon a perturbation approach to deliver a generalization bound that controls the perturbation bias, the statistical learning error, and the optimization error (Theorem 1).

Perturbation makes the learning problem smooth and amenable to gradient descent methods. We choose to study the so-called kernel Sum-of-Squares method [Rudi et al., 2024] in this paper. Furthermore, we prove non-asymptotic guarantees on the performance of the learned policy, as well as convergence rates toward the best policy encoded by (1.5).

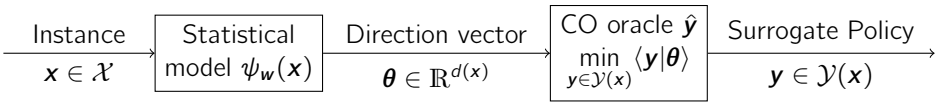

Figure 1: Surrogate policy encoded by the statistical model $\psi_w : x \in \mathcal{X} \mapsto \theta \in \mathbb{R}^{d(x)}$ with combinatorial optimization (CO) layer given by a linear program over solutions $y \in \mathcal{Y}$.

To obtain our performance guarantees, we introduce the uniform weak (UW) moment property. It corresponds to controlling uniformly over w the moments of the distance of $\psi_w(X)$ and the boundaries of the cones of the normal fan of $\mathcal{Y}(x)$. Hence, it quantifies the interplay of the statistical model and (1.4). From this property, we prove an explicit control of the variation of the risk when varying the scale $\lambda > 0$ of the perturbation. Furthermore, we show that the UW property holds under mild assumptions on the statistical model ψ_w on and on $\mathbb{P}_X$. We illustrate this on several examples in Section 3.2.

1.5 Outline

The next section (Section 2) presents our surrogate policy model (Section 2.1), the different kind of risks (Section 2.2) and the main theorem on generalization guarantees (Section 2.3). Section 3 proves the main theorems. It starts with a presentation the conditions required by our analysis (Section 3.1) and a discussion on the UW property (Section 3.2). In Section 3.3 we show a control on the regularized risk under UW property. Section 3.4 focuses on the discrepancy between the risk and its empirical version. Finally, we provide in Section 3.5 the optimization error bound when using kernel Sum-of-Squares.

1.6 Notation

We denote by $\mathcal{L}(U)$ the law of some random variable U and by $\mathbb{P}_U$ (resp. $\mathbb{E}_U$) the probability (resp. expectation) with respect to the law $\mathcal{L}(U)$. We denote the stochastic boundedness by $\mathcal{O}_{\mathbb{P}}$: for a sequence of random variables $(U_n)_n$, notation $U_n = \mathcal{O}_{\mathbb{P}}(1)$ means that for any $\varepsilon > 0$, there exists a finite $M > 0$ and a finite $m \geq 0$ such that for every $n \geq m$, $\mathbb{P}(|U_n| > M) \leq \varepsilon$. For a sequence of non-zero reals $(a_n)_n$, $U_n = \mathcal{O}_{\mathbb{P}}(a_n)$ denotes that $(U_n/a_n) = \mathcal{O}_{\mathbb{P}}(1)$.

The set of instances is denoted by $\mathcal{X}$, and the set of feasible solutions for an instance $x \in \mathcal{X}$ by $\mathcal{Y}(x) \subseteq \mathbb{R}^{d(x)}$. Given an instance $x \in \mathcal{X}$, the objective function is $f^0(\cdot, x)$ whose optimal solution is denoted by $y^0(x)$ or simply y^0. The oscillation of f^0 is defined by

$$\operatorname{osc}(f^0) := \sup_{y,x} \left\{ f^0(y, x) \right\} - \inf_{y,x} \left\{ f^0(y, x) \right\}.$$

The set of parameters is denoted by $\mathcal{W}$ and the statistical model by $\psi_w : x \in \mathcal{X} \mapsto \theta \in \mathbb{R}^{d(x)}$. The solution given by parameter $w \in \mathcal{W}$ is given by $\hat{y}(\psi_w(x))$.

2 The surrogate policy model and its guarantees

2.1 Surrogate policy model

Figure 1 illustrates the policies we consider. We start with a statistical model $x \mapsto \psi_w(x) \in \mathbb{R}^{d(x)}$ that embeds an instance $x \in \mathcal{X}$ into $\mathbb{R}^{d(x)}$ and we set $\theta := \psi_w(x) \in \mathbb{R}^{d(x)}$. We build our surrogate policy $\hat{y}(\psi_w(x)) \in \mathcal{Y}(x)$ by considering a solution to the linear program

$$\max_{y \in \mathcal{C}(x)} \langle y, \theta \rangle \text{ where } \theta := \psi_w(x). \tag{2.1}$$

We minimize over the convex hull

$$\mathcal{C}(\boldsymbol{x}) := \text{conv}(\mathcal{Y}(\boldsymbol{x})).$$

which is a polytope since $\mathcal{Y}(\boldsymbol{x})$ is finite. This surrogate optimization problem is tractable in practice. It is given by the family of surrogate maps $f(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{w}) = \langle \boldsymbol{y}, \psi_{\boldsymbol{w}}(\boldsymbol{x}) \rangle$ linear in $\boldsymbol{\theta} = \psi_{\boldsymbol{w}}(\boldsymbol{x})$.

By construction, the surrogate policy $\hat{\boldsymbol{y}}(\boldsymbol{\theta})$ is an extreme point of $\mathcal{C}(\boldsymbol{x})$. For each $\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})$ one can define the normal cone $\mathcal{F}_{\boldsymbol{y}}$ as the set of directions $\boldsymbol{\theta}$ such that $\boldsymbol{y}$ is a solution to (2.1), equivalently

$$\mathcal{F}_{\boldsymbol{y}} := \left\{ \boldsymbol{\theta} \in \mathbb{R}^{d(\boldsymbol{x})} \;\middle|\; \langle \boldsymbol{y}, \boldsymbol{\theta} \rangle \geq \langle \boldsymbol{y}', \boldsymbol{\theta} \rangle \text{ for all } \boldsymbol{y}' \in \mathcal{C}(\boldsymbol{x}) \right\}$$

The normal fan $\mathcal{N}(\mathcal{C}(\boldsymbol{x}))$ is the collection of all such cones, illustrated on Figure 2.

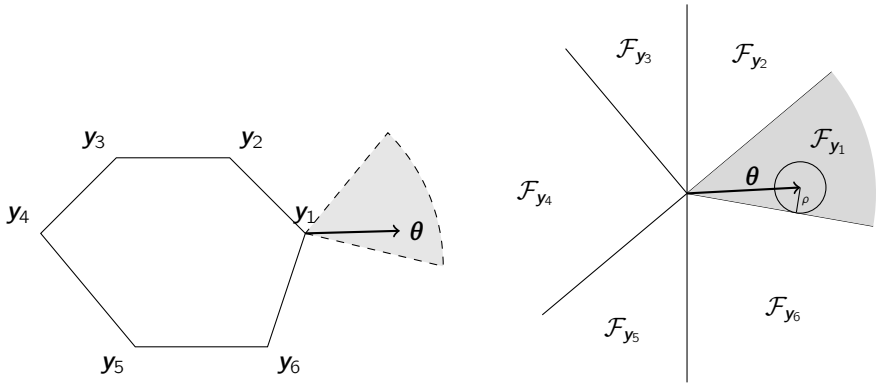

Figure 2: Normal cone at point $\boldsymbol{y}_1$ to the polytope (left) and normal fan with internal radius ρ at point $\boldsymbol{\theta}$ (right).

Example 2 (continued). *In the context of stochastic vehicle scheduling, $\mathcal{Y}(\boldsymbol{x})$ corresponds to the partitions of the vertices V into paths. Solutions are encoded by indicator vectors $\boldsymbol{y} = (y_a)_{a \in A}$ where y_a is a binary variable equal to 1 if arc a is in the solution, and to 0 otherwise. Its convex hull, $\mathcal{C}(\boldsymbol{x})$, is known to be a flow polytope and has a compact formulation.*

Example 3 (continued). *The feasible solutions $\mathcal{Y}(\boldsymbol{x})$ of single machine scheduling are the vectors $\boldsymbol{y}$ encoding a permutation of $\{1, \dots, n\}$. Their convex hull $\mathcal{C}(\boldsymbol{x})$ is a polytope known as the permutahedron.*

For a given $\boldsymbol{x} \in \mathcal{X}$, it can occur that several $\boldsymbol{y} \in \mathcal{Y}$ are optimal, *i.e.*, there exists several $\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})$ such that $\psi_{\boldsymbol{w}}(\boldsymbol{x}) \in \mathcal{F}_{\boldsymbol{y}}$. In this case, $\hat{\boldsymbol{y}}(\psi_{\boldsymbol{w}}(\boldsymbol{x}))$ is defined as a probability distribution on such $\boldsymbol{y}$'s. A formal definition of the surrogate policy $\hat{\boldsymbol{y}}(\psi_{\boldsymbol{w}}(\boldsymbol{x}))$ is given as follows.

Definition 1 (Surrogate policy)**.** *Let $\boldsymbol{\theta} \in \mathbb{R}^{d(\boldsymbol{x})}$, and define the internal cone radius as*

$$\rho(\boldsymbol{\theta}) := \sup \left\{ r > 0 \;\middle|\; \forall \boldsymbol{u} \in B(0,1), \exists \boldsymbol{y} \in \mathcal{Y} \text{ s.t. } \boldsymbol{\theta} \in \mathcal{F}_y \text{ and } \boldsymbol{\theta} + r\boldsymbol{u} \in \mathcal{F}_y \right\},$$

where $B(0,1) \subset \mathbb{R}^{d(\boldsymbol{x})}$ is the standard Euclidean ball. We define the surrogate policy measure

$$\hat{\boldsymbol{y}}(\boldsymbol{\theta}) := \sum_{\boldsymbol{y} \in \mathcal{Y}} p_0(\boldsymbol{y}|\boldsymbol{\theta}) \delta_{\boldsymbol{y}}$$

where $\delta_{\boldsymbol{y}}$ is the Dirac mass at point $\boldsymbol{y}$ and

$$p_0(\boldsymbol{y}|\boldsymbol{\theta}) := \mathbb{P}_U\left[\boldsymbol{\theta} + \rho(\boldsymbol{\theta})U \in \mathcal{F}_{\boldsymbol{y}}\right]$$

where U is uniformly distributed on $B(0,1)$.

Note that $p_0(\boldsymbol{y}|\boldsymbol{\theta})$ measures the proportion of normal cone $\mathcal{F}_{\boldsymbol{y}}$ locally around $\boldsymbol{\theta}$.

Remark 2 (Generic case and an abuse of notation)**.** *Generically, a direction $\boldsymbol{\theta}$ is in the interior of a normal cone $\mathcal{F}_{\boldsymbol{y}}$ and $\boldsymbol{y}' \mapsto p_0(\boldsymbol{y}'|\boldsymbol{\theta})$ is zero except at point $\boldsymbol{y}$. In this case, the measure $\hat{\boldsymbol{y}}(\boldsymbol{\theta})$ is a Dirac mass at point $\boldsymbol{y}$ and, by an abuse of notation, we denote by $\hat{\boldsymbol{y}}(\boldsymbol{\theta})$ the vector $\boldsymbol{y} \in \mathbb{R}^{d(\boldsymbol{y})}$.*
To ease notation, we will consider that, by convention,

$$f(\hat{\boldsymbol{y}}(\boldsymbol{\theta})) := \sum_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})} p_0(\boldsymbol{y}|\boldsymbol{\theta}) f(\boldsymbol{y})$$

for all functions f and all $\boldsymbol{\theta} \in \mathbb{R}^{d(\boldsymbol{x})}$.

2.2 The regularized and empirical risks

In our algorithms, we optimize an empirical and regularized version of the risk $\mathcal{R}$, which we now introduce. We observe n instances $X_1, \dots, X_n$ drawn independently according to a (possibly unknown) distribution $\mathbb{P}_X$. For all $\boldsymbol{w} \in \mathcal{W}$, we denote the policy by $h_{\boldsymbol{w}} : \boldsymbol{x} \mapsto \hat{\boldsymbol{y}}(\psi_{\boldsymbol{w}}(\boldsymbol{x}))$ and introduce the regularized population risk $\mathcal{R}_t$ and its empirical version $\mathcal{R}_{n,t}$ by

$$\mathcal{R}_t(h_{\boldsymbol{w}}) := \mathbb{E}_{X,Z}\big[f^0(\hat{\boldsymbol{y}}(\psi_{\boldsymbol{w}}(X) + tZ(X)), X)\big]\,, \tag{2.2}$$

$$\mathcal{R}_{n,t}(h_{\boldsymbol{w}}) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_Z\Big\{\big[f^0(\hat{\boldsymbol{y}}(\psi_{\boldsymbol{w}}(X_i) + tZ(X_i)), X_i)\big]\Big\}\,, \tag{2.3}$$

where $t \geq 0$ is some regularization parameter and $Z(\cdot)$ is a perturbation defined as follows.

Definition 3 (Law of the perturbation)**.** *Given $\boldsymbol{x} \in \mathcal{X}$, $Z(\boldsymbol{x})$ has the same law as $R \times U$ where the random vector U is uniformly distributed on the Euclidean sphere of $\mathbb{R}^{d(\boldsymbol{x})}$ and the random positive scalar $R \in (0, \infty)$ is independent of U.*

Remark 4. *The perturbation $Z(\boldsymbol{x})$ is a random vector in $\mathbb{R}^{d(\boldsymbol{x})}$ that is uniformly distributed on the Euclidean sphere of $\mathbb{R}^{d(\boldsymbol{x})}$ and is scaled by a random positive scalar R. As such, the perturbed vector $\psi_{\boldsymbol{w}}(\boldsymbol{x}) + tZ(\boldsymbol{x})$ belongs to the normal cone of $\psi_{\boldsymbol{w}}(\boldsymbol{x})$ whenever $tR \leq \rho(\psi_{\boldsymbol{w}}(\boldsymbol{x}))$, where $\rho(\psi_{\boldsymbol{w}}(\boldsymbol{x}))$ is the internal radius of the normal cone at $\psi_{\boldsymbol{w}}(\boldsymbol{x})$. Note that this latter condition does not depend on the direction $U(G)$ of the perturbation $Z(\boldsymbol{x})$.*

For later use in our analysis we introduce the following quantity.

Definition 5 (Perturbed surrogate policy probabilities)**.** *Let $\lambda > 0$ and $\boldsymbol{x} \in \mathcal{X}$. Let $\boldsymbol{\theta} \in \mathbb{R}^{d(\boldsymbol{x})}$ and $\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})$. We define the perturbed surrogate policy probability at solution point $\boldsymbol{y}$ for the direction $\boldsymbol{\theta}$ as*

$$p_\lambda(\boldsymbol{y}|\boldsymbol{\theta}) := \mathbb{E}_Z[p_0(\boldsymbol{y}|\boldsymbol{\theta} + \lambda Z(\boldsymbol{x}))]\,.$$

The (non-perturbed) surrogate policy of $\boldsymbol{x} \in \mathcal{X}$ is defined as a solution $\hat{\boldsymbol{y}}(\psi_{\boldsymbol{w}}(\boldsymbol{x})) \in \mathbb{R}^{d(\boldsymbol{x})}$ to (1.4), and the surrogate policy model is defined by (1.5).

Now, let $\varepsilon_0 \geq 0$ and consider the risk

$$\begin{aligned}
\mathcal{R}_{\varepsilon_0}(\mathcal{H}) &:= \min_{\boldsymbol{w} \in \mathcal{W}} \big\{\mathbb{E}_{X,Z}\big[f^0(\hat{\boldsymbol{y}}(\psi_{\boldsymbol{w}}(X) + \varepsilon_0 Z(X)), X)\big]\big\}\,, \qquad (2.4)\\
&= \min_{\boldsymbol{w} \in \mathcal{W}} \big\{\mathbb{E}_{X,Z}\big[\sum_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})} p_0(\boldsymbol{y}|\psi_{\boldsymbol{w}}(X) + \varepsilon_0 Z(X))\, f^0(\boldsymbol{y}, X)\big]\big\}\,,\\
&= \min_{\boldsymbol{w} \in \mathcal{W}} \big\{\mathbb{E}_{X}\big[\sum_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})} p_{\varepsilon_0}(\boldsymbol{y}|\psi_{\boldsymbol{w}}(X))\, f^0(\boldsymbol{y}, X)\big]\big\}\,,\\
&= \inf_{\boldsymbol{w} \in \mathcal{W}} \mathcal{R}_{\varepsilon_0}(h_{\boldsymbol{w}})\,,\\
&= \mathcal{R}_{\varepsilon_0}(h_{\boldsymbol{w}^*})\,. \qquad (2.5)
\end{aligned}$$

We aim at finding the optimal surrogate policy $h_{\boldsymbol{w}^\star} : \boldsymbol{x} \mapsto \hat{\boldsymbol{y}}(\psi_{\boldsymbol{w}^\star}(\boldsymbol{x}))$, where $\boldsymbol{w}^\star$ is a solution to (2.4), assumed to exist.

Remark 6. *(cases $\varepsilon_0 = 0$ and $\varepsilon_0 > 0$) In the case where $\varepsilon_0 = 0$, note that $\mathcal{R}_0(h_{\boldsymbol{w}}) = \mathbb{E}_X\big[f^0(\hat{\boldsymbol{y}}(\psi_{\boldsymbol{w}}(X)), X)\big]$ and it may fail to be continuous in $\boldsymbol{w}$. Hence it may not be clear that a solution to (2.4) exists.*

In the case $\varepsilon_0 > 0$, we will prove that the risk is continuous and hence a solution always exists, under mild assumptions (Proposition 1). Furthermore, we will also prove that the Property ($\mathrm{UW}_{\varepsilon_0}$), introduced later, holds, under mild assumptions (Proposition 2). So rather than studying a perturbation scaling λ such that $\lambda \in (0, +\infty)$, we suggest instead to fix $\varepsilon_0 > 0$ and focus on $\lambda \in [\varepsilon_0, +\infty)$. The small constant ε_0 should be then understood as a minimal amount of regularization to introduce, playing the role of machine precision.

2.3 Guarantees

Policy model misspecification error. We define the policy model misspecification error as

$$\begin{aligned}\mathcal{E}_{\varepsilon_0}(\mathcal{H}) &:= \mathcal{R}_{\varepsilon_0}(\mathcal{H}) - \mathbb{E}_X\big[f^0(\boldsymbol{y}^0(X), X)\big] \\ &= \mathbb{E}_{X,Z}\big[f^0(\hat{\boldsymbol{y}}(\psi_{\boldsymbol{w}^\star}(X) + \varepsilon_0 Z(X)), X)\big] - \mathbb{E}_X\big[f^0(\boldsymbol{y}^0(X), X)\big] \geq 0\end{aligned}$$

where $\boldsymbol{y}^0(\boldsymbol{x})$ minimizes (1.1).

In $\mathbb{P}_X$-average, it measures, for the loss given by f^0, how the policy $\hat{\boldsymbol{y}}(\psi_{\boldsymbol{w}^\star}(X) + \varepsilon_0 Z(X))$ fails to approximate $\boldsymbol{y}^0(X)$ or equivalently how $\psi_{\boldsymbol{w}^\star}(X) + \varepsilon_0 Z(X)$ fails to point towards normal cones $\mathcal{F}_{\boldsymbol{y}}$ at point $\boldsymbol{y}$ such that $f^0(\boldsymbol{y}, X) - f^0(\boldsymbol{y}^0(X), X)$ is small. To apprehend the term $\mathcal{E}_{\varepsilon_0}(\mathcal{H})$, one needs some regularity-related notion for $\boldsymbol{y}^0$ and f^0. This study is out of the scope of this paper, and we focus on the optimal risk defined in (2.4).

Error bounds We will introduce an estimator $h_{\boldsymbol{w}_{M,n,\lambda}}$ which depends on some complexity integer parameter $M \geq 1$, the sample size n, and a real valued tuning parameter $\lambda > 0$ such that $\lambda \geq \varepsilon_0$. Assume that $\boldsymbol{w}^\star$ exists, then one has the following decomposition of the error, with each term handled separately,

$$\begin{aligned}0 \leq\ & \mathcal{R}_{\varepsilon_0}(h_{\boldsymbol{w}_{M,n,\lambda}}) - \mathbb{E}_X\big[f^0(\boldsymbol{y}^0(X), X)\big] \\ &= \underbrace{\mathcal{R}_{\varepsilon_0}(h_{\boldsymbol{w}_{M,n,\lambda}}) - \mathcal{R}_\lambda(h_{\boldsymbol{w}_{M,n,\lambda}})}_{\text{Theorem 2}} + \underbrace{\mathcal{R}_\lambda(h_{\boldsymbol{w}_{M,n,\lambda}}) - \mathcal{R}_{n,\lambda}(h_{\boldsymbol{w}_{M,n,\lambda}})}_{\text{Theorem 3}} \\ &+ \underbrace{\mathcal{R}_{n,\lambda}(h_{\boldsymbol{w}_{M,n,\lambda}}) - \mathcal{R}_{n,\lambda}(h_{\boldsymbol{w}_{n,\lambda}})}_{\text{Theorem 4}} + \underbrace{\mathcal{R}_{n,\lambda}(h_{\boldsymbol{w}_{n,\lambda}}) - \mathcal{R}_{n,\lambda}(h_{\boldsymbol{w}^\star})}_{\leq 0} \\ &+ \underbrace{\mathcal{R}_{n,\lambda}(h_{\boldsymbol{w}^\star}) - \mathcal{R}_\lambda(h_{\boldsymbol{w}^\star})}_{\text{Theorem 3}} + \underbrace{\mathcal{R}_\lambda(h_{\boldsymbol{w}^\star}) - \mathcal{R}_{\varepsilon_0}(h_{\boldsymbol{w}^\star})}_{\text{Theorem 2}} \\ &+ \mathcal{E}_{\varepsilon_0}(\mathcal{H})\end{aligned}$$

where, for any $\boldsymbol{w} \in \mathcal{W}$, we define $h_{\boldsymbol{w}_{n,\lambda}}$ such that

$$\mathcal{R}_{n,\lambda}(h_{\boldsymbol{w}_{n,\lambda}}) = \min_{\boldsymbol{w} \in \mathcal{W}} \mathcal{R}_{n,\lambda}(h_{\boldsymbol{w}})\,.$$

In practice, we will make some working assumptions, to obtain the following, possibly probabilistic, bounds on each of the terms.

Theorem 1. *Under the conditions given in Section 3.1, the following holds true. Let $\varepsilon_0 \geq 0$ and $\lambda > 0$ be such that $\lambda \geq \varepsilon_0$. Let $\tau \in (0, 1)$. There exists a constant $C > 0$ that depends only on ε, τ and f^0 such that for any $\boldsymbol{w} \in \mathcal{W}$ and $n \geq 1$, one has*

$$|\mathcal{R}_{\varepsilon_0}(h_{\boldsymbol{w}}) - \mathcal{R}_\lambda(h_{\boldsymbol{w}})| = C\lambda^\tau \mathrm{polylog}(\lambda) \qquad \textit{(Theorem 2, Perturbation bias)}$$

$$|\mathcal{R}_\lambda(h_{\boldsymbol{w}}) - \mathcal{R}_{n,\lambda}(h_{\boldsymbol{w}})| = \mathcal{O}_{\mathbb{P}}\Big(\frac{1}{\lambda\sqrt{n}}\Big) \qquad \textit{(Theorem 3, Empirical process)}$$

and, for $h_{\boldsymbol{w}_{n,\lambda}}$ given by the kernel Sum-of-Squares estimate solution to $\min_{\boldsymbol{w}\in\mathcal{W}} \mathcal{R}_{n,\lambda}$,

$$|\mathcal{R}_{n,\lambda}(h_{\boldsymbol{w}_{M,n,\lambda}}) - \mathcal{R}_{n,\lambda}(h_{\boldsymbol{w}_{n,\lambda}})| = \mathcal{O}_{\mathbb{P}}\left[\left(\frac{1}{\lambda(M/\log M)^{\frac{1}{d}}}\right)^{s-\frac{d}{2}}\right] \qquad \textit{by Theorem 4 (K-SoS)}$$

where polylog(λ) *is a polynomial logarithm term and $s > d/2$ is some tuning parameter on the order of regularity of the admissible functions. It yields the following bound on the error between the risk and the regularized risk*

$$0 \leq \mathcal{R}_{\varepsilon_0}(h_{\boldsymbol{w}_{M,n,\lambda}}) - \mathbb{E}_X\left[f^0(\boldsymbol{y}^0(X), X)\right] \leq C\lambda^{\tau}\text{polylog}(\lambda) + \mathcal{O}_{\mathbb{P}}\left[\left(\frac{1}{\lambda\sqrt{n}}\right) + \left(\frac{1}{\lambda(M/\log M)^{\frac{1}{d}}}\right)^{s-\frac{d}{2}}\right] + \mathcal{E}_{\varepsilon_0}(\mathcal{H})\,,$$

where the K-SoS bound is given up to a small a posteriori error, see Remark 11.

Our excess risk bound is of the type *statistical to computational* showing the effect of regularisation λ, sampling n and the complexity of the algorithm M. It shows that, when λ goes to zero, the excess risk goes to zero if $\sqrt{n}$ dominates $1/\lambda$ and M dominates $1/\lambda^d$. The first shows quantitatively how many sample points the practitioner needs when diminishing the regularization λ. Our statistical bound matches the parametric rate $1/\sqrt{n}$ and $1/\lambda$ can be interpreted as a standard error. The latter is a so-called curse of dimension, the complexity of the algorithm is exponential in the dimension d and $1/\lambda$ can be interpreted as a smoothness parameter.

Example 1 (continued). *In our reduction of contextual stochastic optimization to the setting of this article, the definition of X as $(\tilde{X}, \xi)$ and of $f^0(\boldsymbol{y}, \boldsymbol{x})$ as $f^c(\boldsymbol{y}, \tilde{\boldsymbol{x}}, \boldsymbol{\xi})$ may have seemed surprising. The common practice in stochastic optimization would have suggested defining X as $\tilde{X}$ and $f^0(\boldsymbol{y}, \boldsymbol{x})$ as $\mathbb{E}\left[f^c(\boldsymbol{y}, X, \xi)\middle|X = \boldsymbol{x}\right]$. Our reduction has however two advantages.*

First, we do not know the distribution over $(\tilde{X}, \xi)$, but only have access to a training set $\tilde{\boldsymbol{x}}_1, \boldsymbol{\xi}_1, \ldots, \tilde{\boldsymbol{x}}_n, \boldsymbol{\xi}_n$. Hence, we have all the data needed to formulate the empirical risk minimization problem with our reduction, while the alternative reduction would require the conditional distribution on ξ given $\tilde{X} = \boldsymbol{x}_i$, which we do not know.

Second, with our reduction, Theorem 1 applies out of the box to contextual stochastic optimization. In stochastic optimization, we select a decision by coordinating the different scenarios for a given instance/context. In our case, we select instead a policy by coordinating scenarios across different instances/context. Using a machine learning terminology, our reduction is a discriminative approach, while the other one, which would be closer to the common practice in stochastic optimization, is a generative approach. Theorem 1 provides as theoretical basis for this discriminative approach.

3 Main results

We present in this section the results we use to obtain Theorem 1. As discussed above, our analysis involves a parameter $\varepsilon_0 \geq 0$ that accounts for a small perturbation defining the risk $\mathcal{R}_{\varepsilon_0}$. Our main result is an upper bound on the excess risk when Property ($\text{UW}_{\varepsilon_0}$), defined later, is met. We show that this property is always satisfied when $\varepsilon_0 > 0$. In the case $\varepsilon_0 = 0$ there is no regularization on the risk, but one has to check Property ($\text{UW}_{\varepsilon_0}$) to apply our result. We will give four cases in which Property ($\text{UW}_{\varepsilon_0}$) is met when $\varepsilon_0 = 0$.

3.1 The conditions

Partition of the instances On many combinatorial optimization problems, instances $\boldsymbol{x}$ are pairs (G, ξ), where G is a combinatorial object, and ξ_G is a vector of parameters whose dimensions depend only on G, *e.g.*, G is typically a graph and ξ_G contains a vector of parameters (a label) for each vertex or edge of G. Furthermore, we know an embedding of $\mathcal{Y}(\boldsymbol{x})$ in the Euclidean space $\mathbb{R}^{d(\boldsymbol{x})}$ with $d(\boldsymbol{x})$ of polynomial size in $\boldsymbol{x}$. For instance,

solutions of the shortest path problems are the vertices of the path polytope, solutions of the minimum weight spanning tree problem are the vertices of the spanning tree polytope, and so on. More generally, most combinatorial optimization problems admit a mixed integer linear programming formulation. As illustrated by these examples, the set of feasible solutions $\mathcal{Y}(\boldsymbol{x})$ frequently depends only on G. To avoid technicalities, we omit G and ξ_G in the analysis but make the following assumption about a partition of the space of instances.

Assumption (Part)**.** *The following holds.*

- *There is a finite partition $\mathcal{G}$ of $\mathcal{X}$ into $(\mathcal{X}_G)_{G\in\mathcal{G}}$ such that $\mathcal{Y}(\boldsymbol{x})$ are constant and finite for all $\boldsymbol{x}\in\mathcal{X}_G$;*
- *The absolute value of the target function $|f^0|$ is uniformly bounded on $\mathcal{X}_G$ for each $G\in\mathcal{G}$ and, hence, its oscillation is finite, $\operatorname{osc}(f^0)<\infty$, with*

$$\operatorname{osc}(f^0) := \sup f^0 - \inf f^0 \,.$$

Example 2 (continued). *In the context of the stochastic vehicle scheduling, G corresponds to the digraph D we introduced, and $\mathcal{G}$ to all possible digraphs"? Assuming that $\mathcal{G}$ is finite is therefore natural: since no firm is going to serve an infinite number of requests, $\mathcal{G}$ is a subset of the set of acyclic digraphs with at most n^{v} vertices for some n^{v}. Since we are working with the flow polytope, $d(G)$ corresponds to the number of arcs $|A|$ in digraph D. The second part of the assumption holds if costs are positive and if there is a maximum cost that can be encountered per task, which is natural in any real life setting.*

Example 3 (continued). *In the context of single machine scheduling, the set of feasible solutions depends only on the number n of tasks. As a consequence, G can be identified with n, and $\mathcal{G}$ is the subset of $\mathbb{Z}_+$ corresponding to instance sizes. Supposing that it is finite only amounts to having an upper bound on the number of tasks. Finally, $d(G)=n$.*

Remark 7. *Under Assumption (Part), note that, for each $G\in\mathcal{G}$, and all $\boldsymbol{x}\in\mathcal{X}_G$, the embedding of $\mathcal{Y}(\boldsymbol{x})$ into $\mathbb{R}^{d(\boldsymbol{x})}$ and the dimension $\boldsymbol{x}\to d(\boldsymbol{x})$ are constant. We denote by $d(G)$ the value of $d(\boldsymbol{x})$ for $\boldsymbol{x}\in\mathcal{X}_G$. Without loss of generality, we assume that $\mathcal{Y}(\boldsymbol{x})\subseteq\mathbb{R}^{d(\boldsymbol{x})}$ and $\operatorname{Vect}(\mathcal{Y}(\boldsymbol{x}))=\mathbb{R}^{d(\boldsymbol{x})}$.*

Under Assumption (Part), we have the following decomposition of the law $\mathbb{P}_X$ of X.

$$\mathbb{E}_X[t(X)] := \int_{\mathcal{X}} t(\boldsymbol{x})\mathrm{d}\mathbb{P}_X(\boldsymbol{x}) = \sum_{G\in\mathcal{G}} \mathbb{P}[X\in\mathcal{X}_G]\int_{\mathcal{X}_G} t(\boldsymbol{x})\mathrm{d}\mathbb{P}_{\xi_G}(\boldsymbol{x})\,, \tag{3.1}$$

where $\mathbb{P}_{\xi_G}$ is the law of ξ_G, namely the conditional law of $\mathcal{L}(X|X\in\mathcal{X}_G)$, and t is any bounded continuous function. Denote by $\mathbb{P}_X^{(n)}$ its empirical version, given by

$$\mathbb{E}_X^{(n)}[t(X)] := \frac{1}{n}\sum_{i\in[n]} t(X_i) = \frac{1}{n}\sum_{i\in[n]}\sum_{G\in\mathcal{G}} \mathbb{1}_{\{X_i\in\mathcal{X}_G\}} t(X_i). \tag{3.2}$$

The Gaussian perturbation assumption: Throughout this paper, we consider that the perturbation random field $Z(\cdot)$ is a Gaussian random field satisfying the following assumption.

Assumption (Gauss)**.** *We assume that $\sqrt{d(G)}R(G)$ has chi-distribution with $d(G)$ degrees of freedom which is equivalent to*

$$\forall \boldsymbol{x}\in\mathcal{X}_G\,,\quad \sqrt{d(G)}Z(\boldsymbol{x})\sim\mathcal{N}(0,\mathrm{Id}_{d(G)})\,. \tag{3.3}$$

Under Assumptions (Part) and (Gauss), we have the following decomposition of the law $\mathcal{L}(X,Z)$,

$$\mathbb{E}_{X,Z}\big[t(X,Z)\big] := \sum_{G\in\mathcal{G}} \mathbb{P}[X\in\mathcal{X}_G]\int_{\mathcal{X}_G}\Big(\int_{\mathbb{R}^{d(G)}} t(\boldsymbol{x},\boldsymbol{z})\varphi_{d(G)}(\boldsymbol{z})\mathrm{d}\boldsymbol{z}\Big)\mathrm{d}\mathbb{P}_{\xi_G}(\boldsymbol{x})\,, \tag{3.4}$$

where $\mathbb{P}_{\xi_G}$ is the law of ξ_G, namely the conditional law of $\mathcal{L}(X|X\in\mathcal{X}_G)$, $\varphi_{d(G)}$ is the Gaussian density of $\mathcal{N}(0,(1/d(G))\mathrm{Id}_{d(G)})$, and $t(\cdot)$ is any bounded continuous function.

Lipschitz feature map: The following assumption is about the Lipschitz continuity of the feature map.

Assumption (Lip). *For all $\boldsymbol{x} \in \mathcal{X}$, the function $\boldsymbol{w} \in \mathcal{W} \mapsto \psi_{\boldsymbol{w}}(\boldsymbol{x}) \in \mathbb{R}^{d(\boldsymbol{x})}$ is $L_{\mathcal{W}}$-Lipschitz continuous with a constant $L_{\mathcal{W}}$ which does not depend on $\boldsymbol{x}$.*

Example 2 (continued). *In the stochastic vehicle scheduling problem example, we use a generalized linear model with $\theta_a = \boldsymbol{w}^\top \boldsymbol{\phi}_a$, which $\boldsymbol{\phi}_a$ being some features of the connection a, related to the slack between the tasks that are the tails and the head of a. In applications, it is natural to assume that this slack is bounded, which gives our Lipschitz assumption.*

Under these assumptions

Remark 8. *When $\boldsymbol{w}$ encodes the parameters of a neural network with smooth activation functions one can compute the differential with respect to $\boldsymbol{w}$ at input point $\boldsymbol{x}$ of neural network, namely the gradient of $\boldsymbol{w} \mapsto \psi_{\boldsymbol{w}}(\boldsymbol{x})$. By the chain rule, one obtain that partial gradients $\partial\psi_{\boldsymbol{w}}(\boldsymbol{x})/\partial w_i$ are product of derivatives of the activation functions evaluated at layers values, matrices of some of the parameters $\boldsymbol{w}$, and input vector $\boldsymbol{x}$. Since the layers values are themselves composition of activation functions and product of matrices of some parameters $\boldsymbol{w}$ and input vector $\boldsymbol{x}$, they are bounded by compactness of $\mathcal{W}$ and $\mathcal{X}$. Hence, partial gradients $\partial\psi_{\boldsymbol{w}}(\boldsymbol{x})/\partial w_i$ are also bounded by compactness. We deduce that (Lip) is satisfied.*

Existence of solutions: We assume that the optimization problem (2.4) has a solution and that the minimum of the risk has a minimizer.

Assumption (Sol). *There exists $\boldsymbol{w}^\star \in \mathcal{W}$ solution to (2.4) and $\min_{\boldsymbol{w}\in\mathcal{W}} \mathcal{R}_{n,\varepsilon_0}(h_{\boldsymbol{w}})$ has a minimizer.*

Proposition 1. *Let ε_0 be positive and let Assumptions (Gauss) and (Lip) hold. Then, Assumption (Sol) is met.*

Proof. When $\varepsilon_0 > 0$, the risk $\boldsymbol{w} \mapsto \mathcal{R}_{\varepsilon_0}(h_{\boldsymbol{w}})$ is continuous with respect to $\boldsymbol{w}$ (see Lemma 7) and the set $\mathcal{W}$ is compact. Therefore, the infimum in (2.4) is attained. □

The uniform weak (UW) moment property: We introduce a property that is useful for the proofs, and use it to discuss the impact of the perturbation on the policy.

Given a regularization parameter $\lambda > \varepsilon_0 \geq 0$, we will see later that the uniform weak moment property, given below, implies that $\boldsymbol{w} \mapsto |\mathcal{R}_\lambda(h_{\boldsymbol{w}}) - \mathcal{R}_{\varepsilon_0}(h_{\boldsymbol{w}})|$ is uniformly bounded by $\mathcal{O}(\lambda^\tau \mathrm{polylog}(\lambda))$ where $\mathrm{polylog}(\lambda)$ is a polynomial logarithm term and $\tau \in (0,1)$, see Theorem 2.

Property ($\mathrm{UW}_{\varepsilon_0}$). *For all $\tau \in (0,1)$, it exists a positive constant $C_{\varepsilon_0,\tau} > 0$ such that*

$$\forall \boldsymbol{w} \in \mathcal{W}\,,\ \mathbb{E}_{X,Z}\left[\left(\frac{\rho(\psi_{\boldsymbol{w}}(X) + \varepsilon_0 Z(X))}{\sqrt{d(X)}}\right)^{-\tau}\right] \leq C_{\varepsilon_0,\tau}\,. \tag{3.5}$$

3.2 Discussion on the UW property

UW is always met for positive regularization: We have the following proposition showing that, when $\varepsilon_0 > 0$, Property ($\mathrm{UW}_{\varepsilon_0}$) holds with a constant $C_{\varepsilon_0,\tau}$ that depends only on ε_0 and τ.

Proposition 2. *If $\varepsilon_0 > 0$ then, under Assumptions (Part) and (Gauss), Property ($\mathrm{UW}_{\varepsilon_0}$) holds and*

$$C_{\varepsilon_0,\tau} \leq \Big(\int_{\mathbb{R}} |t|^{-\tau}\varphi(t)\mathrm{d}t\Big) \times \varepsilon_0^{-\tau} \times \sum_{G\in\mathcal{G}} \mathbb{P}[X \in \mathcal{X}_G]\,|\mathcal{Y}(G)|^2\,d(G)^\tau\ ,$$

where φ is the standard Gaussian density on $\mathbb{R}$.

Proof. In view of (3.4) and without loss of generality, we consider that $X \in \mathcal{X}_G$ almost surely, for G fixed. Let H be an hyperplane and $d(\theta, H)$ be the Euclidean distance between θ and H. By (3.4), note that

$$\mathbb{E}_{X,Z}\left[\Big(\frac{d(\psi_{\boldsymbol{w}}(X)+\varepsilon_0 Z(X))}{\sqrt{d(X)}}\Big)^{-\tau}\right] = \mathbb{E}_{\xi_G,Z}\left[\Big(\frac{d(G)^{\frac{\tau}{2}}}{d(\psi_{\boldsymbol{w}}(\xi_G)+\varepsilon_0 Z(\xi_G), H)^{\tau}}\Big)\right]$$
$$= d(G)^{\frac{\tau}{2}}\,\mathbb{E}_{\xi_G}\left[\mathbb{E}_Z\Big(\frac{1}{d(\psi_{\boldsymbol{w}}(\xi_G)+\varepsilon_0 Z(\xi_G), H)^{\tau}}\Big|\xi_G\Big)\right].$$

Conditionally to ξ_G, observe that $d(\psi_{\boldsymbol{w}}(\xi_G)+\varepsilon_0 Z(\xi_G), H)^{\tau}$ is the distance between the hyperplane H and a Gaussian with mean $\psi_{\boldsymbol{w}}(\xi_G)$ and variance $(\varepsilon_0^2/d(G))\mathrm{Id}_{d(G)}$. This distance is

$$d(\psi_{\boldsymbol{w}}(\xi_G)+\varepsilon_0 Z(\xi_G), H)^{\tau} = |\langle n_H\,,\,\psi_{\boldsymbol{w}}(\xi_G)+\varepsilon_0 Z(\xi_G)\rangle|^{\tau}$$

where n_H is a unitary orthogonal vector to H. Conditionally to ξ_G, $\langle n_H\,,\,\psi_{\boldsymbol{w}}(\xi_G)+\varepsilon_0 Z(\xi_G)\rangle$ is a Gaussian random variable with mean $\langle n_H\,,\,\psi_{\boldsymbol{w}}(\xi_G)\rangle$ and variance $\varepsilon_0^2/d(G)$.

Observe that the level sets of $t \mapsto |t|^{-\tau}$ are symmetric convex bodies, namely intervals of the form $[-a, a]$ for $a > 0$. By Anderson's lemma [Giné and Nickl, 2021, Theorem 2.4.4], it holds that

$$\mathbb{E}_Z\Big(\frac{1}{d(\psi_{\boldsymbol{w}}(\xi_G)+\varepsilon_0 Z(\xi_G), H)^{\tau}}\Big|\xi_G\Big) \le \mathbb{E}_Z\Big(\frac{1}{d(\varepsilon_0 Z(\xi_G), H)^{\tau}}\Big|\xi_G\Big) = \frac{d(G)^{\frac{\tau}{2}}}{\varepsilon_0^{\tau}}\int_{\mathbb{R}} |t|^{-\tau}\varphi(t)\mathrm{d}t\,.$$

We deduce that

$$\mathbb{E}_{X,Z}\left[\Big(\frac{d(\psi_{\boldsymbol{w}}(X)+\varepsilon_0 Z(X))}{\sqrt{d(X)}}\Big)^{-\tau}\right] \le \frac{d(G)^{\tau}}{\varepsilon_0^{\tau}}\int_{\mathbb{R}} |t|^{-\tau}\varphi(t)\mathrm{d}t < \infty$$

Then observe that for Q a polyhedral cone with non-empty interior we have that Q is the intersection of halfspaces delimited by $H_1, \ldots, H_D$. Note also that $\frac{1}{d(\theta,Q)^{\tau}} \le \frac{1}{d(\theta,H_1)^{\tau}} + \ldots + \frac{1}{d(\theta,H_D)^{\tau}}$, and thus $\theta \mapsto \frac{1}{d(\theta,Q)^{\tau}}$ is integrable with respect to the law of $\psi_{\boldsymbol{w}}(X)+\varepsilon_0 Z(X)$ conditional to $X \in \mathcal{X}_G$. Finally, observe that $\theta \mapsto \rho(\theta)$ is the distance to the boundaries of polyhedral cones. □

The no perturbation case: We would like to show that Property ($\mathrm{UW}_{\varepsilon_0}$) is met when $\varepsilon_0 = 0$ in some cases. Under Assumption (Part) and without loss of generality, we consider that $d(X)$ is constant almost surely, say $d(X) = d$. We consider four scenarios on the family of laws

$$\mathbb{L} := \big\{\mathcal{L}(\psi_{\boldsymbol{w}}(X)) \;:\; \boldsymbol{w} \in \mathcal{W}\big\}$$

where $\mathcal{L}(\psi_{\boldsymbol{w}}(X))$ denotes the law of $\psi_{\boldsymbol{w}}(X)$, given by

(i) There exists a compact $K \subset \mathbb{R}^d$ such that the supports of the laws of $\mathbb{L}$ are included in K and these laws are absolutely continuous, with respect to the Lebesgue measure, with uniformly bounded density;

(ii) There exists a finite set $K \subset \mathbb{R}^d$ such that the laws of $\mathbb{L}$ are finite with support included in K;

(iii) There exists a compact Riemannian submanifold $K \subset \mathbb{R}^d$ (with metric induced by the Euclidean metric) such that the supports of the laws of $\mathbb{L}$ are included in K and these laws are absolutely continuous, with respect to the Riemannian measure, with uniformly bounded density;

(iv) The laws of $\mathbb{L}$ are a mixture of an absolutely continuous law satisfying (i), a discrete law satisfying (ii), and a singular continuous law satisfying (iii).

Now, using that

- $\theta \mapsto \rho(\theta)$ is the distance to the boundaries of polyhedral cones;
- $\frac{1}{d(\theta,Q)^\tau} \leq \frac{1}{d(\theta,H_1)^\tau} + \ldots + \frac{1}{d(\theta,H_D)^\tau}$ where Q a polyhedral cone wich is the intersection of halfspaces delimited by $H_1, \ldots, H_D$;
- $\theta \mapsto \frac{1}{d(\theta,Q)^\tau}$ is integrable with respect to the uniform measure on compacts, hence with respect to any bounded density with compact support;

we deduce that Property ($\mathrm{UW}_{\varepsilon_0}$) holds for the four scenarios above.

Remark 9. *Let us first consider a generalized linear model, such as in our stochastic vehicle scheduling of Example 2, where $\boldsymbol{\theta} = \boldsymbol{\Phi}(x)^\top \boldsymbol{w}$ for a given matrix of features. Then Case* (i) *is satisfied when every feature is continuous and upper bounded and the number of features is rich enough. If the number of features is small, then the image of the instances through the model may be included in a hyperplane, which leads to Case* (iii)*. Case* (ii) *corresponds to features that take integer values and are finite. Using a submanifold instead of a hyperplane, Case* (iii) *enables us to deal with neural networks in $\psi_{\boldsymbol{w}}$.*

Example 2 (continued). *In the stochastic vehicle scheduling problem example, continuous features would typically correspond to features reflecting properties of the delay or the distance, which are by nature continuous and upper bounded. Discrete features can for instance indicate combinatorial properties such as the number of arcs departing from the tail of an arc a.*

3.3 Regularization by perturbation and perturbation bias

We start by introducing

$$V_{\boldsymbol{w}}(\lambda) := \mathbb{E}_X \mathbb{P}_Z \Big[\|Z(X)\|_2 > \frac{\rho(\psi_{\boldsymbol{w}}(X))}{\lambda} \,\Big|\, X \Big] = \mathbb{E}_X \mathbb{P}_R \Big[R(X) > \frac{\rho(\psi_{\boldsymbol{w}}(X))}{\lambda} \,\Big|\, X \Big] ,$$

for $\lambda > 0$ and $\boldsymbol{w} \in \mathcal{W}$. By dominated convergence, note that $V_{\boldsymbol{w}}(\lambda)$ tends to zero as λ goes to zero and that $V_{\boldsymbol{w}}(\lambda)$ tends to one as λ goes to infinity. Note also that the function $\lambda \mapsto V_{\boldsymbol{w}}(\lambda) \in [0,1]$ is non-decreasing.

Our first contribution is a bound on the risk perturbation.

Proposition 3. *Let $\varepsilon_0 \geq 0$ and $\lambda > 0$ be such that $\lambda \geq \varepsilon_0$. It holds that, for all $\boldsymbol{w} \in \mathcal{W}$,*

$$\big|\mathcal{R}_\lambda(h_{\boldsymbol{w}}) - \mathcal{R}(h_{\boldsymbol{w}})\big| \leq 2 \operatorname{osc}(f^0) V_{\boldsymbol{w}}(\lambda) \quad \text{and} \quad \big|\mathcal{R}_\lambda(h_{\boldsymbol{w}}) - \mathcal{R}_{\varepsilon_0}(h_{\boldsymbol{w}})\big| \leq 4 \operatorname{osc}(f^0) V_{\boldsymbol{w}}(\lambda) .$$

Proof. Recall that $p_0(\boldsymbol{y}|\psi_{\boldsymbol{w}}(\boldsymbol{x})) = \mathbb{P}[\psi_{\boldsymbol{w}}(\boldsymbol{x}) + \rho(\psi_{\boldsymbol{w}}(\boldsymbol{x}))U \in \mathcal{F}_{\boldsymbol{y}}]$ where U is uniformly on $B(0,1)$ and

$$p_\lambda(\boldsymbol{y}|\psi_{\boldsymbol{w}}(\boldsymbol{x})) = \mathbb{P}_Z[\psi_{\boldsymbol{w}}(\boldsymbol{x})/\lambda + Z(\boldsymbol{x}) \in \mathcal{F}_{\boldsymbol{y}}] = \mathbb{P}_Z[\psi_{\boldsymbol{w}}(\boldsymbol{x})/\lambda + R(\boldsymbol{x})U(\boldsymbol{x}) \in \mathcal{F}_{\boldsymbol{y}}] .$$

Observe that $p_0(\boldsymbol{y}|\psi_{\boldsymbol{w}}(\boldsymbol{x})) = \mathbb{P}[\psi_{\boldsymbol{w}}(\boldsymbol{x}) + \rho(\psi_{\boldsymbol{w}}(\boldsymbol{x}))U \in \mathcal{F}_{\boldsymbol{y}}] = \mathbb{P}_U[\psi_{\boldsymbol{w}}(\boldsymbol{x}) + \rho(\psi_{\boldsymbol{w}}(\boldsymbol{x}))U(\boldsymbol{x}) \in \mathcal{F}_{\boldsymbol{y}}]$ where $U(\boldsymbol{x})$ is uniformly distributed on the Euclidean sphere $\mathbb{R}^{d(\boldsymbol{x})}$.

Note that

$$\begin{aligned}
&\mathcal{R}_\lambda(h_{\boldsymbol{w}}) - \mathcal{R}(h_{\boldsymbol{w}}) \\
&= \mathbb{E}_X\Big[\sum_{\boldsymbol{y} \in \mathcal{Y}(X)} p_\lambda(\boldsymbol{y}|\psi_{\boldsymbol{w}}(X)) f^0(\boldsymbol{y}, X) - \sum_{\boldsymbol{y} \in \mathcal{Y}(X)} p_0(\boldsymbol{y}|\psi_{\boldsymbol{w}}(X)) f^0(\boldsymbol{y}, X) \Big] \\
&= \mathbb{E}_{X,R,U}\Big[\sum_{\boldsymbol{y} \in \mathcal{Y}(X)} \underbrace{\Big(\mathbb{1}\big(\psi_{\boldsymbol{w}}(\boldsymbol{x})/\lambda + R(\boldsymbol{x})U(\boldsymbol{x}) \in \mathcal{F}_{\boldsymbol{y}}\big) - \mathbb{1}\big(\psi_{\boldsymbol{w}}(\boldsymbol{x}) + \rho(\psi_{\boldsymbol{w}}(\boldsymbol{x}))U(\boldsymbol{x}) \in \mathcal{F}_{\boldsymbol{y}}\big) \Big)}_{A(X,\boldsymbol{y},R,U)} f^0(\boldsymbol{y}, X) \Big] \\
&= \mathbb{E}_{X,R,U}\Big[\sum_{\boldsymbol{y} \in \mathcal{Y}(X)} \Big(\mathbb{1}\big(R(x) \leq \rho(\psi_{\boldsymbol{w}}(x))/\lambda\big) + \mathbb{1}\big(R(x) > \rho(\psi_{\boldsymbol{w}}(x))/\lambda\big) \Big) A(X, \boldsymbol{y}, R, U) f^0(\boldsymbol{y}, X) \Big]
\end{aligned}$$

where we have introduced the random variable $A(X, \boldsymbol{y}, R, U)$. By a zero-sum argument, one can replace $f^0(\boldsymbol{y}, X)$ by $f^0(\boldsymbol{y}, X) - \inf f^0$ in the calculation above.

By definition of ρ, if $R(x) \leq \rho(\psi_{\boldsymbol{w}}(x))/\lambda$ then $\mathbb{1}\big(\psi_{\boldsymbol{w}}(x)/\lambda + R(x)U(x) \in \mathcal{F}_{\boldsymbol{y}}\big)$ is equal to $\mathbb{1}\big(\psi_{\boldsymbol{w}}(\boldsymbol{x}) + \rho(\psi_{\boldsymbol{w}}(\boldsymbol{x}))U \in \mathcal{F}_{\boldsymbol{y}}\big)$. Hence $A(X, \boldsymbol{y}, R, U)\mathbb{1}\big(R(x) \leq \rho(\psi_{\boldsymbol{w}}(x))/\lambda\big) = 0$.

We therefore have

$$
\begin{aligned}
\big|\mathcal{R}_\lambda(h_{\boldsymbol{w}}) - \mathcal{R}(h_{\boldsymbol{w}})\big| &= \big|\mathbb{E}_{X,R,U}\big[\sum_{\boldsymbol{y}\in\mathcal{Y}(X)} \mathbb{1}\big(R(X) > \rho(\psi_{\boldsymbol{w}}(X))/\lambda\big) A(X, \boldsymbol{y}, R, U)(f^0(\boldsymbol{y}, X) - \inf f^0)\big]\big| \\
&\leq \mathbb{E}_{X,R,U}\big[\sum_{\boldsymbol{y}\in\mathcal{Y}(X)} \mathbb{1}\big(R(x) > \rho(\psi_{\boldsymbol{w}}(x))/\lambda\big)|A(X, \boldsymbol{y}, R, U)|(f^0(\boldsymbol{y}, X) - \inf f^0)\big] \\
&\leq \mathbb{E}_{X,R,U}\big[\operatorname{osc}(f^0)\mathbb{1}\big(R(X) > \rho(\psi_{\boldsymbol{w}}(X))/\lambda\big) \underbrace{\sum_{\boldsymbol{y}\in\mathcal{Y}(X)} |A(X, \boldsymbol{y}, R, U)|}_{\leq 2}\big] \\
&\leq 2\operatorname{osc}(f^0)\mathbb{E}_X\mathbb{P}_R\big(R(X) > \rho(\psi_{\boldsymbol{w}}(X))/\lambda | X\big) = 2\operatorname{osc}(f^0)V_{\boldsymbol{w}}(\lambda).
\end{aligned}
$$

Using that the function $\lambda \mapsto V_{\boldsymbol{w}}(\lambda) \in [0, 1]$ is non-decreasing, we have

$$
\begin{aligned}
\big|\mathcal{R}_\lambda(h_{\boldsymbol{w}}) - \mathcal{R}_{\varepsilon_0}(h_{\boldsymbol{w}})\big| &\leq \big|\mathcal{R}_\lambda(h_{\boldsymbol{w}}) - \mathcal{R}(h_{\boldsymbol{w}})\big| + \big|\mathcal{R}(h_{\boldsymbol{w}}) - \mathcal{R}_{\varepsilon_0}(h_{\boldsymbol{w}})\big| \\
&\leq 2\operatorname{osc}(f^0)V_{\boldsymbol{w}}(\lambda) + 2\operatorname{osc}(f^0)V_{\boldsymbol{w}}(\varepsilon_0) \\
&\leq 4\operatorname{osc}(f^0)V_{\boldsymbol{w}}(\lambda)\,,
\end{aligned}
$$

which concludes the proof. □

Corollary 4. *Under Condition (Sol), it holds that*

$$\mathcal{R}_\lambda(\mathcal{H}) - \mathcal{R}(\mathcal{H}) \leq 2\operatorname{osc}(f^0)V_{\boldsymbol{w}^*}(\lambda)\,.$$

Proof. Observe that $\mathcal{R}(\mathcal{H}) = \mathbb{E}_X\big[\sum_{\boldsymbol{y}\in\mathcal{Y}(X)} p_0(\boldsymbol{y}|\psi_{\boldsymbol{w}^*}(X))f^0(\boldsymbol{y}, X)\big]$ and remark that it holds

$$\mathcal{R}_\lambda(\mathcal{H}) \leq \mathbb{E}_X\big[\sum_{\boldsymbol{y}\in\mathcal{Y}(X)} p_\lambda(\boldsymbol{y}|\psi_{\boldsymbol{w}^*}(X))f^0(\boldsymbol{y}, X)\big]\,,$$

which gives the result by substracting the first equality. □

Combining Proposition 3 and Proposition 10 we get the following theorem.

Theorem 2. *Under ($\mathrm{UW}_{\varepsilon_0}$) and (Gauss), the following holds true. Let $\varepsilon_0 \geq 0$ and $\lambda > 0$ be such that $\lambda \geq \varepsilon_0$. Let $\tau \in (0, 1)$. There exists a constant $C > 0$ (that may depend on the constant $C_{\varepsilon_0,\tau}$, appearing in ($\mathrm{UW}_{\varepsilon_0}$), and $\operatorname{osc}(f^0)$) such that, for all $\boldsymbol{w} \in \mathcal{W}$,*

$$\big|\mathcal{R}_\lambda(h_{\boldsymbol{w}}) - \mathcal{R}_{\varepsilon_0}(h_{\boldsymbol{w}})\big| \leq C\,\lambda^\tau \operatorname{polylog}(\lambda)\,,$$

where $\operatorname{polylog}(\lambda)$ *is a polynomial logarithm term.*

Proof. Assumption (Mom.) is implied by Assumption ($\mathrm{UW}_{\varepsilon_0}$), see Proposition 3 and Proposition 10 of Appendix B. □

3.4 Empirical process control

We consider the empirical regularized risk (2.3) and

$$\mathcal{R}_{n,\lambda}(\mathcal{H}) := \min_{\boldsymbol{w}\in\mathcal{W}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_Z\big[f^0\big(\hat{\boldsymbol{y}}(\psi_{\boldsymbol{w}}(X_i) + \lambda Z(X_i)), X_i\big)\big] \right\}. \tag{3.6}$$

We would like to control the random variable

$$\Delta_n = \sup_{\boldsymbol{w}\in\mathcal{W}} \left| \mathcal{R}_{n,\lambda}(\boldsymbol{w}) - \mathcal{R}_{\lambda}(\boldsymbol{w}) \right|.$$

We are going to use a Bernstein inequality for bounded sub-Gaussian random variables. In this analysis, we use the oscillation of f^0, recalled in Assumption (Part), and the Dudley's entropy integral, defined as

$$\mathcal{I}_{\mathcal{W}} := \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{W}, \|\cdot\|, \varepsilon)}\, \mathrm{d}\varepsilon\,, \tag{3.7}$$

where $\mathcal{N}(\mathcal{W}, \|\cdot\|, \varepsilon)$ is ε-covering number of $\mathcal{W}$ with respect to $\|\cdot\|$, for further details the reader may refer to [Wainwright, 2019, Chapter 5].

Remark 10 (Example)**.** *Let $\mathbb{B}$ be the unit ball of $\|\cdot\|$ and assume that there exists $R > 0$ such that $\mathcal{W} \subset R\mathbb{B}$. The volume ratios lemma [Wainwright, 2019, Lemma 5.7] states that*

$$\mathcal{I}_{\mathcal{W}} \leq C d \log(R + 1/R)\,,$$

where $C > 0$ is a universal constant.

Theorem 3. *Let Assumptions (Lip) and (Gauss) hold. Assume that $d(\mathcal{X}) := \max_{\boldsymbol{x}\in\mathcal{X}} d(\boldsymbol{x}) < \infty$ (which is implied by (Part)). Then, for all $\delta \in (0,1)$, it holds that*

$$\Delta_n = \sup_{\boldsymbol{w}\in\mathcal{W}} \left| \mathcal{R}_{n,\lambda}(\boldsymbol{w}) - \mathcal{R}_{\lambda}(\boldsymbol{w}) \right| \leq \frac{\operatorname{osc}(f^0)}{\lambda\sqrt{n}} \left((\ln 2)^{-3/4}\, L_{\mathcal{W}}\, \mathcal{I}_{\mathcal{W}} \sqrt{d(\mathcal{X})} + 4\sqrt{\ln\frac{8}{\delta}} \right), \tag{3.8}$$

with probability higher than $1-\delta$.

Proof. Bounding Δ_n is related to studying the zero-mean random variables

$$V_{i,\boldsymbol{w}} := \frac{1}{n}\Big[\sum_{\boldsymbol{y}\in\mathcal{Y}(X_i)} p_\lambda(\boldsymbol{y}|\psi_{\boldsymbol{w}}(X_i)) f^0(\boldsymbol{y}, X_i) - \mathbb{E}_{(X,Y)}\big[p_\lambda(Y|\psi_{\boldsymbol{w}}(X)) f^0(Y,X)\big]\Big]\,, \tag{3.9}$$

which depend on the random samples $X_i \sim \mathbb{P}_X$. Indeed $\Delta_n = \sup_{\boldsymbol{w}\in\mathcal{W}} \left|\sum_{i\in[n]} V_{i,w}\right|$ for a given draw. Rather than studying the absolute value, we focus on $Z := \sup_{\boldsymbol{w}\in\mathcal{W}} \sum_{i\in[n]} V_{i,\boldsymbol{w}}$ and use a symmetrization argument. Since $\Delta_n = \max(\sup_{\boldsymbol{w}\in\mathcal{W}} \sum_{i\in[n]} V_{i,\boldsymbol{w}}, -\inf_{\boldsymbol{w}\in\mathcal{W}} \sum_{i\in[n]} V_{i,\boldsymbol{w}})$, focusing on the first term, the second one being dealt similarly, we will show that Z is close to its mean $\mathbb{E}Z$ with high probability (through a Bernstein inequality) and then bound the mean $\mathbb{E}Z$ by $\mathcal{O}(1/\sqrt{n})$ (through Hoeffding and Dudley's inequalities).

Let $(X_i') \sim \mathbb{P}_X^{\otimes n}$ be drawn independently from $(X_i)_{i\in[n]}$ and define $V_{i,\boldsymbol{w}}'$ as in Equation (3.9). Then $(V_{i,\boldsymbol{w}})_{i\in[n]}$ and $(V_{i,\boldsymbol{w}}')_{i\in[n]}$ are independent with the same distribution.

Lemma 5. *We have that almost surely*

$$\sup_{\boldsymbol{w}\in\mathcal{W}} \sum_{i\in[n]} (V_{i,\boldsymbol{w}} - V_{i,\boldsymbol{w}}')^2 \leq \frac{(\operatorname{osc}(f^0))^2}{n}\,. \tag{3.10}$$

Proof. We use the fact that $p_\lambda \in [0,1]$ to obtain that

$$\begin{aligned} V_{i,\boldsymbol{w}} - V'_{i,\boldsymbol{w}} &= \frac{1}{n}\Big[\sum_{\boldsymbol{y}\in\mathcal{Y}(X_i)} p_\lambda(\boldsymbol{y}|\psi_{\boldsymbol{w}}(X_i)) f^0(\boldsymbol{y}, X_i) - \sum_{\boldsymbol{y}\in\mathcal{Y}(X'_i)} p_\lambda(\boldsymbol{y}|\psi_{\boldsymbol{w}}(X'_i)) f^0(y, x'_i)\Big]\Big] \\ &\le \frac{1}{n}\Big[\sup_{\boldsymbol{x},\boldsymbol{x}'\in\mathcal{X},\, \boldsymbol{y},\boldsymbol{y}'\in\mathcal{Y}} f^0(\boldsymbol{y},\boldsymbol{x}) - f^0(\boldsymbol{y}',\boldsymbol{x}')\Big]\Big] \\ &= \frac{\operatorname{osc}(f^0)}{n}\,, \end{aligned}$$

Since $(V'_{i,\boldsymbol{w}} - V_{i,\boldsymbol{w}})$ is equal in law to $(V_{i,\boldsymbol{w}} - V'_{i,\boldsymbol{w}})$, this yields (3.10). □

Owing to Lemma 5, we can thus apply Boucheron et al. [2013, Theorem 12.3, Page 333] since

$$\mathbb{E}_{X'_i\sim\mathbb{P}_X}[\sup_{\boldsymbol{w}\in\mathcal{W}} \sum_{i\in[n]} (V_{i,\boldsymbol{w}} - V'_{i,\boldsymbol{w}})^2] \le \frac{\operatorname{osc}(f^0)}{n}\,.$$

Consequently, for all $t \ge 0$, we have

$$\mathbb{P}[Z \ge \mathbb{E}Z + t] \le 4e^{-\frac{nt^2}{16(\operatorname{osc}(f^0))^2}}\,. \tag{3.11}$$

We are now going to show that $\mathbb{E}Z = \mathcal{O}(1/\sqrt{n})$ through a chaining argument. Fix $\boldsymbol{w}, \boldsymbol{w}' \in \mathcal{W}$. By Lemma 9, note that $\boldsymbol{w} \mapsto \sum_{\boldsymbol{y}\in\mathcal{Y}(\boldsymbol{x})} p_\lambda(\boldsymbol{y}|\psi_{\boldsymbol{w}}(\boldsymbol{x}))$ is Lipschitz with constant $L_{\mathcal{W}}\sqrt{d(\boldsymbol{x})}/\lambda$. We obtain that for all $i \in [n]$, as $\frac{1}{n}\sum_{\boldsymbol{y}\in\mathcal{Y}(X_i)} p_\lambda(\boldsymbol{y}|\psi_{\boldsymbol{w}}(X_i)) = 1$,

$$\begin{aligned} V_{i,\boldsymbol{w}} - V_{i,\boldsymbol{w}'} &= \frac{1}{n} \sum_{\boldsymbol{y}\in\mathcal{Y}(X_i)} (p_\lambda(\boldsymbol{y}|\psi_{\boldsymbol{w}}(X_i)) - p_\lambda(\boldsymbol{y}|\psi_{\boldsymbol{w}'}(X_i)) f^0(\boldsymbol{y}, X_i) \\ &= \frac{1}{n} \sum_{\boldsymbol{y}\in\mathcal{Y}(X_i)} (p_\lambda(\boldsymbol{y}|\psi_{\boldsymbol{w}}(X_i)) - p_\lambda(\boldsymbol{y}|\psi_{\boldsymbol{w}'}(X_i))(f^0(\boldsymbol{y}, X_i) - \inf_{\boldsymbol{x}',\boldsymbol{y}'} f^0(\boldsymbol{y}',\boldsymbol{x}')) \\ &\le \frac{1}{n} L_{\mathcal{W}} \frac{\sqrt{d(\mathcal{X})}}{\lambda} \operatorname{osc}(f^0) \|\boldsymbol{w} - \boldsymbol{w}'\| =: a_n\,, \end{aligned}$$

and similarly $V_{i,\boldsymbol{w}'} - V_{i,\boldsymbol{w}} \le a_n$. For $V_{\boldsymbol{w}} := \sum_{i\in[n]} V_{i,\boldsymbol{w}}$, we deduce that the random variable $(V_{\boldsymbol{w}} - V_{\boldsymbol{w}'})$ is sub-Gaussian as a sum of zero-mean independent sub-Gaussians [Vershynin, 2018, Proposition 2.6.1, Page 29], and it holds that for the associated ψ_2-norm [Vershynin, 2018, Example 2.5.8, Page 28],

$$\|V_{\boldsymbol{w}} - V_{\boldsymbol{w}'}\|_{\psi_2} \le \Big[\frac{1}{\sqrt{\ln 2}} \sum_{i\in[n]} \|V_{i,\boldsymbol{w}} - V_{i,\boldsymbol{w}'}\|^2_{\psi_2}\Big]^{1/2} \le \frac{a_n\sqrt{n}}{(\ln 2)^{1/4}} = \frac{L_{\mathcal{W}}\sqrt{d(\mathcal{X})}}{(\ln 2)^{1/4}\lambda\sqrt{n}} \operatorname{osc}(f^0)\|\boldsymbol{w} - \boldsymbol{w}'\|\,,$$

with, for a sub-Gaussian X, the ψ_2-norm is defined as $\|X\|_{\psi_2} := \inf\{t \ge 0 \,|\, \mathbb{E}e^{X^2/t^2} \le 2\}$. In other words, we have shown that the zero-mean random variable $V_{\boldsymbol{w}}$ has sub-Gaussian increments. We can thus apply Dudley's inequality [Vershynin, 2018, Theorem 8.1.3, Page 188],

$$\mathbb{E}V = \mathbb{E}[\sup_{\boldsymbol{w}\in\mathcal{W}} V_{\boldsymbol{w}}] \le \frac{L_{\mathcal{W}}\sqrt{d(\mathcal{X})}}{(\ln 2)^{3/4}\lambda\sqrt{n}} \operatorname{osc}(f^0)\,\mathcal{I}_{\mathcal{W}}\,.$$

We proceed similarly with $\inf_{\boldsymbol{w}\in\mathcal{W}} \sum_{i\in[n]} V_{i,\boldsymbol{w}}$. We obtain that, setting $t = \sqrt{\frac{16(\operatorname{osc}(f^0))^2}{n} \ln\frac{4}{\delta}}$ in (3.11), with probability $1 - 2\delta$,

$$\Delta_n = \max\Big\{ \sup_{\boldsymbol{w}\in\mathcal{W}} \sum_{i\in[n]} V_{i,\boldsymbol{w}}, - \inf_{\boldsymbol{w}\in\mathcal{W}} \sum_{i\in[n]} V_{i,\boldsymbol{w}} \Big\} \le \mathbb{E}V + t$$

Using the bound on $\mathbb{E}V$, this yields (3.8). □

3.5 Optimization error bound on the Kernel-SoS case

The last part of the proof consists in bounding the optimization error when minimizing $\mathcal{R}_{n,\lambda}$. Up to our knowledge, many global optimization algorithms converge because they provably do an exhaustive search, as is the case for DIRECT for instance [Jones et al., 1993, Section 5]. Consequently, their convergence bounds are most often affected by the curse of dimensionality. Here we have an advantage since $\mathcal{R}_{n,\lambda}$ is a smooth function thanks to the parameter $\lambda > 0$. This incites us to use a specific kernel-based approach recently suggested in Rudi et al. [2024] to leverage the smoothness of the function $\mathcal{R}_{n,\lambda}$. To alleviate notation, we set

$$R_{n,\lambda}(\boldsymbol{w}) := \mathcal{R}_{n,\lambda}(h_{\boldsymbol{w}}).$$

The key idea will be to find an (approximate) Sum-of-Squares representation for $R_{n,\lambda}$ while minimizing $R_{n,\lambda}$. Fix $s > 1 + d/2$ and let $\mathcal{H}_\phi$ be the Sobolev space over $\mathcal{W}$ of smoothness s. Denote by k its reproducing kernel, i.e. $k(\cdot, \boldsymbol{w}) \in \mathcal{H}_\phi$ such that for all $g \in \mathcal{H}_\phi, \boldsymbol{w} \in \mathcal{W}$, we have $g(\boldsymbol{w}) = \langle g, k(\cdot, \boldsymbol{w})\rangle_{\mathcal{H}_\phi}$. Let $\phi : \mathcal{W} \to \mathcal{H}_\phi$ be the kernel embedding, i.e. $\phi(\boldsymbol{w}) = k(\cdot, \boldsymbol{w})$. Set $\tilde{s} = s - d/2 \geq 1$. Let $S^+(\mathcal{H}_\phi)$ be the set of self-adjoint positive semidefinite operators from $\mathcal{H}_\phi$ to $\mathcal{H}_\phi$.

Assumption (kSoS)**.** *We assume that $\mathcal{W}$ is the closure of a union of closed balls $\cup_{\boldsymbol{w}\in\mathcal{S}}\mathbb{B}(\boldsymbol{w}, r)$ for some $r > 0$ and a bounded set $\mathcal{S} \subset \mathbb{R}^d$. Assume furthermore that $R_{n,\lambda} \in C^{s+3}(\mathcal{W}, \mathbb{R})$ and that $R_{n,\lambda}$ has a kernel Sum-of-Square representation, i.e. there exists $A_{n,\lambda} \in S^+(\mathcal{H}_\phi)$ such that*

$$R_{n,\lambda}(\boldsymbol{w}) - \min_{\boldsymbol{w}'\in\mathcal{W}} R_{n,\lambda}(\boldsymbol{w}') = \langle \phi(\boldsymbol{w}), A_{n,\lambda}\phi(\boldsymbol{w})\rangle_{\mathcal{H}_\phi}, \ \forall \boldsymbol{w} \in \mathcal{W}. \tag{3.12}$$

For the following, we fix the $A_{n,\lambda}$ with minimal trace satisfying (3.12).

The requirement on $\mathcal{W}$ is to use scattering bounds and is quite mild for compact sets. We mostly ensure $R_{n,\lambda} \in C^{s+3}(\mathcal{W}, \mathbb{R})$ by regularizing $R_{n,0}$ through $\lambda > 0$ and using a smooth feature map $\boldsymbol{w} \mapsto \psi_{\boldsymbol{w}}$, e.g. $\psi_{\boldsymbol{w}}(\boldsymbol{x}) = \Phi_x \boldsymbol{w}$ with $\Phi_x \in \mathbb{R}^{d(x)\times d}$. By Rudi et al. [2024, Theorem 3] (see also Marteau-Ferey et al. [2024]), the existence of $A_{n,\lambda}$ is for instance guaranteed if $R_{n,\lambda}$ and its second order partial derivatives belong to $\mathcal{H}_\phi$, and $R_{n,\lambda}$ has a finite number of minimizers, all with positive Hessian. We will see an explicit example below.

Recall that a global optimization problem can always be cast as a convex one under the form

$$\begin{aligned} \max_{c\in\mathbb{R},} \quad & c \\ \text{s.t.} \quad & c \leq R_{n,\lambda}(\boldsymbol{w}), \ \forall\, \boldsymbol{w} \in \mathcal{W}. \end{aligned}$$

We consider the kernel Sum-Of-Squares approximation of it, namely

$$\min_{\substack{c\in\mathbb{R},\\ A\in S^+(\mathcal{H}_\phi)}} \quad -c + \lambda_\phi \operatorname{Tr}(A) \tag{3.13a}$$

$$\text{s.t.} \quad R_{n,\lambda}(\tilde{\boldsymbol{w}}_m) - c = \langle \phi(\tilde{\boldsymbol{w}}_m), A\phi(\tilde{\boldsymbol{w}}_m)\rangle_{\mathcal{H}_\phi}, \ \forall\, m \in [M] \tag{3.13b}$$

where $\lambda_\phi > 0$ is some parameter, the $(\tilde{\boldsymbol{w}}_m)_{m\in[M]}$ are sampled uniformly at random in $\mathcal{W}$.

Remark 11. *A few words are in order. We want a $c \in \mathbb{R}$ such that $w \mapsto R_{n,\lambda}(\boldsymbol{w}) - c \geq 0$. On the other hand, for any $A \in S^+(\mathcal{H}_\phi)$, the function $w \mapsto \langle \phi(\boldsymbol{w}), A\phi(\boldsymbol{w})\rangle_{\mathcal{H}_\phi}$ is nonnegative, and under some assumptions on ϕ and $\mathcal{H}_\phi$ [Marteau-Ferey et al., 2020], these functions are pointwise dense in the nonnegative continuous functions. So we look for one such A for which the two functions are equal at least at the sampled points and we penalize the extra variable A by its trace. For (3.13), the Newton-based algorithm detailed in Rudi et al. [2024, Section 6] outputs the infimum value estimate $\hat{R} = \hat{c}$, the corresponding $\hat{A}$ and a candidate $\hat{\boldsymbol{w}}_{n,\lambda} = \sum_{m\in[M]} \alpha_m \tilde{\boldsymbol{w}}_m$ where each α_m is the Lagrange multiplier of the equality constraint (3.13b) at $\tilde{\boldsymbol{w}}_m$. Contrary to Rudi et al. [2024], an extra term $|R_{n,\lambda}(\hat{\boldsymbol{w}}_{n,\lambda}) - \hat{R}|$ appears in our analysis. One expects that this error goes to zero as M grows large, and we referred to it as a posteriori error.*

Theorem 4. *Fix $\delta \in (0,1)$. With the above notation, under Assumption (kSoS), there exists explicit constants $\bar{M} \in \mathbb{N}$ and $\bar{C} \in \mathbb{R}_+$ depending on $s, d, \delta, r, \mathrm{diam}(\mathcal{W})$ such that for $M \geq \bar{M}$ and $\lambda_\phi \geq \bar{C} M^{-\tilde{s}/d} \left(\log \frac{M}{\delta}\right)^{\tilde{s}/d}$ we have, with probability at least $1-\delta$,*

$$|R_{n,\lambda}(\hat{\boldsymbol{w}}_{n,\lambda}) - R_{n,\lambda}(\boldsymbol{w}_{n,\lambda})| \leq R_{n,\lambda}(\hat{\boldsymbol{w}}_{n,\lambda}) - \hat{R} + \lambda_\phi \left(\mathrm{Tr}\left(A_{n,\lambda}\right) + |R_{n,\lambda}|_{\mathcal{W},\lceil\tilde{s}\rceil}\right), \tag{3.14}$$

depending on the Sobolev norm $|R_{n,\lambda}|_{\mathcal{W},\lceil\tilde{s}\rceil} = \sup_{|\alpha|=\tilde{s}} \sup_{\boldsymbol{w}'\in\mathcal{W}} |D^\alpha R_{n,\lambda}(\boldsymbol{w}')|$.

The estimate (3.14) has the advantage that if $s \gg d$ then the smoothness mitigates the curse of dimensionality. Indeed, to achieve an a priori error ϵ on the last term of (3.14), which we denote $\lambda_\phi C_{n,\lambda}$, we can take $M \propto \epsilon^{-d/\tilde{s}}$ rather than the more usual, and much larger, $M \propto \epsilon^{-d}$. Note however that the constants $\bar{M}$ and $\bar{C}$ can still depend exponentially in the dimension, their closed-form expressions can be found in Rudi et al. [2024, Theorem 5]. Furthermore the Sobolev norm $|R_{n,\lambda}|_{\mathcal{W},\lceil s-d/2\rceil}$ and the cost of the Sum-of-Square representation $\mathrm{Tr}\left(A_{n,\lambda}\right)$ depend crucially on the noise parameter λ since the latter plays a role on the regularity of $R_{n,\lambda}$, and we expect both terms to diverge when λ goes to 0. On the other hand, as just said, λ_ϕ can be taken very small when the number of sampling points M is large, so that it compensates the other terms. In other words the computational complexity, which is polynomial in M, of solving larger SDP problems will offset the cases when λ is taken small and the function $R_{n,\lambda}$ is less smooth. We face thus a classical tradeoff between having a better estimate of $R_{n,0}$ or doing fewer computations.

The case of generalized linear embeddings

Assume $\boldsymbol{\theta} = \psi_{\boldsymbol{w}}(\boldsymbol{x}) = \Phi_x \boldsymbol{w}$ with $\Phi_x \in \mathbb{R}^{d(x)\times d}$ and the added noise $Z(x)$ is Gaussian with zero mean, then we can provide more explicit estimates for $\mathrm{Tr}\left(A_{n,\lambda}\right)$ and $|R_{n,\lambda}|_{\mathcal{W},\lceil s-d/2\rceil}$.

Decompose $Z(x) = \Phi_x Z'(x) + Z''(x)$ with Z' Gaussian over $\mathbb{R}^d$ with covariance Σ_x and Z'' Gaussian with values in the complement $\mathrm{Im}(\Phi_x)^\perp$ to the range of Φ_x. Then

$$\mathbb{E}_{Z(x)}\left[f^0(\hat{\boldsymbol{y}}(\psi_{\boldsymbol{w}}(\boldsymbol{x}) + \lambda Z(\boldsymbol{x})), \boldsymbol{x})\right] = \mathbb{E}_{Z'(x)}\mathbb{E}_{Z''(x)}\left[f^0(\hat{\boldsymbol{y}}(\Phi_x(\boldsymbol{w} + Z'(x)) + \lambda Z''(x)), \boldsymbol{x})\right] \tag{3.15}$$

Let $\tilde{R}_{n,\lambda,x} = \mathbb{E}_{Z''(x)}\left[f^0(\hat{\boldsymbol{y}}(\Phi_x \boldsymbol{w} + \lambda Z''(x)), \boldsymbol{x})\right]$. We then observe that the objective function is a sum of Gaussian convolutions

$$R_{n,\lambda}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^n \tilde{R}_{n,\lambda,x_i} \star \mathcal{N}_{\lambda\Sigma_{x_i}}(\boldsymbol{w}) \tag{3.16}$$

so $R_{n,\lambda}$ is smooth. Moreover the convolution commutes with the derivation, and $|g \star g'|_\infty \leq |g|_\infty |g'|_1$ for any functions g, g'. Moreover $|R_{n,\lambda,x_i}|_\infty \leq |f^0|_\infty$, whence for $\tilde{s} = s - d/2$,

$$\begin{aligned}
|R_{n,\lambda}|_{\mathcal{W},\lceil\tilde{s}\rceil} &\leq \frac{1}{n}\sum_{i=1}^n \sup_{|\alpha|=\tilde{s}} \sup_{\boldsymbol{\theta}\in\Omega} |\underbrace{D^\alpha R_{n,\lambda,x_i} \star \mathcal{N}_{\lambda\Sigma_{x_i}}(\boldsymbol{w})}_{= R_{n,\lambda,x_i} \star D^\alpha \mathcal{N}_{\lambda\Sigma_{x_i}}}| \\
&\leq |f^0|_\infty \sup_{i\in[n],|\alpha|=\tilde{s}} |D^\alpha \mathcal{N}_{\lambda\Sigma_{x_i}}|_1 \\
&= \mathcal{O}\left(\lambda^{-\tilde{s}}\right).
\end{aligned}$$

Moreover, defining $\mathcal{M}_{\lambda,i} = \sqrt{\mathcal{N}_{\lambda\Sigma_{x_i}}}$, clearly $\mathcal{M}_{\lambda,i} \in \mathcal{H}_\phi$ and we can obtain a closed-form sum-of-squares expression of $R_{n,\lambda}$ for the chosen kSoS feature map ϕ, so that we justify the assumption (3.12). Indeed by the reproducing property, we have

$$\begin{aligned}
R_{n,\lambda}(\boldsymbol{w}) &= \frac{1}{n}\sum_{i=1}^n \left\langle \phi(\boldsymbol{w}), \left[\int_{\mathbb{R}^d} \mathcal{M}_{\lambda,i}(\boldsymbol{w}', \cdot) \otimes \mathcal{M}_{\lambda,i}(\boldsymbol{w}', \cdot) R_{n,0}(\boldsymbol{w}') d\boldsymbol{w}'\right] \phi(\boldsymbol{w}) \right\rangle_{H_\phi} \\
&=: \left\langle \phi(\boldsymbol{w}), \tilde{A}_{n,\lambda}\phi(\boldsymbol{w}) \right\rangle_{H_\phi}.
\end{aligned}$$

and since $A_{n,\lambda}$ was chosen as the operator with minimal trace

$$\mathrm{Tr}(A_{n,\lambda}) \leq \mathrm{Tr}(\tilde{A}_{n,\lambda}) = \int_{\mathbb{R}^d} \langle \mathcal{M}_{\lambda,i}(\boldsymbol{w}', \cdot), \mathcal{M}_{\lambda,i}(\boldsymbol{w}', \cdot) \rangle_{H_\phi} R_{n,0}(\boldsymbol{w}') d\boldsymbol{w}' .$$

Since the RKHS associated with ϕ is the Sobolev space of order s, then we have a characterization of the norm through the Fourier transform

$$\langle \mathcal{M}_{\lambda,i}(\boldsymbol{w}', \cdot), \mathcal{M}_{\lambda,i}(\boldsymbol{w}', \cdot) \rangle_{H_\phi} = \int_{\mathbb{R}^d} (1 + |\xi|^2)^s |\hat{\mathcal{M}}_{\lambda,i}(\boldsymbol{w}', \xi)|^2 d\xi = \mathcal{O}(\lambda^{-\tilde{s}}).$$

Example 2 (continued). *The stochastic vehicle scheduling problem (StoVSP) is an instance of generalized linear embedding. As such, the previous calculation holds.*

References

EURO Meets NeurIPS 2022 Vehicle Routing Competition. https://euro-neurips-vrp-2022.challenges.ortec.com/.

Maria-Florina Balcan. Data-driven algorithm design. In Tim Roughgarden, editor, Beyond the Worst-Case Analysis of Algorithms, chapter 28, pages 626–645. Cambridge University Press, 2021.

Maria-Florina Balcan, Dan DeBlasio, Travis Dick, Carl Kingsford, Tuomas Sandholm, and Ellen Vitercik. How much data is sufficient to learn high-performing algorithms? generalization guarantees for data-driven algorithm design. In Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing, pages 919–932. ACM, 2021. ISBN 978-1-4503-8053-9.

Léo Baty, Kai Jungel, Patrick S. Klein, Axel Parmentier, and Maximilian Schiffer. Combinatorial Optimization-Enriched Machine Learning to Solve the Dynamic Vehicle Routing Problem with Time Windows. Transportation Science, 2024.

Yoshua Bengio, Andrea Lodi, and Antoine Prouvost. Machine learning for combinatorial optimization: A methodological tour d'horizon. European Journal of Operational Research, 290(2):405–421, 2021. ISSN 0377-2217.

Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, and Francis Bach. Learning with differentiable perturbed optimizers. In Advances in Neural Information Processing Systems 33 (NeurIPS), pages 9508–9519, 2020.

Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. The Annals of Statistics, 44(2):813–852, 2016. ISSN 00905364.

Mathieu Blondel, André F. T. Martins, and Vlad Niculae. Learning with fenchel-young losses. J. Mach. Learn. Res., 21:35:1–35:69, 2020.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. Concentration inequalities: a nonasymptotic theory of independence. Oxford University Press, 1st ed edition, 2013. ISBN 9780199535255.

Djalil Chafaï, Olivier Guédon, Guillaume Lecué, and Alain Pajor. Interactions between compressed sensing random matrices and high dimensional geometry, volume 37. Société Mathématique de France Paris, 2012.

Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A consistent regularization approach for structured prediction. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, pages 4412–4420, 2016.

Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A general framework for consistent structured prediction with implicit loss embeddings. J. Mach. Learn. Res., 21:98:1–98:67, 2020.

Guillaume Dalle, Léo Baty, Louis Bouvier, and Axel Parmentier. Learning with combinatorial optimization layers: a probabilistic approach. ArXiv preprint, abs/2207.13513, 2022.

Adam N. Elmachtoub and Paul Grigas. Smart "Predict, then Optimize". Management Science, 2021. ISSN 0025-1909.

Evarist Giné and Richard Nickl. Mathematical foundations of infinite-dimensional statistical models. Cambridge university press, 2021.

Rishi Gupta and Tim Roughgarden. A pac approach to application-specific algorithm selection. In Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science, pages 123–134, 2016.

D. R. Jones, C. D. Perttunen, and B. E. Stuckman. Lipschitzian optimization without the lipschitz constant. Journal of Optimization Theory and Applications, 79(1):157–181, 1993. ISSN 1573-2878.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. Annals of statistics, pages 1302–1338, 2000.

Ulysse Marteau-Ferey, Francis R. Bach, and Alessandro Rudi. Non-parametric models for non-negative functions. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Second order conditions to decompose smooth functions as sums of squares. SIAM Journal on Optimization, 34(1):616–641, 2024.

Alex Nowak-Vila, Francis Bach, and Alessandro Rudi. A general theory for structured prediction with smooth convex surrogates. ArXiv preprint, abs/1902.01958, 2019.

Sebastian Nowozin and Christoph H. Lampert. Structured Learning and Prediction in Computer Vision. Foundations and Trends® in Computer Graphics and Vision, 6(3–4):185–365, 2011. ISSN 1572-2740, 1572-2759.

Axel Parmentier. Learning to Approximate Industrial Problems by Operations Research Classic Problems. Operations Research, 2021a. ISSN 0030-364X.

Axel Parmentier. Learning structured approximations of operations research problems. ArXiv preprint, abs/2107.04323, 2021b.

Axel Parmentier and Vincent T'Kindt. Learning to solve the single machine scheduling problem with release times and sum of completion times. ArXiv preprint, abs/2101.01082, 2021.

Marin Vlastelica Pogancic, Anselm Paulus, Vít Musil, Georg Martius, and Michal Rolinek. Differentiation of blackbox combinatorial solvers. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020.

Meng Qi, Paul Grigas, and Zuo-Jun Max Shen. Integrated conditional estimation-optimization. arXiv preprint arXiv:2110.12351, 2021.

Alessandro Rudi, Ulysse Marteau-Ferey, and Francis Bach. Finding global minima via kernel approximations. Mathematical Programming, 2024. ISSN 1436-4646.

Utsav Sadana, Abhilash Chenreddy, Erick Delage, Alexandre Forel, Emma Frejinger, and Thibaut Vidal. A survey of contextual optimization methods for decision-making under uncertainty. European Journal of Operational Research, 2024.

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

Roman Vershynin. High-dimensional probability: an introduction with applications in data science. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, 2018. ISBN 9781108231596.

Ky Khac Vu, Claudia d'Ambrosio, Youssef Hamadi, and Leo Liberti. Surrogate-based methods for black-box optimization. International Transactions in Operational Research, 24(3):393–424, 2017.

Martin J Wainwright. High-dimensional statistics: A non-asymptotic viewpoint, volume 48. Cambridge university press, 2019.

Supplemental Materials

A Technical lemmas

Lemma 6. *Let $u : \mathbb{R}^d \to \mathbb{R}$ be an integrable bounded function, Z a standard normal random vector of $\mathbb{R}^d$, $\sigma > 0$ a positive real number, and $G(\boldsymbol{w}) := \mathbb{E}_Z[u(\boldsymbol{w} + \sigma Z)]$. Then $\boldsymbol{w} \mapsto G(\boldsymbol{w})$ is $\sqrt{d}\|u\|_\infty/\sigma$-Lipschitz.*

Proof. Let h be the density of σZ. One has $G(\boldsymbol{w}) = \int_{\mathbb{R}^d} u(\boldsymbol{w}+\boldsymbol{v})h(\boldsymbol{v})\mathrm{d}\boldsymbol{v} = \int_{\mathbb{R}^d} u(\boldsymbol{v})h(\boldsymbol{v}-\boldsymbol{w})\mathrm{d}\boldsymbol{v}$. By dominated convergence, it holds that

$$\nabla G(\boldsymbol{w}) = -\int_{\mathbb{R}^d} u(\boldsymbol{v})\nabla h(\boldsymbol{v}-\boldsymbol{w})\mathrm{d}\boldsymbol{v} = -\int_{\mathbb{R}^d} u(\boldsymbol{w}+\boldsymbol{v})\nabla h(\boldsymbol{v})\mathrm{d}\boldsymbol{v}\,.$$

By Jensen inequality and standard calculation with the Gaussian density, one has

$$\|\nabla G(\boldsymbol{w})\| \le \|u\|_\infty \int_{\mathbb{R}^d} \|\nabla h(\boldsymbol{v})\|\mathrm{d}\boldsymbol{v} \le \frac{\sqrt{d}\|u\|_\infty}{\sigma}\,,$$

which gives the result. □

Lemma 7. *Under Assumptions (Lip) and (Gauss), the following holds. When $\varepsilon_0 > 0$, the risk $\boldsymbol{w} \mapsto \mathcal{R}_{\varepsilon_0}(h_{\boldsymbol{w}})$ is continuous with respect to $\boldsymbol{w}$.*

Proof. Let $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y}, \boldsymbol{\theta} \in \mathbb{R}^{d(\boldsymbol{x})}$. Invoke Lemma 6 with $u(\boldsymbol{\theta}) = p_0(\boldsymbol{y}|\boldsymbol{\theta}) \in [0,1]$ to get that $\boldsymbol{\theta} \mapsto \mathbb{E}_Z[\sum_{\boldsymbol{y}\in\mathcal{Y}(\boldsymbol{x})} p_0(\boldsymbol{y}|\boldsymbol{\theta}+\varepsilon_0 Z(\boldsymbol{x}))\, f^0(\boldsymbol{y},\boldsymbol{x})]$ is continuous. Now, note that $\boldsymbol{w} \mapsto \psi_{\boldsymbol{w}}(\boldsymbol{x})$ is $L_{\mathcal{W}}$-Lipschitz by Assumption (Lip), hence $\boldsymbol{w} \mapsto \mathbb{E}_Z[\sum_{\boldsymbol{y}\in\mathcal{Y}(\boldsymbol{x})} p_0(\boldsymbol{y}|\psi_{\boldsymbol{w}}(\boldsymbol{x}) + \varepsilon_0 Z(\boldsymbol{x}))\, f^0(\boldsymbol{y},\boldsymbol{x})]$ is continuous. Taking the expectation, by dominated convergence, we deduce the result. □

Lemma 8. *Let $\lambda > 0$, $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})$. Then, the function $\boldsymbol{\theta} \in \mathbb{R}^{d(\boldsymbol{x})} \mapsto \sum_{\boldsymbol{y}\in\mathcal{Y}(\boldsymbol{x})} p_\lambda(\boldsymbol{y}|\boldsymbol{\theta})$ is $\sqrt{d(\boldsymbol{x})}/\lambda$-Lipschitz, under Assumption (Gauss).*

Proof. Note that $\sum_{\boldsymbol{y}\in\mathcal{Y}(\boldsymbol{x})} p_\lambda(\boldsymbol{y}|\boldsymbol{\theta}) = \mathbb{E}_Z[u(\boldsymbol{\theta}+\lambda Z)]$ where $u(\cdot) = \sum_{\boldsymbol{y}\in\mathcal{Y}(\boldsymbol{x})} p_0(\boldsymbol{y}|\cdot) \in [0,1]$ and apply Lemma 6. □

Lemma 9. *The function $\boldsymbol{w} \in \mathcal{W} \mapsto \sum_{\boldsymbol{y}\in\mathcal{Y}(\boldsymbol{x})} p_\lambda(\boldsymbol{y}|\psi_{\boldsymbol{w}}(\boldsymbol{x})) \in [0,1]$ is $L_{\mathcal{W}}\sqrt{d(\boldsymbol{x})}/\lambda$-Lipschitz, under Assumptions (Lip) and (Gauss).*

Proof. Let $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y}, \boldsymbol{\theta} \in \mathbb{R}^{d(\boldsymbol{x})}$. Invoke Lemma 8 to get that $\boldsymbol{\theta} \mapsto p_\lambda(\boldsymbol{y}|\boldsymbol{\theta})$ is $\sqrt{d(\boldsymbol{x})}/\lambda$-Lipschitz. Now, observe that $\boldsymbol{w} \mapsto \psi_{\boldsymbol{w}}(\boldsymbol{x})$ is $L_{\mathcal{W}}$-Lipschitz to conclude. □

B Perturbation bias with remarks

Proposition 10. *If Assumption (Gauss) holds, then, for all $\lambda \in (0,1)$ and all $\boldsymbol{w} \in \mathcal{W}$,*

$$V_{\boldsymbol{w}}(\lambda) \le \inf_{q\in(0,1)} \left\{ \mathbb{P}_X\Big(\frac{\rho(\psi_{\boldsymbol{w}}(X))}{\sqrt{d(X)}} < \lambda^q\Big) + \exp\Big(-\frac{1}{10\lambda^{2(1-q)}}\Big)\right\}.$$

Furthermore, under the following **Assumption** (Orl.),

$$\exists \tau > 0 \text{ s.t. } \mathbb{E}_X\left[\exp\Big(\big(\frac{\rho(\psi_{\boldsymbol{w}}(X))}{\sqrt{d(X)}}\big)^{-\tau}\Big)\right] < \infty\,, \tag{Orl.}$$

it holds

$$V_{\boldsymbol{w}}(\lambda) = \mathcal{O}_{\lambda\to 0}\Big(\exp(-\lambda^{-\frac{2\tau}{2+\tau}})\Big)\,;$$

And under the following **Assumption** (Mom.),

$$\exists \tau > 0 \text{ s.t. } \mathbb{E}_X\Big[\Big(\frac{\rho(\psi_{\boldsymbol{w}}(X))}{\sqrt{d(X)}}\Big)^{-\tau}\Big] < \infty\,, \tag{Mom.}$$

it holds

$$V_{\boldsymbol{w}}(\lambda) = \mathcal{O}_{\lambda\to 0}\big(\lambda^{\tau}\text{polylog}(\lambda)\big)\,,$$

where $\text{polylog}(\lambda)$ *is a polynomial logarithm term.*

Proof. Invoke [Laurent and Massart, 2000, Lemma 1] to get that, for $\boldsymbol{x} \in \mathcal{X}$, for all $t \geq d(\boldsymbol{x})$,

$$\mathbb{P}[\|Z(\boldsymbol{x})\|_2^2 \geq t] \leq \exp(-0.1\, t)\,,$$

for the Gaussian model (see Assumption ((Gauss))). Hence, for all $\lambda \in (0,1)$, all $q \in (0,1)$,

$$\forall \boldsymbol{x} \text{ s.t. } \rho(\psi_{\boldsymbol{w}}(\boldsymbol{x})) \geq \lambda^q\sqrt{d(\boldsymbol{x})}\,, \quad \mathbb{P}\Big[\|Z(\boldsymbol{x})\|_2 \geq \frac{\rho(\psi_{\boldsymbol{w}}(\boldsymbol{x}))}{\lambda}\Big] \leq \exp\Big(-0.1\,\frac{\rho^2(\psi_{\boldsymbol{w}}(\boldsymbol{x}))}{\lambda^2}\Big)\,.$$

Hence It follows that, for all $\lambda \in (0,1)$ and all $q \in (0,1)$,

$$\begin{aligned}
V(\lambda) &= \mathbb{E}_X\mathbb{P}_Z\Big[\|Z\|_2 \geq \frac{\rho(\psi_{\boldsymbol{w}}(X))}{\lambda}\,\Big|\,X\Big]\,,\\
&\leq \mathbb{P}_X\Big[\rho(\psi_{\boldsymbol{w}}(X)) < \lambda^q\sqrt{d(X)}\Big] + \mathbb{E}_X\Big[\exp\Big(-0.1\,\frac{\rho^2(\psi_{\boldsymbol{w}}(X))}{\lambda^2}\Big)\mathbf{1}_{\{\rho(\psi_{\boldsymbol{w}}(X))\geq\lambda^q\sqrt{d(X)}\}}\Big]\,,\\
&\leq \mathbb{P}_X\Big[\rho(\psi_{\boldsymbol{w}}(X)) < \lambda^q\sqrt{d(X)}\Big] + \mathbb{E}_X\Big[\exp\Big(-\frac{d(X)}{10\lambda^{2(1-q)}}\Big)\Big]\,,\\
&\leq \mathbb{P}_X\Big[\rho(\psi_{\boldsymbol{w}}(X)) < \lambda^q\sqrt{d(X)}\Big] + \exp\Big(-\frac{1}{10\lambda^{2(1-q)}}\Big)\,,
\end{aligned}$$

the last inequality stemming from $d(X) \geq 1$ $\mathbb{P}_X$-almost surely.

Assume (Orl.), by Markov's inequality it holds

$$\mathbb{P}_X\Big[\rho(\psi_{\boldsymbol{w}}(X)) < \lambda^q\sqrt{d(X)}\Big] \leq C\exp\big(-\lambda^{-\tau q}\big)\,.$$

Taking $q = 2/(2+\tau)$ gives the result.

Assume (Mom.), by Markov's inequality it holds

$$\mathbb{P}_X\Big[\rho(\psi_{\boldsymbol{w}}(X)) < \lambda^q\sqrt{d(X)}\Big] \leq C\,\lambda^{\tau q}\,.$$

Taking $q = 1 - \log(-10\tau\log\lambda)/(-2\log(\lambda))$ gives the result (note that $q < 1$, and $q > 0$ for λ sufficient small). □

Remark 12 (On the conditions (Orl.) and (Mom.))**.** *These conditions describe the tail behavior of the random variable* $R := \sqrt{d(X)}/\rho(\psi_{\boldsymbol{w}}(X))$*. Condition* (Orl.) *is of Orlicz-norm type, it states that* R *belongs to* τ*-Orlicz space and, as such as, it has* τ*-exponentially light tail, see for instance [Chafaï et al., 2012, Section 1.1] for a review on Orlicz spaces. Hence, the distribution of the random variable* $\rho(\psi_{\boldsymbol{w}}(X))/\sqrt{d(X)} = 1/R$ *is* τ*-exponentially light at edge support point zero. The random variable* $\rho(\psi_{\boldsymbol{w}}(X))/\sqrt{d(X)} = 1/R$ *is close to zero when switching from a normal cone to another. Roughly speaking, Condition* (Orl.) *means that the distribution of* $\psi_{\boldsymbol{w}}(X)$ *does not put too much mass around decision boundaries (the border of the normal cones), the learnt direction* $\psi_{\boldsymbol{w}}(X)$ *points towards the interior of the normal cone far from its boundaries.*

Condition (Mom.) *is a moment type assumption on* R. *It is weaker than Condition* (Orl.). *Indeed, by [Chafaï et al., 2012, Theoreme 1.1.5], note that* (Orl.) *is equivalent to*

$$\forall p \geq \tau\,, \quad \left[\mathbb{E}_X\Big(\frac{\rho(\psi_{\boldsymbol{w}}(X))}{\sqrt{d(X)}}\Big)^{-p}\right]^{\frac{1}{p}} \leq C p^{\tau}\,,$$

for some positive constant $C > 0$. *Hence, Condition* (Orl.) *implies* (Mom.). *Condition* (Mom.) *gives a slower rate of convergence of* $V(\lambda)$ *than Condition* (Orl.).

Note that both convergence rates are decreasing in τ *and one cannot have better rate than the sub-Gaussian type bound* $\mathcal{O}_{\lambda\to 0}\big(\exp(-\lambda^{-2})\big)$.