Distributionally Robust Optimization via Regularized Robust Optimization

William Piat* Jalal Fadili† Frédéric Jurie‡ Sébastien Da Veiga§

Abstract

In this paper, we aim at solving distributionally robust optimization problems motivated by application in robust machine learning. For this, we propose a novel SGD-type computationally tractable and provably convergent algorithm without any need of convexity/concavity assumptions unlike most works in the literature. To achieve this, the distributionally robust optimization is first approached with a point-wise counterpart at controlled accuracy. Second, to avoid solving the generally intractable inner maximization problem, we use entropic regularization and Monte Carlo integration. The approximation errors induced by these steps are quantified and thus can be controlled by making the regularization parameter decay and the number of integration samples increase at an appropriate rate. This paves the way to minimizing our objective with stochastic (sub)gradient descent whose convergence guarantees to critical points are established. To support these theoretical findings, compelling numerical experiments on simulated and benchmark datasets are carried out and confirm the practical benefits of our approach.

Keywords Robust optimization, DRO, Smoothing, Regularization, Robust learning, Neural networks, SGD.

1 Introduction

The need of robust models arises when we are considering modeling in the face of uncertainties. Building a reliable decision-making system in the face of uncertain inputs is central to many critical applications: not only the system has to prove it operates correctly on its operational design setting, but it also has to remain stable under some perturbation. In the literature, stability over some kind of perturbation is referred to as robustness and is one of the main challenges in many areas of science and engineering, for instance in optimization, operations research, management sciences, finance, logistics and supply chain or machine learning, to name a few (see [1, 2] for a comprehensive review). Since there are growing applications to critical systems, it is crucial to prove that a statistical model can operate under a given level of uncertainty. On some critical cases the model has to remain stable given any possible point in the uncertain set, as if the perturbation was tailored by an adversary to perturb the decision-making. In this context, Robust Optimization allows to optimize under uncertainty without an explicit model of the uncertainties and thus aims at limiting the scope of actions of any adversary perturbing the model.

*Safran Tech, Digital Sciences and Technologies Department, Rue des Jeunes Bois, Châteaufort, 78114 Magny-Les-Hameaux, France, william.piat@safrangroup.com.

†Normandie Univ, ENSICAEN, CNRS, GREYC, Caen, France, Jalal.Fadili@greyc.ensicaen.fr.

‡Normandie Univ, ENSICAEN, CNRS, GREYC, Caen, France, frederic.jurie@unicaen.fr.

§ENSAI, CREST, CNRS, Rennes, France, sebastien.da-veiga@ensai.fr.

1.1 Problem statement

Let $(\mathcal{Z}, \mathsf{d})$ be a (data) metric space with $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^m$, where $\mathcal{X}$ (resp. $\mathcal{Y}$) is the input (resp. output) space. Let ρ_0 be a probability measure on $\mathcal{Z}$, $\Theta \subset \mathbb{R}^p$ the parameter/action space, $\mathcal{L} : \Theta \times \mathcal{Z} \to \mathbb{R}_+$ a loss function such that $\mathcal{L}(\theta, \cdot)$ is ρ_0-measurable and integrable for all $\theta \in \Theta$. Throughout, we assume that Θ is closed. Consider the optimization problem:

$$\min_{\theta \in \Theta} \mathbb{E}_{z \sim \rho_0} \left[\mathcal{L}(\theta, z)\right]. \tag{1}$$

Robust Optimization (RO) is one contemporary robustification approach to deal with the presence of data perturbations, adversarial attacks, or uncertainties in (1) [1, 2, 3]. The origin of the RO approach can be traced back to the classical economic paradigm of a two-person zero-sum game formulated as a min-max problem (see e.g., [4] the recent review paper on min-max problems and their applications from a signal processing and machine learning perspective, and [5] for a mathematical finance perspective). In this framework, an agent, considered as a defender, is subject to degradation of its performance by a secondary player, the attacker. The defending agent (here θ), whose goal is to minimize an objective function under action constraints Θ, aims at guarding against the degradation of the objective by optimizing its worst value under perturbation without changing the feasibility set of the actions. Put formally, the robust counterpart of (1) reads

$$\min_{\theta \in \Theta} \mathbb{E}_{z \sim \rho_0} \left[\max_{\mathsf{d}(z,z') \leq \varepsilon} \mathcal{L}(\theta, z')\right], \tag{2}$$

where $\varepsilon > 0$ is the size of uncertainty/perturbation/attack, and the perturbation acts pointwise on z in the adversarial risk, which justifies the terminology Pointwise Robust Optimization (PRO) for problem (2). A common choice is $\mathsf{d}(z, z') = \left\|z - z'\right\|_q$ where $\left\|\cdot\right\|_q$ is the ℓ_q norm on $\mathbb{R}^m$ with $q \geq 1$ with the usual adaptation for $q = \infty$.

PRO is one way of quantifying the impact of an adversary or perturbation, and other notions of adversarial risk have been proposed in the literature. In particular, in many areas, such as machine learning, the existence and pervasiveness of adversarial examples point to the limitations of the usual independent and identically distributed (i.i.d.) model of perturbations. This suggests another approach, known as Distributionally Robust Optimization (DRO), which asks to not only perform well on a fixed problem instance (parameterized by a fixed distribution ρ_0), but simultaneously for a range of problems, each determined by a distribution around ρ_0. DRO takes the form

$$\min_{\theta \in \Theta} \max_{\mathsf{D}(\rho, \rho_0) \leq \varepsilon} \mathbb{E}_{z \sim \rho} \left[\mathcal{L}(\theta, z)\right], \tag{3}$$

where D is a discrepancy on the space of probability measures supported on $\mathcal{Z}$. The choice of D affects the richness of the uncertainty set and the tractability of the resulting optimization problem. Typical choices are the Wasserstein distance [6, 7, 8, 9, 10], the Maximum Mean Discrepancy (MMD) [11] or ϕ-divergences including the Kullback-Leibler divergence [12, 13, 14]. There are also other ways to parametrize the perturbation set in terms of the distribution moments, support, etc., [15, 16]. The Wasserstein distance has become very successful in this context, and unlike other distances/divergences, it enjoys the remarkable property that its ball around ρ_0 includes measures having a different support, which allows robustness to unseen data. This is typically the case when ρ_0 is the empirical measure on an observed subset of the data. In the rest of this paper, we focus on Wasserstein balls.

Compared to PRO, DRO has the chief advantage of allowing to consider a larger range of perturbations, and in areas such as machine learning, DRO has been widely studied. On the other hand, one observes that the objective in the inner problem of the saddle point problem (3) is concave (actually linear) in ρ. Though this

property is apparently appealing, this problem operates in infinite dimension (space of probability measures on $\mathcal{Z}$), and thus DRO faces a first important challenge to solve compared to the finite-dimensional PRO problem.
The natural question that arises is whether one can have a surrogate of (3) which operates in finite-dimension under minimal assumptions on the problem data (for instance $\mathcal{L}$), and if this can come up with provable guarantees. For instance, is there a relationship between DRO (3) and PRO (2) (or alike) and hence can we solve the latter as a surrogate for the former ? We will show later that this is indeed the case.
On the other hand, the rigorous treatment of (2), though in finite dimension, remains very challenging for general losses $\mathcal{L}$, especially in absence of the important properties of joint convexity-concavity in (θ, z) and smoothness which are key to design efficient and provably convergent algorithms [1]. These assumptions are however stringent and unrealistic in applications we have in mind, for instance in the so-called adversarial training with neural networks [17, 18, 19, 20]. Iterative solvers used by many authors do not come up with any guarantee. In fact, in such applications, the inner maximization problem is generally non-concave in z and is even provably NP-hard with certain activations such as ReLU [9]. The goal pursued in this paper is thus to design algorithms to solve (2), as a surrogate for (3) under minimal assumptions on the problem data (for instance, without need of joint convexity-concavity) while enjoying convergence guarantees.

1.2 Contributions

Our main contributions in this work are:

1. The DRO problem (3), when D is the Wasserstein distance with Lipschitz continuous ground cost, is approached with a PRO counterpart of the constrained form (2) with a controlled accuracy that depends on the perturbation radius.

2. To avoid solving the generally intractable inner maximization problem in (2), we first smooth the latter using entropic regularization and then use Monte Carlo integration to approximate integrals. We conduct an error analysis to precisely quantify these approximation errors and provide error bounds both on the objective values and its subgradients. Relying on the theory of Γ-convergence we show in particular that the minimizers of the approximate problems converge to those of PRO.

3. Capitalizing on the above results, we propose provably convergent stochastic (sub)gradient descent (SGD) algorithms to solve the PRO problem. The first algorithm supposes access to an oracle of the inner maximization problem. To avoid the latter, which can be challenging, we also provide an inexact SGD algorithm with asymptotically vanishing error term. This error originates from the regularization and integration sampling parameters, and making them decay at an appropriate rate, convergence guarantees to critical points are established without any need of convexity-concavity assumptions on the loss $\mathcal{L}$.

1.3 Relation to prior work

There is a substantial body of work on robust optimization dedicated to robust learning. Here we only review those closely related to ours. Many works have studied instances of DRO for which tractable algorithms can be designed. For D chosen as a ϕ-divergence, and under some assumptions on $\mathcal{L}$, [1, 13, 21] propose convex optimization approaches.

For the Wasserstein distance, several authors convert the DRO problem through duality into a regularized empirical risk minimization problem closely related to PRO [22, 6, 7, 8, 9, 23, 11] (see detailed discussion in

Section 3). Stochastic gradient descent was applied in [9] to this penalized form and convergence guarantees are established under the assumptions that the gradient of the loss $\mathcal{L}$ is bi-Lipschitz and the ground cost c is strongly convex. However, this approach resorts to an oracle corresponding to solving an inner supremum problem. This is again a challenging problem and even NP-hard. Stochastic coordinate descent was also advocated in [11] but without any guarantee. In this work, we treat a much larger class of losses and costs and use smoothing to translate the inner maximization problem into an integration problem that we approximate with Monte Carlo integration. Overall, this allows us to apply stochastic gradient descent while being able to prove convergence to critical points of (2). While we were finalizing a first version of the paper, we became aware of the work of [24] who also used entropic smoothing to learn an optimally robust randomized mixture of classifiers. Their setting and motivation is however different and their algorithm does not enjoy convergence guarantees.

There is another line of research about DRO where D is the entropically regularized Wasserstein distance [25, 26, 27]. In all these papers, the focus is on establishing strong duality and deriving the dual problem of DRO in this case under various more or less stringent assumptions. The key property of the dual problem with the entropic regularization of the Wasserstein distance is that there is no inner maximization problem anymore, which is is replaced by a smooth approximation of the LIE (Log-Integration-Exp) type. This was used in [25, 26] for computational purposes but with little or even no guarantees. This LIE smoothing appears reminiscent of the one we propose for the PRO problem but the two approaches have marked differences. First, our base measure in the integration (within the $\log$) is the uniform one. It is the same as in [25], but different from the one in [26] where it is a Gibbs measure associated to the Wasserstein ground cost and where the entropic regularization parameter plays the role of temperature. Second, while we address directly the PRO problem with the uncertainty set in a constrained form, the latter is in a penalized form in the above papers. We thus do not need to optimize or choose any penalization multiplier. Third, and most importantly, we proposed an SGD-type algorithm and provide its convergence guarantees which are lacking in those papers.

1.4 Organization of the paper

Section 2 summarizes the key prerequisites and notations that are necessary to our exposition. Section 3 shows mild conditions under which DRO can be reasonably approximated using PRO. Section 4 is devoted to our smoothing approach and its key theoretical properties. In Section 5, we turn to studying provably convergent algorithms to solve the PRO problem. Finally, we illustrate these results on some use cases (Section 6) that show the advantages of using smoothing over other heuristics.

2 Preliminaries and assumptions

2.1 Preliminaries

Throughout, $\|\cdot\|_q$, $q \in [1, +\infty]$ is the ℓ_q norm, $\mathbb{B}_r^q(x)$ is the ℓ_q ball of radius $r \geq 0$ centered at x. The subscript q will be omitted when $q = 2$. For $N \in \mathbb{N}$, $[N]$ is the set of integers $\{1, \ldots, N\}$. $\mu_{\mathcal{L}}()$ is the Lebesgue measure/volume of a set. For the nonempty set $\mathcal{C}$, $\iota_{\mathcal{C}}$, is its indicator function (taking 0 on $\mathcal{C}$ and $+\infty$ otherwise), and $\mathrm{dist}(x, \mathcal{C}) = \inf_{z \in \mathcal{C}} \|x - z\|$ is the distance function to $\mathcal{C}$. The set of nearest points of x in $\mathcal{C}$ are denoted by $\mathrm{P}_{\mathcal{C}}(x)$. $\mathscr{C}^s$ is the class of s-continuously differentiable functions and $\mathscr{C}$ is the space of continuous functions. For any sequence $(\theta_k)_{k \in \mathbb{N}}$, we denote $\mathfrak{C}((\theta_k)_{k \in \mathbb{N}})$ the set of its cluster points.

Probability measures For a subset $\mathcal{C} \subset \mathbb{R}^m$, let $\mathcal{B}$ be the Borel σ-algebra on $\mathcal{C}$. $\mathcal{M}_+(\mathcal{C})$ denotes the cone of non-negative measures on $(\mathcal{C}, \mathcal{B})$ equipped with the finite total variation norm. We also define $\mathcal{P}(\mathcal{C})$ the space of Borel probability measures on $\mathcal{C}$

$$\mathcal{P}(\mathcal{C}) \stackrel{\text{def}}{=} \left\{\varphi \in \mathcal{M}_+(\mathcal{C}) : \int_{\mathcal{C}} \mathrm{d}\varphi(x) = 1\right\}.$$

δ_x is the Dirac measure at x.

For any $(\nu, \mu) \in \mathcal{P}(\mathcal{C})$, the Kullback-Leibler divergence between μ and ν is

$$\mathrm{KL}(\mu, \nu) = \begin{cases} \int_{\mathcal{C}} \log\left(\frac{\mu(x)}{\nu(x)}\right) \mathrm{d}\mu(x) & \text{if } \mu \ll \nu \text{ and } \int_{\mathcal{C}} \left|\log\left(\frac{\mu(x)}{\nu(x)}\right)\right| \mathrm{d}\mu(x) < \infty \\ +\infty & \text{otherwise,} \end{cases}$$

where $\ll$ stands for absolute continuity of measures.

For $(\nu, \mu) \in \mathcal{P}(\mathcal{C})$, denote $\Pi(\mu, \nu)$ their couplings, i.e., joint probability measures π on $\mathcal{Z}^2$ whose marginals are μ and ν. The Wasserstein distance between μ and ν with ground/transportation cost c is

$$W_{\mathsf{c}}(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{Z}^2} \mathsf{c}(z, z') \mathrm{d}\pi(z, z').$$

When $\mathsf{c}(z, z') = \|z - z'\|_q^q$, $q \geq 1$, then $W_{\mathsf{c}}^{1/q}$ is indeed a distance, known as the q-Wasserstein distance. We will denote it W_q.

Γ- or epi-convergence We will invoke the notion of Γ-convergence, which plays a fundamental role in convergence of optimization problems (values and extrema points). In finite dimension, Γ-convergence of a sequence of functions corresponds to convergence of their epigraphs. The interested reader may refer to [28] for a comprehensive treatment.

Tameness We will need the notion of tame functions (and sets). A rich family will be provided by semialgebraic functions, i.e., functions whose graph is defined by some Boolean combination of real polynomial equations and inequalities [29]. Definable functions on an o-minimal structure over $\mathbb{R}$ correspond in some sense to an axiomatization of some of the prominent geometrical properties of semialgebraic geometry [30, 31]. O-minimality includes many important structures such as globally subanalytic sets or sets belonging to the log-exp structure hence covering the vast majority of applications in learning, including neural network learning with various activations and loss functions. A slightly more general notion is that of a tame function, which is a function whose graph has a definable intersection with every bounded box. We then use the terminology definable for both. Given the variety of optimization problems that can be formulated within the framework of definable functions and sets, our convergence results will be stated for this class. The reader unfamiliar with these notions can just replace definability by semialgebraicity.

We now summarize a few properties of the Clarke subdifferential that will be useful to us in this paper; see [32].

Proposition 2.1. *Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz continuous functions. Then*

(i) $\partial^C(\lambda f)(x) = \lambda \partial^C f(x)$, $\lambda \in \mathbb{R}$.

(ii) $\partial^C(f + g)(x) \subset \partial^C f(x) + \partial^C g(x)$.

(iii) *Consider the family of functions $(f_t)_{t\in T}$, where T is a compact space and $t \mapsto f_t(x)$ is upper semi-continuous. Suppose that for each t, $f_t : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz continuous. Let $f(x) = \max_{t\in T} f_t(x)$. Let S be a subset of full Lebesgue measure. Then*

$$\partial^C f(x) \subset \operatorname{conv}\left\{\lim_{k\to\infty} \nabla f_{t_k}(x_k) : \ x_k \to x, x_k \in S, t_k \in T, f_{t_k}(x) \to f(x)\right\}. \tag{4}$$

If moreover, the functions f_t are of class $\mathscr{C}^1$ such that $f_t(x)$ and $\nabla f_t(x)$ depend continuously on (t,x)[1]*, then*

$$\partial^C f(x) = \left\{\int_T \nabla f_t(x) \mathrm{d}\mu(t) : \ \mu \in \mathcal{P}\Big(\operatorname*{Argmax}_{t\in T} f_t(x)\Big)\right\}. \tag{5}$$

Remark 2.2. *We have made no effort to further weaken the assumptions in the calculus rules of Proposition 2.1 since they are sufficient for our purpose.*

2.2 Main assumptions

In order to help the reading, we summarize the working assumptions on c and $\mathcal{L}$ that will be used throughout the paper (for some results only a subset of these assumptions will be needed). In the sequel, the perturbation radius ε is fixed and given.

(H.1) $\mathcal{Z}$ is a compact set and the ground cost $\mathsf{c}(z,z') = \|z - z'\|^q$, $q \geq 1$, where $\|\cdot\|$ is a norm on $\mathbb{R}^m$. In the sequel, we denote the set $\mathcal{C}_\varepsilon$ as a ball of $\mathcal{Z}$ in the norm $\|\cdot\|$ of radius ε, i.e.

$$\mathcal{C}_\varepsilon \stackrel{\text{def}}{=} \{z \in \mathcal{Z} : \ \|z\| \leq \varepsilon\}.$$

(H.2) $\mathcal{L}$ is continuous on $\Theta \times \mathcal{Z}$ and bounded from below.

(H.3) For each bounded subset $\Xi \subset \Theta$, there exists a strictly increasing function $\phi_{\mathcal{Z}} : \mathbb{R}_+ \to \mathbb{R}_+$ with $\phi_{\mathcal{Z}}(0) = 0$ such that $|\mathcal{L}(\theta,z) - \mathcal{L}(\theta,z')| \leq \phi_{\mathcal{Z}}(\|z - z'\|)$ for all $\theta \in \Xi$ and all $(z,z') \in \mathcal{Z}^2$.

(H.4) For each bounded subset $\Xi \subset \Theta$, $\exists l_\Xi > 0$ such that $|\mathcal{L}(\theta,z) - \mathcal{L}(\theta',z)| \leq l_\Xi \|\theta - \theta'\|$, $\forall(\theta,\theta') \in \Xi^2$ and $\forall z \in \mathcal{C}_\varepsilon + \mathcal{Z}$.

(H.5) For $z \in \mathcal{Z}$, $\mathcal{L}(\cdot,z)$ is differentiable on an open set containing Θ with $\nabla_\theta\mathcal{L}(\cdot,\cdot)$ continuous in both of its arguments, and $\nabla_\theta\mathcal{L}(\cdot,z)$ is locally Lipschitz continuous uniformly in z, i.e., for any bounded set Ξ, $\exists L_\Xi > 0$ such that $\|\nabla_\theta\mathcal{L}(\theta,z) - \nabla_\theta\mathcal{L}(\theta',z)\| \leq L_\Xi \|\theta - \theta'\|$, $\forall(\theta,\theta') \in \Xi^2$ and $\forall z \in \mathcal{C}_\varepsilon + \mathcal{Z}$. Let $L_{\Xi,\mathcal{Z}} \stackrel{\text{def}}{=} \max_{\theta\in\Xi, z\in\mathcal{Z}+\mathcal{C}_\varepsilon} \|\nabla_\theta\mathcal{L}(\theta,z)\|$.

Remark 2.3 (Discussion of the assumptions)**.**

- *One can see that $\mathcal{C}_\varepsilon$ is convex, compact and full dimensional. (H.1) corresponds to the case where $W_{\mathsf{c}}^{1/q}$ is the q-Wasserstein distance W_q. This assumption is not restrictive, and was made to make the presentation lighter. It is also sufficient for our purposes and it covers most applications we have in mind, e.g. robust training in machine learning as reviewed in the introduction. Many of our results can however be extended rather easily to more general costs c as soon as the corresponding set $\mathcal{C}_\varepsilon$ is compact, convex and full dimensional.*

[1] Functions f such that these assumptions are verified are known as lower-$\mathscr{C}^1$ functions; see [33].

- *The compactness assumption on $\mathcal{Z}$ is crucial. First, and naturally, it ensures, together with (H.2), that the inner maximization problem in (8) is well-defined. This assumption is also needed for the max formula (5) to apply with $T = \mathcal{C}_\varepsilon + z$, for any $z \in \mathcal{Z}$. Second, this assumption entails, with continuity, that the constants l_Ξ and $L_{\Xi,\mathcal{Z}}$ are finite. Again, our setting is very general (non-smooth, non-convex) and we need minimal compactness assumptions to prove guarantees. Note that this compactness assumption holds in many applications including machine learning considered in this paper.*

- *(H.3)-(H.4) are mild. They hold for instance of $\mathcal{L}$ is continuously differentiable in both arguments. This covers many problems of interest including in machine learning that will be the focus of our application. (H.3) also holds if, e.g., $\mathcal{L}(\theta, \cdot)$ is locally Hölder continuous. Note also that (H.3)-(H.4) imply (H.2) if $\Theta = \Xi$.*

- *(H.4) says that $\mathcal{L}(\theta, z)$ is locally Lipschitz continuous in θ uniformly in z. This is a minimal requirement for the max formula (4) to apply.*

- *(H.5) is a strengthened version of (H.4), and will allow, among others, to apply the max formula in (5) of Proposition 2.1(iii).*

- *Our assumptions are much weaker than the strong concavity assumption of $\mathcal{L}(\theta, \cdot) - \gamma\|z - \cdot\|^q$, on $\mathcal{Z}$, for any $\gamma > 0$, that most existing works in the literature generally assume.*

3 From DRO to PRO

The goal here is to show how to go from the DRO problem (3) to the PRO one (2), provably, by bounding the corresponding objectives. This paves the way to using PRO as a surrogate for DRO provided that ε is not too large.

Proposition 3.1. *Suppose that (H.1), (H.2) and (H.3) hold. Then, for any $\theta \in \Xi$, where Ξ is a bounded subset of Θ,*

$$0 \leq \sup_{W_q(\rho,\rho_0)\leq\varepsilon} \mathbb{E}_{z\sim\rho}[(\mathcal{L}(\theta, z)] - \mathbb{E}_{z\sim\rho_0}\left[\max_{z'\in\mathcal{C}_\varepsilon} \mathcal{L}(\theta, z + z')\right] \leq C_{q,l_\mathcal{Z}}\varepsilon,$$

where $C_{q,l_\mathcal{Z}}$ is a non-negative constant that depends only on q and $l_\mathcal{Z}$.

See Appendix A.1 for the proof.

In [9] (see also [7]), using Lagrangian duality arguments, it was shown that

$$\sup_{W_\mathsf{c}(\rho,\rho_0)\leq\varepsilon} \mathbb{E}_{z\sim\rho}[(\mathcal{L}(\theta, z)] = \inf_{\gamma\geq 0}\left(\gamma\varepsilon + \mathbb{E}_{z\sim\rho_0}\left[\sup_{z'\in\mathcal{Z}}\left(\mathcal{L}(\theta, z') - \gamma\mathsf{c}(z, z')\right)\right]\right), \tag{6}$$

For a fixed parameter $\gamma > 0$, this can be seen as a penalized form of the constrained form (2). Though (6) is an identity rather than a bound, it faces a few algorithmic challenges to solve. For instance, the joint presence of the expectation and the inner maximization problems makes minimization of the multiplier γ a difficult task. One can of course think of a simple procedure such as bisection but this will necessitate extra-smoothness assumptions and to solve for θ for each value of γ on the bisection. If $\mathcal{L}(\theta, \cdot)$ has a Lipschitz continuous gradient, and c is strongly convex in its second argument, then it can be easily shown that for γ large enough, $\mathcal{L}(\theta, \cdot) - \mathsf{c}(z, \cdot)$ is strongly concave. This has been leveraged by [9] to use gradient descent to minimize over θ, but only for a fixed (large enough) parameter γ. But still, choosing γ is not easy.

Taking c as in (H.1), and ρ_0 the empirical measure on n points, the following bound was established in [23, 11],

$$0 \leq \sup_{W_q(\rho,\rho_0)\leq\varepsilon} \mathbb{E}_{z\sim\rho}[(\mathcal{L}(\theta, z)] - \sup_{(z'_i)_i : \frac{1}{n}\sum_{i=1}^{n}\|z'_i - z_i\|^q \leq \varepsilon^q} \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\theta, z'_i) \leq l_{\mathcal{Z}}/n. \tag{7}$$

As in our case, this gives access to a uniform bound, but which depends now on n rather than ε (and thus gets tighter as n increases). However, the price to pay is that, unlike problem (6), the inner maximization in the surrogate problem (7) is coupled between all variables in the objective and constraints, necessitating to optimize on a variable in $\mathbb{R}^{mn}$ rather than $\mathbb{R}^m$.

4 Entropic regularization

Despite formulating surrogates as devised in (6) and (7) and discussed above, solving the resulting min-max problems remains a very challenging task unless stringent joint convexity/concavity assumptions are made. This is the motivation behind our smoothing hereafter.

In view of Proposition 3.1, the PRO problem (2) now reads

$$\min_{\theta\in\Theta} \mathbb{E}_{z\sim\rho_0}\left[\max_{z'\in\mathcal{C}_\varepsilon} \mathcal{L}(\theta, z + z')\right]. \tag{8}$$

We will use the shorthand notation[2]

$$g(\theta) \overset{\text{def}}{=} \max_{z'\in\mathcal{C}_\varepsilon} \mathcal{L}(\theta, z + z'). \tag{9}$$

As $\mathcal{C}_\varepsilon$ is compact by definition, and recalling the continuity assumption (H.2), the set of maximizers in g is a nonempty compact set. We will then rewrite function g as

$$g(\theta) = \max_{\mu\in\mathcal{P}(\mathcal{C}_\varepsilon)} \int_{\mathcal{C}_\varepsilon} \mathcal{L}(\theta, z + z')\mathrm{d}\mu(z'), \tag{10}$$

and the integral is a duality pairing between $\mathcal{P}(\mathcal{C}_\varepsilon)$ and $\mathscr{C}(\mathcal{C}_\varepsilon)$. We will show that (10) and its subgradients can be provably approximated following a two-step strategy: first, an *entropic regularization* followed by *Monte-Carlo sampling* to approximate the integrals.

4.1 Regularized objective

Problem (10) is concave in μ, but operates in infinite-dimension and is thus hard to solve. Approximating (10) by an atomic measure supported on a finite set (i.e., replace $\mathcal{P}(\mathcal{C}_\varepsilon)$ by a finite-dimensional simplex by sampling N points at random in $\mathcal{C}_\varepsilon$), as done by some authors (see e.g., [34]), suffers an exponential dependence in $1/m$. Indeed, an analysis using Lipschitzianity of $\mathcal{L}$ shows that this method achieves an approximation rate of $O(N^{-1/m})$, and essentially, this cannot be improved. Rather, we will consider the following regularized version of (10), namelyAs for g, g_τ also depends on z but we drop this in the notation.

$$g_\tau(\theta) \overset{\text{def}}{=} \max_{\mu\in\mathcal{P}(\mathcal{C}_\varepsilon)} \int_{\mathcal{C}_\varepsilon} \mathcal{L}(\theta, z + z')\mathrm{d}\mu(z') - \tau\mathrm{KL}(\mu, \nu), \tag{11}$$

[2] Keep in mind that g depends on z but we drop this to lighten the notation. This will have no incidence on our claims and analysis.

where $\tau > 0$ is the regularization parameter, and ν is a reference measure on $\mathcal{C}_\varepsilon$. Entropic regularization has been used in several fields including optimal transport [35] and semi-infinite programming [36].

In the sequel we will set ν as the uniform measure $\mu_{\mathcal{U}}$ on $\mathcal{C}_\varepsilon$, that is $\mathrm{d}\nu(z') = \mu_{\mathcal{L}}(\mathcal{C}_\varepsilon)^{-1}\mathrm{d}z'$ for all $z' \in \mathcal{C}_\varepsilon$. The KL regularization term then prevents solutions to be atomic measures as such measures are not absolutely continuous with respect to the uniform one.

Remarkably, (11) is well-posed under mild conditions and has a unique solution taking an explicit form.

Proposition 4.1. *Assume that (H.1)-(H.2) hold. Then (11) has a unique solution and*

$$g_\tau(\theta) = \tau \log\left(\mathbb{E}_{z'\sim\mu_{\mathcal{U}}}\left[\exp\left(\frac{\mathcal{L}(\theta, z+z')}{\tau}\right)\right]\right) = \tau \log\left(\frac{\int_{\mathcal{C}_\varepsilon} \exp\left(\frac{\mathcal{L}(\theta,z+z')}{\tau}\right)\mathrm{d}z'}{\mu_{\mathcal{L}}(\mathcal{C}_\varepsilon)}\right). \tag{12}$$

The proof can be found in Appendix A.2. It turns out that convexity of $\mathcal{C}_\varepsilon$ is not needed. Note also that this is a generalization of the standard log-sum-exp formula for softmax smoothing of the maximum of a finite number of functions, where now the sum is replaced by an expectation wrt to the base measure $\mu_{\mathcal{U}}$.

4.2 Consistency of the regularization

We now turn to describing how g_τ in (11) is a good surrogate for g in (8). We provide both a qualitative result as the regularization parameter τ vanishes, and a quantitative convergence rate.

Theorem 4.2. *Assume that (H.1)-(H.2) hold. Then, the following statements hold:*

(i) $(\mathbb{E}_{z\sim\rho_0}[g_\tau])_{\tau\geq 0}$ *is increasing as* $\tau \searrow 0$ *with* $\mathbb{E}_{z\sim\rho_0}[g_\tau(\theta)] \leq \mathbb{E}_{z\sim\rho_0}[g(\theta)]$ *for any* $\theta \in \Theta$*. If* Θ *is a compact set, then* $\mathbb{E}_{z\sim\rho_0}[g_\tau] + \iota_\Theta$ Γ*-converges to* $\mathbb{E}_{z\sim\rho_0}[\sup_\tau g_\tau] + \iota_\Theta$ *as* $\tau \to 0^+$.

(ii) If, moreover, (H.3) holds, then for any $\tau \in]0,1]$ *and* $\theta \in \Xi$*, where* Ξ *is a compact subset of* Θ*,*

$$g(\theta) - h(\tau) \leq g_\tau(\theta) \leq g(\theta), \tag{13}$$

where

$$h(\tau) = m\tau\log(\tau^{-1}) + \phi_{\mathcal{Z}}\left(\tau D_{\mathcal{C}_\varepsilon}\right), \tag{14}$$

and $D_{\mathcal{C}_\varepsilon}$ *is the diameter of* $\mathcal{C}_\varepsilon$*. Thus* $\mathbb{E}_{z\sim\rho_0}[g_\tau] + \iota_\Xi$ Γ*-converges to* $\mathbb{E}_{z\sim\rho_0}[g] + \iota_\Xi$ *as* $\tau \to 0^+$.

See Appendix A.3 for the proof.

Remark 4.3.

1. *The second term in* $h(\tau)$ *depends on* $\phi_{\mathcal{Z}}$ *and reflects the impact of the smoothness of the loss* $\mathcal{L}(\theta,\cdot)$ *on how well* g_τ *approximates* g*. When* $\mathcal{L}(\theta,\cdot)$ *is* κ*-Hölderian with* $\kappa \in]0,1]$*, this scales as* $O(\tau^\kappa)$.

2. *For the dependence on the dimension of the convergence rate, the first term grows linearly. For the second term, the function* $\phi_{\mathcal{Z}}$ *in (H.3) may also hide dependence on the dimension. This again emphasizes the role of the smoothness modulus of the loss in controlling the robustness of the model.*

In view of Theorem 4.2, we obtain the following key result, which relates the minimizers of the smoothed problem to those of the original PRO problem.

Theorem 4.4. *Suppose that* (H.1), (H.2) *and* (H.3) *hold, where* Θ *is a compact set. Let* $\theta_\tau^\star \in \operatorname{Argmin}_{\theta\in\Theta} g_\tau(\theta)$. *Then the following holds:*

(i) $\lim_{\tau\to 0^+} \min_{\theta\in\Theta} \mathbb{E}_{z\sim\rho_0}\left[g_\tau(\theta)\right] = \min_{\theta\in\Theta} \mathbb{E}_{z\sim\rho_0}\left[g(\theta)\right]$.

(ii) Each cluster point of $\theta_\tau^\star$, *as* $\tau \to 0^+$, *lies in* $\operatorname{Argmin}_{\theta\in\Theta} \mathbb{E}_{z\sim\rho_0}\left[g(\theta)\right]$.

(iii) In particular, if $\operatorname{Argmin}_{\theta\in\Theta} \mathbb{E}_{z\sim\rho_0}\left[g(\theta)\right] = \{\theta^\star\}$, *then* $\lim_{\tau\to 0^+} \theta_\tau^\star = \theta^\star$.

See Appendix A.4 for the proof.

Remark 4.5. *It is well-known that convergence of minimal values and minimizers from* Γ*-convergence needs uniform compactness materialized by equi-coercivity. Compactness of* Θ *ensures this and that the claims of Theorem* 4.2 *hold true.*

4.3 Consistency of the Monte Carlo integration

Computing the values $g_\tau(\theta)$ in (12) necessitates to compute a possibly high dimensional integral. Our goal is to approximate the latter with Monte Carlo integration by uniformly drawing independent samples $(z_k')_{k=1}^N$ in the set $\mathcal{C}_\varepsilon$. This gives the approximation

$$g_{\tau,N}(\theta) \overset{\text{def}}{=} \tau \log\left(\sum_{k=1}^{N} \frac{\exp\left(\frac{\mathcal{L}(\theta,z+z_k')}{\tau}\right)}{N}\right). \tag{15}$$

We now provide an error bound for such an approximation. We denote $\Delta\mathcal{L}(\theta) \overset{\text{def}}{=} \overline{\mathcal{L}}(\theta) - \underline{\mathcal{L}}(\theta)$, where $\underline{\mathcal{L}}(\theta) \overset{\text{def}}{=} \inf \mathcal{L}(\theta, \mathcal{Z} + \mathcal{C}_\varepsilon)$ and $\overline{\mathcal{L}}(\theta) \overset{\text{def}}{=} \sup \mathcal{L}(\theta, \mathcal{Z} + \mathcal{C}_\varepsilon)$. For Ξ a compact subset of Θ, will also define $\underline{\mathcal{L}}^\Xi \overset{\text{def}}{=} \min \underline{\mathcal{L}}(\Xi)$, $\overline{\mathcal{L}}^\Xi \overset{\text{def}}{=} \max \overline{\mathcal{L}}(\Xi)$, and $\Delta\mathcal{L}^\Xi \overset{\text{def}}{=} \overline{\mathcal{L}}^\Xi - \underline{\mathcal{L}}^\Xi$ is the oscillation of $\mathcal{L}$ on Ξ. The lower bound $\underline{\mathcal{L}}^\Xi$ is well-defined thanks to assumption (H.2). We also have $\overline{\mathcal{L}}^\Xi < +\infty$ thanks to continuity of $\mathcal{L}$ on $\Theta \times \mathcal{Z}$ by (H.2) and compactness of $\mathcal{Z}$ by (H.1). In turn, $\Delta\mathcal{L}^\Xi < +\infty$, i.e., $\mathcal{L}$ has bounded oscillation on Ξ.

Theorem 4.6. *Suppose that* (H.1), (H.2) *and* (H.3) *are verified. Then, the following holds for any compact set* $\Xi \subset \Theta$.

(i) We have

$$\sup_{\theta\in\Xi} \mathbb{E}\left[\left|g_{\tau,N}(\theta) - g(\theta)\right|\right] \leq h(\tau) + \frac{\tau e^{\frac{\Delta\mathcal{L}^\Xi}{\tau}}}{\sqrt{N}}, \tag{16}$$

where $h(\tau)$ *is given in* (14).

(ii) If, moreover, (H.4) *holds, then for any* $t > 0$

$$\sup_{\theta\in\Xi} \left|g_{\tau,N}(\theta) - g(\theta)\right| \leq h(\tau) + \tau e^{\frac{\Delta\mathcal{L}^\Xi}{\tau}} \sqrt{\frac{(p+t)\log N}{2N}} + e^{\frac{\Delta\mathcal{L}^\Xi}{\tau}} \frac{4 l_\Xi D_\Xi}{N}, \tag{17}$$

with probability at least $1 - 2N^{-t}$, *where* D_Ξ *is the diameter of* Ξ.

(iii) Assume, in addition, that

(H.6) *τ is a function of N, say τ_N, with $\tau_N \to 0$ and $\tau_N e^{\frac{\Delta\mathcal{L}^\Xi}{\tau_N}}\sqrt{\frac{\log N}{N}} + \frac{e^{\frac{\Delta\mathcal{L}^\Xi}{\tau_N}}}{N} \to 0$ as $N \to +\infty$.*

Then,

(a)

$$\sup_{\theta\in\Xi} \mathbb{E}\left[|g_{\tau,N}(\theta) - g(\theta)|\right] \underset{N\to+\infty}{\longrightarrow} 0.$$

(b) If (H.4) also holds then the event

$$g_{\tau_N,N}(\theta) \underset{N\to+\infty}{\longrightarrow} g(\theta) \quad \textit{for all} \quad \theta \in \Xi$$

holds almost surely.

The proof can be found in Appendix A.5.

Remark 4.7. *It is important to realize that the control in Theorem 4.6(ii) and (iii)(b) is uniform in all $\theta \in \Xi$. This is the reason we need the additional regularity assumption (H.4). Observe also that the convergence in expectation in (iii)(a) and that in almost sure sense in (iii)(b) are not equivalent. In fact, (iii)(b) implies (iii)(a) by boundedness and the dominated convergence theorem. The converse is not true. In our statements above,* sup *is actually a* max *since the inner objective is continuous (see the proof of Theorem 4.2) and Ξ is compact.*

For fixed τ, the convergence rate in (17) is $O(N^{-1/2})$ (up to a logarithmic term). But one has to keep in mind that we have not used any smoothness property of $\mathcal{L}(\theta,\cdot)$ in our simple (uniform) Monte Carlo sampling integration. This could be possibly improved using the rich theory of Monte Carlo integration, see e.g., [37, 38], but probably at the price of a higher computation cost. Such improvements may also potentially necessitate more sophisticated deviation bounds in the proof instead of Hoeffding's inequality that we use here.

Note that the variance of the samples appears implicitly in (17) through the exponential term. One can alternatively use Bernstein's inequality instead of Hoeffding's inequality to show that

$$\sup_{\theta\in\Xi} |g_{\tau,N}(\theta) - g_\tau(\theta)| = O\left(\tau\frac{\sigma(\theta)}{\bar{S}(\theta)}\sqrt{\frac{t\log N}{N}}\right)$$

with high probability, where

$$\bar{S}(\theta) = \frac{1}{\mu_{\mathcal{L}}(\mathcal{C}_\varepsilon)}\int_{\mathcal{C}_\varepsilon} e^{\frac{\mathcal{L}(\theta,z')}{\tau}}\mathrm{d}z' \text{ and } \sigma^2(\theta) = \frac{1}{\mu_{\mathcal{L}}(\mathcal{C}_\varepsilon)}\int_{\mathcal{C}_\varepsilon} e^{2\frac{\mathcal{L}(\theta,z')}{\tau}}\mathrm{d}z' - \bar{S}(\theta)^2.$$

The error bound (17) reveals an exponential dependence in τ, and thus for the right-hand side to vanish as $\tau \to 0^+$ and $N \to +\infty$, there is a trade-off between N and τ to annihilate the exponential term and make the right hand side of (17) converge to 0. Taking $\tau \sim (\kappa\log N)^{-1}$, for any $\kappa > 0$ large enough such that $\kappa\Delta\mathcal{L}^\Xi < 1/2$, the convergence rate in (17) is dominated by the first term $h(\tau)$ which scales as $O\left((\log N)^{-1}\right)$. This choice of τ however depends on Ξ which is not always desirable unless Ξ is known a priori. The alternative choice $\tau = c/\log(\log N)$, $c > 0$, is independent of Ξ but entails a slower convergence rate in (17) (and (16)) of $O\left((\log(\log N))^{-1} + \phi_{\mathcal{Z}}\left((\log(\log N))^{-1}\right)\right)$. These slow rates anyway reflect the difficulty of approximating the function g in (8) without any smoothness prior.

4.4 Consistency of the subgradient estimate

Equipped with the above results, a natural strategy now is to solve the PRO problem (8) by using $g_{\tau,N}$ in (15) as a (provably controlled) approximation of g. Towards this goal, we would like to apply a first-order scheme, typically (sub)gradient descent. Such a scheme will involve a first-order oracle on $g_{\tau,N}$, the gradient

$$\nabla g_{\tau,N}(\theta) = \sum_{k=1}^{N} \nabla_\theta \mathcal{L}(\theta, z + z'_k) \frac{\exp\left(\frac{\mathcal{L}(\theta,z+z'_k)}{\tau}\right)}{\sum_{j=1}^{N} \exp\left(\frac{\mathcal{L}(\theta,z+z'_j)}{\tau}\right)}. \tag{18}$$

The natural question that arises is whether (18) behaves well as τ vanishes and is a consistent approximation of a Clarke subgradient of g (the latter being accessible via the formula (5) in Proposition 2.1 under mild assumptions). The result in Theorem 4.8 shows that this is indeed the case under appropriate assumptions on $\mathcal{L}$ that are slightly stronger than those required for consistency of the (zero-th order oracle) function values established in Theorem 4.6. In turn, Assumption (H.6) will also be slightly strengthened to:

(H$'$.6) τ is a function of N, say τ_N, with $\tau_N \to 0$ and $e^{\frac{\Delta_{\mathcal{L}}^{\Xi}}{\tau_N}} \sqrt{\frac{\log N}{N}} + \frac{\tau_N^{-1} e^{\frac{\Delta_{\mathcal{L}}^{\Xi}}{\tau_N}}}{N} \to 0$ as $N \to +\infty$.

Theorem 4.8. *Suppose that (H.1), (H.2), (H.3), (H.4) and (H.5) hold. Then $\mathcal{L}$ and $g_{\tau,N}$ are continuously differentiable, and the following holds for any compact set $\Xi \subset \Theta$.*

(i) We have

$$\sup_{\theta\in\Xi} \mathbb{E}\left[\operatorname{dist}\left(\nabla g_{\tau,N}(\theta), \partial^C g(\theta)\right)\right] \leq \frac{2L_{\Xi,\mathcal{Z}} e^{\frac{\Delta_{\mathcal{L}}^{\Xi}}{\tau}}}{\sqrt{N}} + \psi(\tau). \tag{19}$$

where

$$\psi(\tau) = o_\tau(1) \ \ \textit{and} \ \ \partial^C g(\theta) = \left\{ \int_{\mathcal{C}_\varepsilon} \nabla_\theta \mathcal{L}(\theta, z + z') \mathrm{d}\mu(z') : \ \mu \in \mathcal{P}\Big(\operatorname{Argmax} \mathcal{L}(\theta, z + \mathcal{C}_\varepsilon)\Big) \right\}. \tag{20}$$

(ii) For any $t > 0$ and N large enough,

$$\sup_{\theta\in\Xi} \operatorname{dist}\left(\nabla g_{\tau,N}(\theta), \partial^C g(\theta)\right) \leq 4L_{\Xi,\mathcal{Z}} e^{\frac{\Delta_{\mathcal{L}}^{\Xi}}{\tau}} \sqrt{\frac{2(p+t)\log N}{N}} + e^{\frac{\Delta_{\mathcal{L}}^{\Xi}}{\tau}} \frac{4\left(\tau^{-1} L_{\Xi,\mathcal{Z}}(l_\Xi + 1) + L_\Xi\right) D_\Xi}{N} + \psi(\tau). \tag{21}$$

with probability at least $1 - 4N^{-t}$.

(iii) Assume, moreover, (H$'$.6) is verified, then

(a) $\sup_{\theta\in\Xi} \mathbb{E}\left[\operatorname{dist}\left(\nabla g_{\tau,N}(\theta), \partial^C g(\theta)\right)\right] \underset{N\to+\infty}{\longrightarrow} 0.$

(b) The event

$$\sup_{\theta\in\Xi} \operatorname{dist}\left(\nabla g_{\tau_N,N}(\theta), \partial^C g(\theta)\right) \underset{N\to+\infty}{\longrightarrow} 0 \quad \textit{for all} \quad \theta \in \Xi.$$

holds almost surely.

The proof is deferred to Appendix A.6.

Clearly, Theorem 4.8(i)-(ii) tell us that simultaneously for all $\theta \in \Xi$, $\nabla g_{\tau,N}(\theta)$ is within a ball around the Clarke subdifferential of g at any $\theta \in \Xi$ both in expectation and high probability. The result also quantifies its radius and how it vanishes with N and τ. Arguing as for Theorem 4.6, this radius vanishes as $N \to +\infty$ by taking $\tau \sim (\kappa \log N)^{-1}$, for any κ such that $\kappa \Delta \mathcal{L}^{\Xi} \in]0, 1/2 - \alpha]$, $\alpha \in]0, 1/2[$. The convergence rate of the first term in (21) is then nearly $O\left(N^{-\alpha}\right)$ (up to logarithmic factors). An alternative choice that does not depend on Ξ is $\tau = c/\log(\log N)$, $c > 0$, entailing a convergence rate of the first term in (21) of $O(N^{-1/2})$ up to a polylogarithmic factor. As far as the $o_\tau(1)$ term in (21) is concerned, we do not have a quantitative estimate for the corresponding rate in general. Under additional assumptions, a close inspection of the proof Lemma A.4 reveals that the convergence rate in τ is at least

$$O\left(\tau^{-\frac{m-m_1}{2}} e^{-\frac{\kappa}{\tau}} + \sum_{i>1} \tau^{\frac{m_1-m_i}{2}} + \tau^{1/2}\right) = O\left(\tau^{1/2}\right).$$

This is again the term that will dominate the convergence behaviour of (21) and would vanish slowly for the choice $\tau = 1/\log(\log N)$.

Remark 4.9. *It is worth noting that the statements of Lemma A.2, hence Theorem 4.8, only ensure subsequential convergence of the gradients whose cluster points are Clarke subgradients of g. This is sufficient for our purposes, and for instance when analyzing convergence properties of our SGD algorithms (see Section 5). Showing global convergence of the gradients, together with a precise form of the limit Clarke subgradient and the corresponding convergence rate in τ is possible, but is a more complicated problem. This in turn necessitates extra regularity assumptions. For the interested reader, this is detailed in Lemma A.4 in the appendix.*

5 Robust optimization algorithm via SGD

We are now ready to describe our algorithmic framework to solve the PRO problem (8) based on the stochastic subgradient (SGD) method. To make the presentation easier, we assume that $\Theta = \mathbb{R}^p$ though our algorithmic framework will be extended later in Section 5.2 to the constrained case over a convex compact set by including a projection step. We now consider the standard setting where ρ_0 in (8) is the empirical measure on $\mathcal{Z}$, hence leading to the finite sum minimization problem

$$\min_{\theta \in \mathbb{R}^p} \left\{ G(\theta) \stackrel{\text{def}}{=} \frac{1}{M} \sum_{i=1}^{M} g_i(\theta) \right\} \quad \text{where} \quad g_i(\theta) \stackrel{\text{def}}{=} \max_{u \in \mathcal{C}_\varepsilon} \mathcal{L}(\theta, z_i + u), \tag{22}$$

5.1 Without smoothing

As revealed by Proposition 2.1, the Clarke subdifferential has only inclusion rules under finite sum and pointwise maximization. Thus, if in addition to (H.2), is $\mathcal{L}(\cdot, z_i + u)$ is assumed locally Lipschitz continuous for each (z_i, u), then the calculus rules in Proposition 2.1 give us the inclusion

$$\partial^C G(\theta) = \partial^C \left(\frac{1}{M} \sum_{i=1}^{M} g_i(\theta) \right) \subset \frac{1}{M} \sum_{i=1}^{M} \partial^C g_i(\theta) \subset \frac{1}{M} \sum_{i=1}^{M} \mathcal{D}_i(\theta), \tag{23}$$

where

$$\mathcal{D}_i(\theta) = \operatorname{conv}\left\{ \lim_{k\to\infty} \nabla_\theta \mathcal{L}(\theta_k, z_i + u_k) : \ \theta_k \to \theta, \theta_k \in S, u_k \in \mathcal{C}_\varepsilon, \mathcal{L}(\theta, z_i + u_k) \to g_i(\theta) \right\}.$$

Remark 5.1. *The inclusion above, for instance the one of the sum rule, is strict in many situations of interest in applications. For the sum rule, one may consider other generalized derivatives or even other (but closely related) fields than the Clarke subdifferential, e.g. the conservative fields proposed in [39, 40]. These fields enjoy nice sum and chain rules and coincide with the Clarke subdifferential almost everywhere. However, calculus rules for conservative fields under pointwise maximization involving an inner* constrained *maximization oracle is still an open problem.*

We are now naturally led to consider the set of critical points:

$$\text{crit–}G \overset{\text{def}}{=} \left(\frac{1}{M} \sum_{i=1}^{M} \mathcal{D}_i \right)^{-1} (0). \tag{24}$$

Clearly, this set is larger than the set of critical points $(\partial^C G)^{-1}(0)$.

Let $(B_k)_{k\in\mathbb{N}}$ be a sequence of nonempty mini-batches sampled independently, uniformly at random in $[M]$. We can then devise the following iteration

$$\theta_{k+1} = \theta_k - \gamma_k d_k, \quad \text{where} \quad d_k \in \frac{1}{|B_k|} \sum_{i\in B_k} \mathcal{D}_i(\theta_k), \tag{25}$$

and $(\gamma_k)_{k\in\mathbb{N}}$ is a positive step sequence decaying at an appropriate rate. A natural question now is whether the sequence $(\theta_k)_{k\in\mathbb{N}}$ in (25) enjoys some convergence guarantees to the set of critical points in crit–G. For this, we will rely on the stochastic approximation method for differential inclusions with compact and convex-valued operators developed in [41], and used recently in [39, 42]. The idea is to view $(\theta_k)_{k\in\mathbb{N}}$ as a discrete-time stochastic process which asymptotically behaves as the (absolutely continuous) solution trajectories of the differential inclusion

$$\begin{cases} 0 \in \dot\theta(t) + \frac{1}{M}\sum_{i=1}^{M} \mathcal{D}_i(\theta(t)) & \text{for almost every } t \in \mathbb{R}, \\ \theta(0) = \theta_0, \end{cases} \tag{26}$$

whose stationary solutions are the critical points in (24). A key argument to invoke the results of [41], is to build an appropriate Lyapunov function and show that the function G is path differentiable, that is, it obeys the chain rule

$$\frac{\mathrm{d}}{\mathrm{d}t} G(\theta(t)) = \langle \dot\theta(t), v \rangle, \quad \forall v \in \frac{1}{M} \sum_{i=1}^{M} \mathcal{D}_i(\theta(t)). \tag{27}$$

While path differentiability can be shown (see later) for the finite sum under tameness/definability for the Clarke subdifferential, it seems very difficult to deal with the pointwise maximization and to prove path differentiability of g_i with the field $\mathcal{D}_i$.

The situation however changes if we work under (a part of) assumption (H.5), in which case (20) applies and (23) becomes

$$\partial^C G(\theta) = \partial^C \left(\frac{1}{M} \sum_{i=1}^{M} g_i(\theta) \right) \subset \frac{1}{M} \sum_{i=1}^{M} \partial^C g_i(\theta), \tag{28}$$

where

$$\partial^C g_i(\theta) = \left\{ \int_{\mathcal{C}_\varepsilon} \nabla_\theta \mathcal{L}(\theta, z_i + u) \mathrm{d}\mu(u) : \ \mu \in \mathcal{P}\Big(\operatorname{Argmax} \mathcal{L}(\theta, z_i + \mathcal{C}_\varepsilon) \Big) \right\}. \tag{29}$$

This gives the scheme in Algorithm 1.

Algorithm 1: SGD for PRO without smoothing.

```
Input: Step-sizes $(\gamma_k)_{k\in\mathbb{N}}$ according to (31); mini batch sizes $(b_k)_{k\in\mathbb{N}}$; $\varepsilon > 0$;
Input: Initialization $\theta_0$;
for $k = 0, \ldots$ do
    Draw independently uniformly at random a mini-batch $B_k \subset [M]$ of size $b_k$;
    for $i \in B_k$ do
        Solve $\bar{u}_i \in \operatorname{Argmax}_{u\in\mathcal{C}_\varepsilon} \mathcal{L}(\theta_k, z_i + u)$.
    $d_k = \frac{1}{b_k}\sum_{i\in B_k} \nabla_\theta \mathcal{L}(\theta_k, z_i + \bar{u}_i)$ ;
    $\theta_{k+1} = \theta_k - \gamma_k d_k$.
```

This algorithm enjoys the following guarantees.

Theorem 5.2. *Assume that (H.1)-(H.2) hold, that $\mathcal{L}(\cdot, z_i + u)$ is locally Lipschitz continuous for each $(z_i, u) \in \mathcal{Z} \times \mathcal{C}_\varepsilon$, and that $\mathcal{L}(\cdot, z)$ is differentiable with $\nabla_\theta \mathcal{L}(\cdot, \cdot)$ continuous in both its arguments on $\Theta \times \mathcal{Z}$. Then*

$$\text{crit–}G = \left(\frac{1}{M} \sum_{i=1}^{M} \partial^C g_i \right)^{-1} (0). \tag{30}$$

Suppose moreover that $\|\cdot\|$ and $\mathcal{L}$ are definable in an o-minimal structure., and that the step-sizes satisfy

$$\sum_{k\in\mathbb{N}} \gamma_k = +\infty \quad \textit{and} \quad \gamma_k = o\left(\frac{1}{\log k}\right). \tag{31}$$

Consider the sequence $(\theta_k)_{k\in\mathbb{N}}$ generated by Algorithm 1. Suppose that $(\theta_k)_{k\in\mathbb{N}}$ is almost surely bounded. We have, almost surely, that the set of cluster points $\emptyset \neq \mathfrak{C}((\theta_k)_{k\in\mathbb{N}}) \subset$ crit–G, that $(G(\theta_k))_{k\in\mathbb{N}}$ converges and G is constant on $\mathfrak{C}((\theta_k)_{k\in\mathbb{N}})$.

See Appendix A.7 for the proof.

A caveat of Algorithm 1 is that one has to solve the inner maximization problems to compute the subgradient approximation d_k as dictated by (29). We recall that this can be computationally challenging in general and iterative schemes do not come with any guarantees unless stringent assumptions are imposed on $\mathcal{L}$. To avoid this, one can appeal to the smoothing strategy as we develop now.

5.2 With smoothing

Denote for short the smoothed objective

$$G_{\tau,N}(\theta) \overset{\text{def}}{=} \frac{1}{M}\sum_{i=1}^{M} g^i_{\tau,N}(\theta), \quad \text{crit–}G_{\tau,N} = \left(\frac{1}{M} \sum_{i=1}^{M} \nabla g^i_{\tau,N} \right)^{-1} (0),$$

where $g^i_{\tau,N}$ is the smoothed version of g_i; see (15). In this section, our aim is to capitalize on the results of Theorem 4.8. However, this must be done with care. Indeed, the gradient of $G_{\tau,N}(\theta)$ is an inexact version of an actual Clarke subgradient element of $\frac{1}{M}\sum_{i=1}^{M}\partial^C g_i(\theta)$, and the crucial idea is then to make this error vanish or sufficiently small by appropriately choosing the smoothing parameter τ and the number of integration points N (see (H′.6)). In principle, from Theorem 4.8, this error term can be uniformly bounded for all θ in some compact set Ξ, and the bound depends on several constants which themselves depend Ξ. To apply this, one may think that it would be sufficient to assume that the algorithm sequence $(\theta_k)_{k\in\mathbb{N}}$ to live in Ξ to be done. There are however several issues with this reasoning. First $(\theta_k)_{k\in\mathbb{N}}$ will be a random process in our stochastic algorithms, and assuming that $(\theta_k)_{k\in\mathbb{N}}$ is bounded almost surely as usually done would entail that Ξ is itself random. One may turn to the more stringent assumption of uniform boundedness of the sequence of iterates, i.e., that Ξ is non-random. This is however not sufficient either as the choice of the sequence τ should avoid dependence on Ξ, if not known a priori, to generate $(\theta_k)_{k\in\mathbb{N}}\subset\Xi$ as this would induce a causality dilemma.

To tackle these issues, we have then elaborated two strategies. In the first one, τ and N are fixed but Ξ is unknown, in which case one can only get approximate critical points[3]. In the second algorithm, τ and N will vary with iterations while Ξ is known and fixed, to make the subgradient error term vanish and hence ensure convergence to critical points of the original objective. This will necessitate, however, to add a projection step on Ξ in the algorithm and to carefully account for the subgradient error term and this projector in the convergence analysis.

5.2.1 Fixed parameters

The idea here is to choose τ and N respectively small and large, but fixed. We propose the SGD algorithm:

```
Algorithm 2: SGD for PRO with smoothing.
  Input: Step-sizes $(\gamma_k)_{k\in\mathbb{N}}$ according to (31); $N$ integration samples $(u_j)_{j\in[N]}$ drawn independently
         uniformly at random in $\mathcal{C}_\varepsilon$; smoothing parameter $\tau$; mini batch sizes $(b_k)_{k\in\mathbb{N}}$;
  Input: Initialization $\theta_0$;
  for $k = 0, \ldots$ do
      Draw independently uniformly at random a mini-batch $B_k \subset [M]$ of size $b_k$ ;
      $d_k = \frac{1}{b_k}\sum_{i\in B_k}\nabla g^i_{\tau,N}(\theta_k)$, with $\nabla g^i_{\tau,N}(\theta_k) \overset{\mathrm{def}}{=} \sum_{j=1}^{N}\nabla_\theta\mathcal{L}(\theta, z_i+u_j)\frac{e^{\frac{\mathcal{L}(\theta,z_i+u_j)}{\tau}}}{\sum_{l=1}^{N}e^{\frac{\mathcal{L}(\theta,z_i+u_l)}{\tau}}}$ ;
      $\theta_{k+1} = \theta_k - \gamma_k d_k$.
```

We will denote ω_S the event of sampling the N integration points $(u_j)_{j\in[N]}$. Denote

$$\varsigma(N,\tau,t,\Xi) \overset{\mathrm{def}}{=} 4L_{\Xi,\mathcal{Z}}e^{\frac{\Delta\mathcal{L}^\Xi}{\tau}}\sqrt{\frac{2(p+t)\log N}{N}} + e^{\frac{\Delta\mathcal{L}^\Xi}{\tau}}\frac{4\left(\tau^{-1}L_{\Xi,\mathcal{Z}}(l_\Xi+1)+L_\Xi\right)D_\Xi}{N} + \psi(\tau).$$

For $\kappa > 0$, we also define the κ-critical set of G as

$$\mathrm{crit}_\kappa\text{–}G \overset{\mathrm{def}}{=} \left\{\theta : \ \mathrm{dist}\left(0, \frac{1}{M}\sum_{i=1}^{M}\partial^C g_i(\theta)\right) \le \kappa\right\}.$$

Recalling (30) from Theorem 5.2, crit_κ–G can be seen as κ-approximate critical points of G.

Theorem 5.3. *Assume that (H.1), (H.2), (H.3), (H.4) and (H.5) hold. Suppose moreover that $\mathcal{L}$ is definable in an o-minimal structure than contains also the* log-exp *structure. Run Algorithm 2 with γ_k chosen according to (31). The following holds,*

[3] An alternative strategy would be also to extend the results of [43] to the non-smooth case, but this would only allow for the choice $\tau = 1/\log(\log(N))$.

(i) *Conditioned on ω_S, if $(\theta_k)_{k\in\mathbb{N}}$ is almost surely bounded, then we have almost surely that the set of cluster points $\emptyset \neq \mathfrak{C}((\theta_k)_{k\in\mathbb{N}}) \subset \mathrm{crit}\text{–}G_{\tau,N}$, that $(G_{\tau,N}(\theta_k))_{k\in\mathbb{N}}$ converges and $G_{\tau,N}$ is constant on $\mathfrak{C}((\theta_k)_{k\in\mathbb{N}})$.*

(ii) *Let Ξ be any compact subset of $\mathrm{crit}\text{–}G_{\tau,N}$. For any $t > 0$ and N large enough, the following holds with probability at least $1 - 4N^{-t}$ on the event ω_S,*

$$\sup_{\bar{\theta}_{\tau,N}\in\Xi} \mathrm{dist}\left(0, \frac{1}{M}\sum_{i=1}^{M} \partial^C g_i(\bar{\theta}_{\tau,N})\right) \leq \varsigma(N,\tau,t,\Xi), \tag{32}$$

i.e., $\bar{\theta}_{\tau,N} \in \mathrm{crit}_{\varsigma(N,\tau,t,\Xi)}\text{–}G$ for all $\bar{\theta}_{\tau,N} \in \Xi$.

Assume that $\mathrm{crit}_\kappa\text{–}G$ is compact for some $\kappa > 0$. If τ and N are chosen small and large enough, respectively, such that $\varsigma(N,\tau,\Xi) \leq \kappa$, then with the same probability on ω_S, one has

$$\sup_{\bar{\theta}_{\tau,N}\in\Xi} \mathrm{dist}\left(\bar{\theta}_{\tau,N}, \mathrm{crit}\text{–}G\right) \leq \varphi(\varsigma(N,\tau,\Xi)) \leq \varphi(\kappa), \tag{33}$$

for a nonnegative increasing definable function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ continuous at 0 and which vanishes only at 0.

See Appendix A.8 for the proof.

A few remarks are in order.

Remark 5.4.

- *There exists an o-minimal structure containing globally subanalytic functions and the graph of $\exp$; see [31]. This family is very large and covers all applications we have in mind.*
- *The bounds (32) and (33) can also be established in expectation by invoking (19).*
- *The statement of Theorem 5.3(ii) asserts that bounded $\mathrm{crit}\text{–}G_{\tau,N}$ points are approximate $\mathrm{crit}\text{–}G$ points with high probability. As discussed previously (see page 13), there are different ways of choosing τ such that the rhs of (32) is small, and each choice will entail a different convergence rate to zero. Moreover, the choice that entails the fastest convergence would depend on Ξ. On the other hand, the choice $\tau = 1/\log(\log N)$ entails a slower rate but has the advantage of not depending on Ξ.*
- *N can be selected considering the dimension of the problem as advocated by the literature on Monte Carlo or Quasi Monte Carlo integration [37, 38].*

5.2.2 Iteration-dependent parameters

We now propose an alternative algorithm where Ξ is *fixed a priori* by the user. For this one solves the constrained version of (22) over a nonempty convex compact set Ξ, i.e.,

$$\min_{\theta\in\Xi} \left\{ G(\theta) \stackrel{\mathrm{def}}{=} \frac{1}{M}\sum_{i=1}^{M} g_i(\theta) \right\} \quad \text{where} \quad g_i(\theta) \stackrel{\mathrm{def}}{=} \max_{u\in\mathcal{C}_\varepsilon} \mathcal{L}(\theta, z_i + u).$$

This is a standard approach in many fields such as machine learning. This gives the scheme summarized in Algorithm 3 which incorporates the projector on Ξ, $\mathrm{P}_\Xi : \mathbb{R}^p \to \Xi$, which is single-valued by convexity of Ξ. Hence, by construction, we have $(\theta_k)_{k\in\mathbb{N}} \subset \Xi$ and uniform boundedness is in force.

Algorithm 3: Projected SGD for PRO with smoothing.

```
Input: Step-sizes $(\gamma_k)_{k\in\mathbb{N}}$ according to (34); number of integration points $(N_k)_{k\in\mathbb{N}}$; smoothing
       parameters $(\tau_k)_{k\in\mathbb{N}}$; mini batch sizes $(b_k)_{k\in\mathbb{N}}$; $\varepsilon > 0$;
Input: Initialization $\theta_0$;
for $k = 0, \ldots$ do
    Draw independently uniformly at random a mini-batch $B_k \subset [M]$ of size $b_k$ ;
    Draw $N_k$ samples $(u_j)_{j\in[N_k]}$ independently uniformly at random in $\mathcal{C}_\varepsilon$ ;
    $d_k = \frac{1}{b_k}\sum_{i\in B_k} \nabla g^i_{\tau_k,N_k}(\theta_k)$, with $\nabla g^i_{\tau_k,N_k}(\theta_k) \overset{\text{def}}{=} \sum_{j=1}^{N_k} \nabla_\theta \mathcal{L}(\theta, z_i+u_j)\frac{e^{\frac{\mathcal{L}(\theta,z_i+u_j)}{\tau_k}}}{\sum_{l=1}^{N_k} e^{\frac{\mathcal{L}(\theta,z_i+u_l)}{\tau_k}}}$ ;
    $\theta_{k+1} = \mathrm{P}_\Xi(\theta_k - \gamma_k d_k)$.
```

Because of the presence of the projector P_Ξ, the set of critical points we will be interested in now is

$$\text{crit–}G_\Xi = \left(\frac{1}{M}\sum_{i=1}^{M} \partial^C g_i + N_\Xi\right)^{-1}(0),$$

where N_Ξ is the normal cone to the set Ξ in the sense of convex analysis. Note that N_Ξ is not compact-valued. We will need to strengthen (31) to

$$\sum_{k\in\mathbb{N}} \gamma_k = +\infty \quad \text{and} \quad \sum_{k\in\mathbb{N}} \gamma_k^2 < +\infty. \tag{34}$$

We will assume the following on the parameters τ_k and N_k:

(H.7) As $k \to +\infty$, $\tau_k \to 0$ and $\frac{e^{\frac{\Delta\mathcal{L}^\Xi}{\tau_k}}}{\sqrt{N_k}} \to 0$.

We then have the following convergence result.

Theorem 5.5. *Assume that (H.1), (H.2), (H.3) and (H.5) hold. Suppose moreover that Ξ is a nonempty convex compact set and that $\|\cdot\|$, $\mathcal{L}$ and Ξ are definable in an o-minimal structure. Suppose that γ_k and (τ_k, N_k) (34) and (H.7) respectively. Run Algorithm 3 with this choice of parameters (γ_k, τ_k, N_k). We have, almost surely, that the set of cluster points $\emptyset \neq \mathfrak{C}((\theta_k)_{k\in\mathbb{N}}) \subset \text{crit–}G_\Xi$, that $(G(\theta_k))_{k\in\mathbb{N}}$ converges and G is constant on $\mathfrak{C}((\theta_k)_{k\in\mathbb{N}})$.*

See Appendix A.9 for the proof.

Remark 5.6. *Other algorithms can be used instead of projected SGD provided they are proven to converge with appropriate generalized subgradients (see [44] for instance).*

6 Numerical results

In this section we report experimental results on a classification problem using neural networks[4]. We compare our training algorithms to other training methods. Though we illustrate our algorithm on a classification problem, our abstract setting is general enough so that the algorithm applies to any robust optimization problem, even beyond machine learning. In fact, the core idea is to choose the loss $\mathcal{L}$ to adapt to the problem at hand. For instance, for the standard regression problem, one takes the usual quadratic loss, or the logistic loss for logistic regression.

[4] Implementations were made available at https://github.com/williampiat3/DistributionalRobustViaPointwiseSmoothing.

6.1 Experimental protocol

Dataset Experiments were all performed on the Avila dataset, introduced in [45]. This dataset is representative of a classification task – writer identification in medieval manuscripts through page layout features – with 8 input features and 12 classes. We centered and normalized the input features in a preprocessing step. The label distribution is uneven for the twelves classes (A:41%, B:0.048%, C:0.99%, D:3.4%, E:10%, F:19%, G:4.3%, H:5.0%, I:8.0%, W:0.43%, X:5.0%, Y:2.6%), with a class A that is far more present than the other labels. This dataset was selected for its moderate input dimension to keep Monte Carlo sampling needed for computing (15) reasonable, and because it has unevenly distributed labels which will help highlighting the compromise between generalization and robust learning.

Neural network classifier We build a 3 layer MLP network $f : \Theta \times \mathbb{R}^8 \to \mathbb{R}^{12}$ with two hidden layers of 200 neurons each and an output layer, resulting in $p = 44000$ parameter vector θ; i.e., $\Theta \subset \mathbb{R}^{44000}$. To comply with our regularity assumptions, we used the ELU activation function [46]. The loss used for training the network is the cross-entropy loss after a softmax step on the network output. More precisely, for a training example $z = (x, y)$, where $x \in \mathbb{R}^8$ is a feature vector and $y \in [12]$ is the (true) label, the loss is given by

$$\mathcal{L}(\theta, z) = -f(\theta, x)_y + \log\left(\sum_{j=1}^{12} e^{f(\theta,x)_j}\right) = \log\left(1 + \sum_{j \neq y} e^{f(\theta,x)_j - f(\theta,x)_y}\right)$$

where the subscript y (resp. j) stands for the y-th (resp. j-th) entry of the vector $f(\theta, x) \in \mathbb{R}^{12}$. Observe that this loss also verifies assumptions (H.1)-(H.5) and $\underline{\mathcal{L}}^{\Xi} = 0$.

Training methods For comparison purposes, we trained the MLP network with three different methods:

- Non-robust training: this corresponds to minimizing the non-robust loss $\frac{1}{M}\sum_{i=1}^{M} \mathcal{L}(\theta, z_i)$ using SGD;
- Adversarial PGD training [18]: this is a heuristic algorithm proposed to solve the PRO problem (22). It alternates between an SGD step on the parameters θ and a projected gradient ascent (PGD[5]) step as attempt to solve the inner maximization problem in (22), i.e., to find an adversarial attack;
- Our robust training method. In our experiments, we used Algorithm 2 to solve (22).

Our aim is to show that we can provably train a robust model using our algorithm, while being competitive with current state of the art procedures for robustifying neural networks, for different values of the perturbation radius ε. Throughout this section, we take $\mathcal{C}_\varepsilon = \mathbb{B}^q_\varepsilon(0)$ with typically $q = +\infty$ (see Table 1). All experiments were carried out with training methods run with the same number of epochs/iterations, batch size and therefore the same number of updates (see Table 1 for details).

Performance metrics For a pair $(x, y) \in \mathbb{R}^8 \times [12]$, let $F_\theta(x) = \text{Argmax}_{j\in[12]} f(\theta, x)_j$ be the predicted label. We denote $\mathbb{1} : \mathbb{R}^2 \to \{0, 1\}$ the mapping that returns 1 if its arguments are equal and 0 otherwise. In the numerical results, we will report three different performance metrics.

[5] The inner maximization problem involves indeed an ascent step and the acronym should rather be PGA. But we stick with the original acronym of [18] though it is not precise.

- **Test Accuracy**: We define it as the accuracy on a test dataset $\{(x_i, y_i) : i \in [N_{\text{test}}]\}$:

$$\text{Test Accuracy} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \mathbb{1}(F_\theta(x_i), y_i).$$

- **Adversarial Accuracy**: This metric is meant to represent the accuracy on an adversarial set given by applying a white-box PGD attack [18]. The attack depends on the loss function $\mathcal{L}$, a ball $\mathbb{B}^q_\varepsilon(0)$ in which the attack is constrained, and a tuple (x_i, y_i) of data points to be attacked. It reads

$$\text{Adversarial Accuracy} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \mathbb{1}(F_\theta(\hat{x}_i), y_i) \quad \text{where} \quad \hat{x}_i = \text{PGD}(\mathcal{L}, \mathbb{B}^q_\varepsilon(0), x_i, y_i).$$

- **Worst-case Robustness Accuracy**: This is defined as the worst-case accuracy when the data points undergo perturbations within a ball $\mathbb{B}^q_\varepsilon(0)$ (the same ball as for the PGD attack). More precisely, recall that $\mu_{\mathcal{U}}$ is the uniform measure on $\mathbb{B}^q_\varepsilon(0)$. For N perturbation realizations $(u_j)_{j=1}^N$ drawn i.i.d. from $\mu_{\mathcal{U}}$, we compute this metric as

$$\text{Robust Accuracy} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \min_{j \in [N]} \mathbb{1}(F_\theta(x_i + u_j), y_i).$$

  For large enough number of samples N, this metric is intended to show robustness as promoted when solving the PRO problem during the training.

In addition to these metrics, we also compute an estimate of the upper-bound of the Lipschitz constant of the neural network. To this end, we chose the LipSDP method [47] which is efficient, accurate and adequate for the size and structure of the networks used.

All above metrics were computed for each of the three training methods, for different values of ε. We made 100 runs for each configuration with different initializations to account for statistical variability. The training and test sets have been sampled once for all runs to ensure that the only source of variability originates from differences in the initialization of the training methods. The plots we will display show the median value and the quantiles at 0.1, 0.25, 0.75 and 0.9.

Remark 6.1. *Ultimately, the ideal metric for quantifying robustness is the population robust* 0-1 *loss. This is however computationally intractable. It is closely linked to the adversarial frequency [48]*

$$\mathbb{E}_{(x,y)\sim\rho_0}\left[\min_{u \in \mathbb{B}^q_\varepsilon(x)} \mathbb{1}(F_\theta(u), y)\right].$$

The robust accuracy attempts to estimate this quantity by drawing random samples in both the min and expectation. This however may suffer the curse of dimensionality when estimating the min value. The adversarial accuracy uses a heuristic in the form of an adversarial attack obtained by PGD [18], but as we argued above, the latter does not enjoy any convergence guarantee. The robust accuracy metric appears as a better representative metric of the robust behavior of a training method with the proviso that N is large.

6.2 Influence of the smoothing parameter and number of integration samples

As advocated by Theorem 5.3, the number of samples N in the Monte Carlo integration must be large enough and the smoothing sequence τ has to be take small enough (see the discussion in Remark 5.4). Of course, in practice, the choice of N is limited by obvious hardware memory reasons. Some workarounds can be found to extend the batch size, however they drastically increase the computation time. Therefore we need to check experimentally how the algorithm behaves in practice with different values of τ and N.

For each value of τ and N, the Monte Carlo integration samples in our robust training method are drawn in the hypercube $\mathbb{B}^{\infty}_{0.05}(0)$. The inner maximization problem is assessed and averaged on the test set by resorting to a fixed sampling of the perturbation set. This sampling is chosen larger (10^6) than the largest value explored in the set so as to maximize the chance that the metric is evaluated accurately when testing.

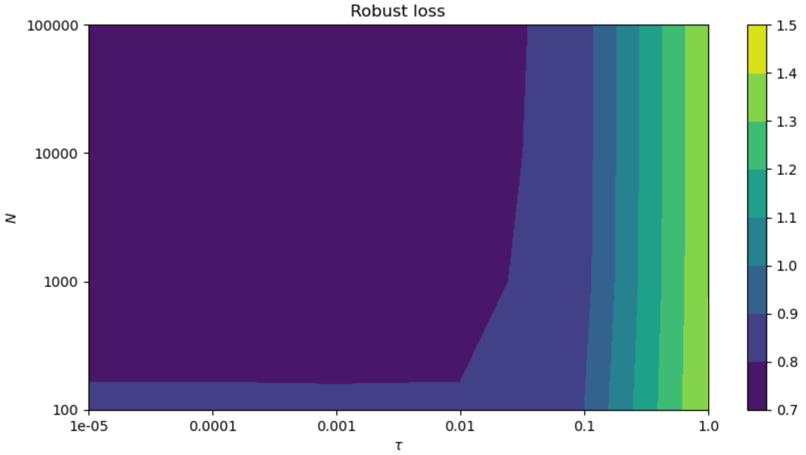

Figure 1: Influence of τ (x axis) and N (y axis) on the robust loss on test set (color scale). The darker the better.

Figure 1 shows the value of the robust loss for different values of τ and N, after a fixed number of iterations (see Table 1). The observed behavior is indeed as expected: the more we shrink τ and increase the number of samples N the smaller the error on the robust loss.

6.3 Robustness to perturbations

In this section, we present a few experiments illustrating the robustness to perturbations of a classifier learned with our training method, and provide comparison to the adversarial PGD training method (see above for details).

For these experiments, we set $\mathbb{B}^{\infty}_{\varepsilon}(0)$ as the perturbation set in the adversarial PGD training method and in our robust training method. In the latter, and motivated by the results of Figure 1, we kept fixed the number of integration samples $N = 5.10^5$ and the smoothing/regularization parameter $\tau = 10^{-4}$. See Section B for the details on the choice the parameters in these experiments.

We plot the evolution of the test accuracy (Figure 2), adversarial accuracy (Figure 3) and worst case accuracy (Figure 4) against the robustness radius ε. The plain line shows the median value and the shaded areas correspond to the quantiles. We also display in Figure 5 the evolution of the Lipschitz constant of the learned network estimated using LipSDP [47].

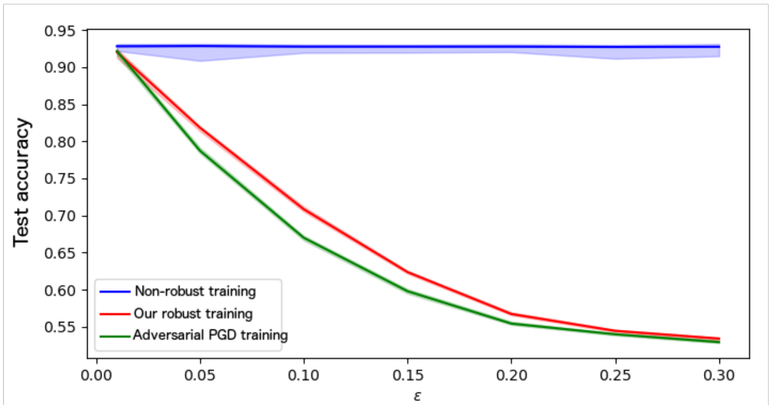

Figure 2: Test accuracy metric of the three training methods on Avila dataset: Non-robust training (blue), Adversarial PGD training (green), Our robust training (red), with median (plain line) and 10% and 90% quantiles over 50 trials.

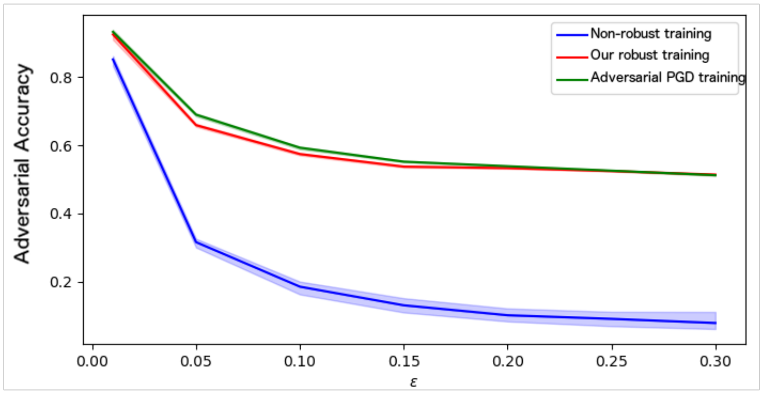

Figure 3: Adversarial accuracy metric of the three training methods on Avila dataset: Non-robust training (blue), Adversarial PGD training (green), Our robust training (red), with median (plain line) and 10% and 90% quantiles over 50 trials.

For the sake of completeness, and due to the fact that the loss values do not account for the complexity of the distribution of the predictions, we also provide three confusion matrices (i.e., percentage of predicted labels per each class in the test set). We normalized the confusion matrices column-wise as it allows to assess which percentage of the real label was split to which predicted label. The confusion matrices were computed for our three training methods and we set $\varepsilon = 0.3$. The results are displayed in Figure 6.

On all the figures above, the behaviour of our robust training method is competitive compared to the popular PGD-based adversarial training, and differences are in general small. We note that our robust training is better on the test accuracy for moderate perturbations (Figure 2), while adversarial PGD training appears to be slightly better in terms of adversarial accuracy (Figure 3). We note, as expected, a decrease in accuracy on the test dataset as the perturbation radius ε increases, but a better tolerance to perturbations/attacks since increasing the perturbation radius ε, we make the learned neural network-based classifier stable to larger adversarial attacks. This is symptomatic of the trade-off between robustness and generalization, see e.g., [49, 50, 51, 52]. Note also that the variability across runs is small (and highest for the non-robust training), confirming that the performance of all the training methods is reproducible from run to run and for different initializations.

The decrease in accuracy of our robust training method and the adversarial PGD training one on the test set can be further understood when considering the confusion matrices in Figure 6. Indeed, we see that these two robust learning methods aggregated the labels that were close to label A, namely labels C, D, E, F, G and H: this is due to class A being overly represented as it leans toward aggregating samples that are close. The labelling performed by the robustly trained networks on a test point assigns the label that has the greatest mass in the ε perturbation ball around the test point. This clearly means that some feature vectors originally in classes C, D, E, F, G and H are in fact within a ball $\mathbb{B}^q_\varepsilon(0)$ around those of label A. Class B has a very

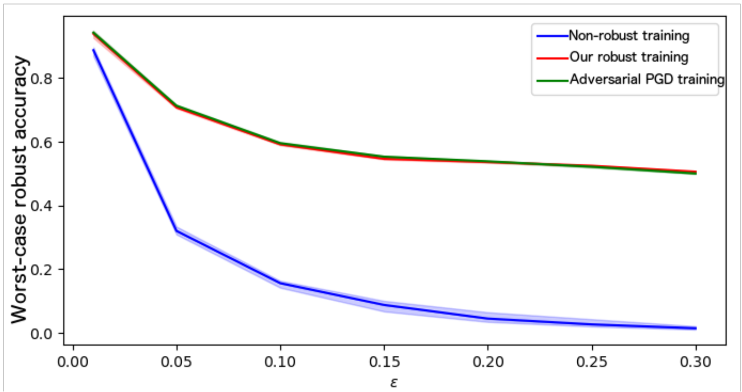

Figure 4: Worst-case robustness accuracy metric of the three training methods on Avila dataset: Non-robust training (blue), Adversarial PGD training (green), Our robust training (red), with median (plain line) and 10% and 90% quantiles over 50 trials.

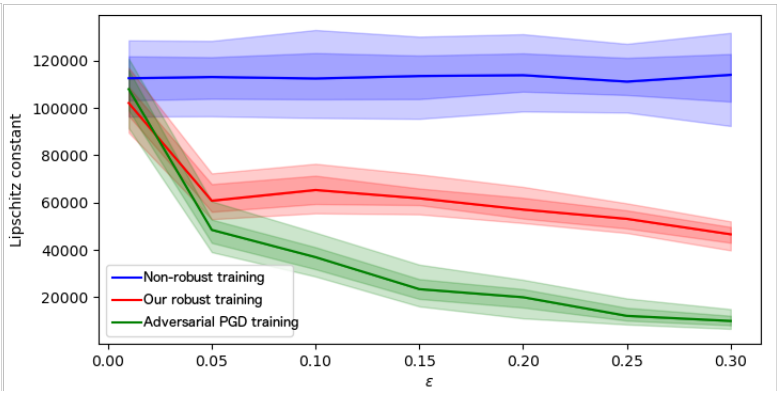

Figure 5: Lipschitz constant upper bound of the learned network by three training methods on Avila dataset: Non-robust training (blue), Adversarial PGD training (green), Our robust training (red), with median (plain line) and 10%, 25%, 75% and 90% quantiles over 50 trials.

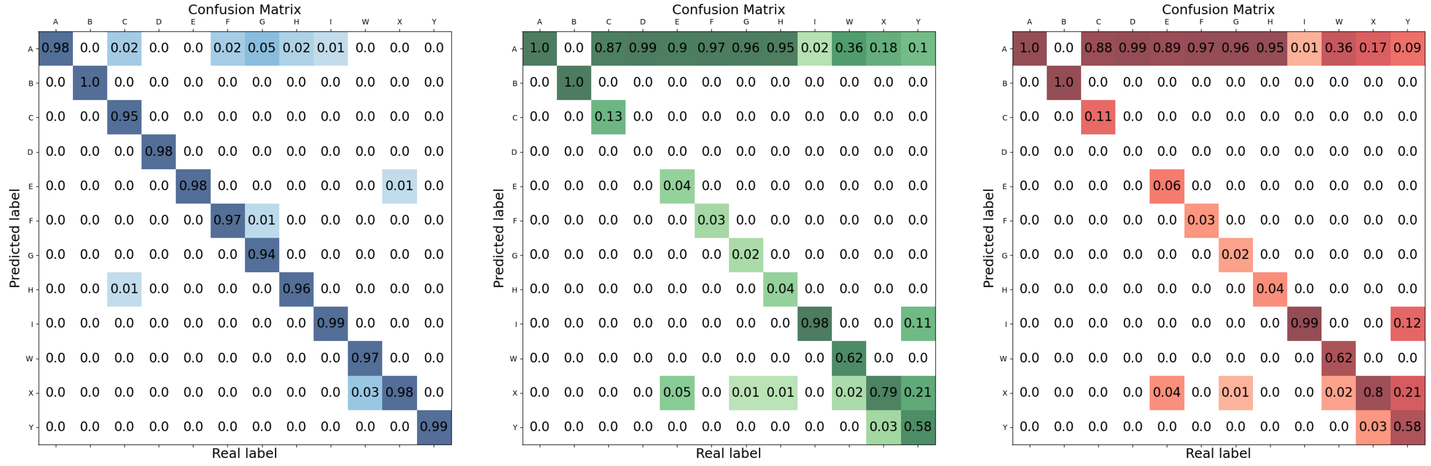

Figure 6: Confusion matrix for Non-robust training (left), Adversarial PGD training (middle) and Our robust training (right). All matrices are normalized column-wise to display the percentage of predicted labels per each class in the test set.

limited number of samples (5, only 0.048% of the dataset), however, it appears to be far enough from the other classes not to be confused with them.

As far as Figure 5 is concerned, one can clearly (and unsurprisingly) see that the our robust training method and the adversarial PGD training one tend to monotonically reduce the Lipschitz constant of the learned networks as ε increases. The fact that, on an adversarial set, the non-robust training method performs poorly compared to robustness-aware training methods is expected (Figure 3). Observe also that the performance of our robust training method is similar to the adversarial PGD training method on this dataset. Of course, adversarial attacks favor those models that have already been trained in an adversarial setting. The main drawback of our robust training method remains its computational cost compared to the adversarial PGD training method. For instance, under the hyperparameters in Section B with $\varepsilon = 0.3$, both training methods were performed on the same GPU A100 with the same number of epochs and updates. Adversarial PGD training took 47 min whereas our robust training took 11 hours. The former remains a very effective method for solving empirically the PRO problem though it lacks theoretical guarantees in general. Let us stress that, despite this increased cost, our robust training method can be more effectively parallelized as it only requires one expensive forward pass and one expensive backward pass whereas the adversarial PGD training method requires multiple iterations of both.

7 Conclusion

Numerically solving distributionally robust optimization problems (DRO) is a very challenging task in robust optimization that has ramifications in many fields including machine learning. While the DRO has been approached in the literature under stringent and unrealistic assumptions, this paper takes a new perspective and demonstrates that the DRO problem with a sufficiently small error can be approached with a pointwise counterpart (PRO). In order to solve the latter, we have designed a novel SGD-type algorithm that builds on an appropriate smoothing of the inner maximization problem of PRO and Monte Carlo sampling, hence avoiding the intractable inner maximization problem. Our approach is one of the first few to enjoy provable convergence guarantees without the need of high smoothness or convexity/concavity assumptions. On the numerical side, when applied to a real-world supervised neural network-based classification problem, our robust training has shown a similar or slightly better performance compared to state-of-the-art adversarial training.

The main limitation of our algorithm is that it comes at the expense of an overall higher computational cost. For instance, in large-scale machine learning applications with overparametrized neural networks involving a very large number of parameters and a high-dimensional input space, scalability of our robust training framework remains a challenge, particularly due to the simplicity of our Monte Carlo integration step. To overcome this limitation, one potential avenue for future investigation is the use of more sophisticated sampling strategies, such as those based on Langevin diffusion.

Data Availability

The Avila dataset [45] used in this study is available publicly on the UCI repository [53] at the link https://archive.ics.uci.edu/ml/datasets/Avila. It consists of page features from an XII century giant Latin copy of the Bible and the identity of the copyist that produced the page.

References

[1] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*, volume 28 of *Princeton Series in Applied Mathematics*. Princeton University Press, 2009.

[2] Dimitris Bertsimas, David B. Brown, and Constantine Caramanis. Theory and Applications of Robust Optimization. *SIAM Rev.*, 53(3):464–501, 2011.

[3] Dimitris Bertsimas and Melvyn Sim. The price of robustness. *Operations Research*, 52(1):35–53, 2004.

[4] Meisam Razaviyayn, Tianjian Huang, Songtao Lu, Maher Nouiehed, Maziar Sanjabi, and Mingyi Hong. Nonconvex min-max optimization: Applications, challenges, and recent theoretical advances. *IEEE Signal Processing Magazine*, 37(5):55–66, 2020.

[5] Kerem Uğurlu. Terminal wealth maximization under drift uncertainty. *Optimization*, 74(7):1743–1761, 2025.

[6] Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani, and Daniel Kuhn. Distributionally Robust Logistic Regression. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1576–1584, 2015.

[7] Jose H. Blanchet, Yang Kang, and Karthyek Rajhaa A. M. Robust Wasserstein profile inference and applications to machine learning. *J. Appl. Probab.*, 56(3):830–857, 2019.

[8] Jose H. Blanchet and Karthyek R. A. Murthy. Quantifying Distributional Model Risk via Optimal Transport. *Math. Oper. Res.*, 44(2):565–600, 2019.

[9] Aman Sinha, Hongseok Namkoong, and John C. Duchi. Certifying Some Distributional Robustness with Principled Adversarial Training. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[10] Ruidi Chen and Ioannis Ch. Paschalidis. Distributionally Robust Learning. *Foundations and Trends® in Optimization*, 4(1-2):1–243, 2020.

[11] Matthew Staib and Stefanie Jegelka. Distributionally Robust Optimization and Generalization in Kernel Methods. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9131–9141, 2019.

[12] Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Manag. Sci.*, 59(2):341–357, 2013.

[13] Hongseok Namkoong and John C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2208–2216, 2016.

[14] John C. Duchi, Peter W. Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Math. Oper. Res.*, 46(3):946–969, 2021.

[15] J. Goh and M. Sim. Distributionally robust optimization and its tractable approximations. *Operations Research*, 58(4):902–917, 2010.

[16] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):95–612, 2010.

[17] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[19] Nicholas Carlini and David A. Wagner. Towards Evaluating the Robustness of Neural Networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57. IEEE Computer Society, 2017.

[20] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. Ensemble Adversarial Training: Attacks and Defenses. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[21] John C. Duchi and Hongseok Namkoong. Variance-based Regularization with Convex Objectives. *J. Mach. Learn. Res.*, 20:68:1–68:55, 2019.

[22] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations. *Math. Program.*, 171(1-2):115–166, 2018.

[23] Rui Gao and Anton J. Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. arXiv:1604.02199 [math.OC], 2016.

[24] Laurent Meunier, Meyer Scetbon, Rafael Pinot, Jamal Atif, and Yann Chevaleyre. Mixed Nash Equilibria in the Adversarial Examples Game. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 7677–7687. PMLR, 2021.

[25] Jose Blanchet and Yang Kang. Semi-supervised learning based on distributionally robust optimization. *Data Analysis and Applications: Computational, Classification, Financial, Statistical and Stochastic Methods*, 5:1–33, 2020.

[26] Jie Wang, Rui Gao, and Yao Xie. Sinkhorn distributionally robust optimization. *CoRR*, abs/2109.11926, 2021.

[27] Waïss Azizian, Franck Iutzeler, and Jérôme Malick. Regularization for wasserstein distributionally robust optimization. arXiv:2205.08826, May 2022.

[28] G.D. Maso. *An Introduction to* Γ*-Convergence*. Progress in Nonlinear Differential Equations and Their Applications. Birkhäuser Boston, 2012.

[29] Michel Coste. *An introduction to semialgebraic geometry*. Dottorato di ricerca in matematica / Università di Pisa, Dipartimento di Matematica. Istituti Editoriali e Poligrafici Internazionali, Pisa, 2000.

[30] Michel Coste. *An introduction to o-minimal geometry*. Dottorato di ricerca in matematica / Università di Pisa, Dipartimento di Matematica. Istituti Editoriali e Poligrafici Internazionali, Pisa, 2000.

[31] Lou van den Dries and Chris Miller. Geometric categories and o-minimal structures. *Duke Mathematical Journal*, 84(2):497 – 540, 1996.

[32] Frank H. Clarke. *Optimization and Nonsmooth Analysis*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1990.

[33] R. T. Rockafellar and R. Wets. *Variational analysis*, volume 317. Springer Verlag, 1998.

[34] Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D. Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 14905–14916, Vancouver, BC, Canada, 2019.

[35] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[36] Bo Wei, William B. Haskell, and Sixiang Zhao. An inexact primal-dual algorithm for semi-infinite programming. *Math. Methods Oper. Res.*, 91(3):501–544, 2020.

[37] Russel E. Caflisch. Monte Carlo and quasi-Monte Carlo methods. *Acta Numerica*, 7:1–49, January 1998. Publisher: Cambridge University Press.

[38] Josef Dick, Frances Y. Kuo, and Ian H. Sloan. High-dimensional integration: The quasi-Monte Carlo way. *Acta Numerica*, 22:133–288, May 2013.

[39] Jérôme Bolte and Edouard Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Math. Program.*, 188(1):19–51, 2021.

[40] Jérôme Bolte and Edouard Pauwels. A mathematical model for automatic differentiation in machine learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[41] Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. Stochastic Approximations and Differential Inclusions. *SIAM J. Control. Optim.*, 44(1):328–348, 2005.

[42] C. Castera, J. Bolte, C. A. Sing-Long Févotte, and E. Pauwels. An inertial newton algorithm for deep learning. *Journal of Machine Learning Research*, 22(134):1–31, 2021.

[43] Vladislav B. Tadić and Arnaud Doucet. Asymptotic bias of stochastic gradient search. *Ann. Appl. Probab.*, 27(6):3255–3304, 2017.

[44] Andrzej Ruszczynski. Convergence of a stochastic subgradient method with averaging for nonsmooth nonconvex constrained optimization. *Optim. Lett.*, 14(7):1615–1625, 2020.

[45] C. De Stefano, M. Maniaci, F. Fontanella, and A. Scotto di Freca. Reliable writer identification in medieval manuscripts through page layout features: The Avila Bible case. *Engineering Applications of Artificial Intelligence*, 72:99–110, June 2018.

[46] Djork-Arno Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[47] Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George J. Pappas. Efficient and Accurate Estimation of Lipschitz Constants for Deep Neural Networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11423–11434, 2019.

[48] Osbert Bastani, Yani Ioannou, Leonidas Lampropoulos, Dimitrios Vytiniotis, Aditya V. Nori, and Antonio Criminisi. Measuring neural net robustness with constraints. 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 2613–2621, 2016.

[49] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness May Be at Odds with Accuracy. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[50] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C. Duchi, and Percy Liang. Adversarial Training Can Hurt Generalization. In *ICML 2019 Workshop on Identifying and Understanding Deep Learning Phenomena*, 2019. arXiv: 1906.06032.

[51] Y.Y. Yang, C. Rashtchian, H. Zhang, R.R Salakhutdinov, and K Chaudhur. A closer look at accuracy vs. robustness. In *Advances in neural information processing systems*, volume 33, pages 8588–8601, 2020.

[52] Elvis Dohmatob and Alberto Bietti. On the (non-)robustness of two-layer neural networks in different learning regimes. arXiv:2203.11864, Mar 2022.

[53] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[54] Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: a Hitchhiker's Guide*. Springer, Berlin; London, 2006.

[55] Chii-Ruey Hwang. Laplace's method revisited: Weak convergence of probability measures. *The Annals of Probability*, 8(6):1177–1182, 1980. Publisher: Institute of Mathematical Statistics.

[56] Dennis D. Cox, Robert M. Hardt, and Petr Klouček. Convergence of Gibbs Measures Associated with Simulated Annealing. *SIAM Journal on Mathematical Analysis*, 39(5):1472–1496, January 2008.

[57] Olivier Catoni. Simulated annealing algorithms and markov chains with rare transitions. In Jacques Azéma, Michel Émery, Michel Ledoux, and Marc Yor, editors, *Séminaire de Probabilités XXXIII*, pages 69–119, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg.

[58] Jonas Moritz Kohler and Aurelien Lucchi. Sub-sampled cubic regularization for non-convex optimization. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 1895–1904. JMLR.org, 2017.

[59] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inf. Theor.*, 57(3):1548–1566, March 2011.

[60] Damek Davis, Dmitriy Drusvyatskiy, Sham M. Kakade, and Jason D. Lee. Stochastic Subgradient Method Converges on Tame Functions. *Found. Comput. Math.*, 20(1):119–154, 2020.

[61] H. J. Kushner and G. G. Yin. *Stochastic Approximation Algorithms and Applications*, volume 35 of *Applications of Mathematics*. Springer, 1997.

[62] Jae Hyoung Lee and Tien-Son Pham. Openness, höder metric regularity, and höder continuity properties of semialgebraic set-valued maps. *SIAM Journal on Optimization*, 32(1):56–74, 2022.

[63] Jérôme Bolte, Tam Le, Eric Moulines, and Edouard Pauwels. Inexact subgradient methods for semialgebraic functions. *Mathematical Programming*, 2025. arXiv:2404.19517.

[64] R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, 4th edition, 2010.

[65] J. M. Lee. *Introduction to smooth manifolds*. Springer, 2003.

[66] H. Federer. Curvature measures. *Trans. Amer. Math. Soc.*, 93:418–491, 1959.

[67] Herbert Federer. *Geometric Measure Theory*. Classics in Mathematics. Springer Berlin Heidelberg, 1996.

[68] D. Salas and L. Thibault. On characterizations of submanifolds via smoothness of the distance function in Hilbert spaces. *Journal of Optimization Theory and Applications*, 182(1):189–210, 2019.

[69] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and D. Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, volume 9 of *JMLR Proceedings*, pages 249–256. JMLR.org, 2010.

A Proofs

A.1 Proof of Proposition 3.1

For any $z \in \mathcal{Z}$ and $\varepsilon \geq 0$, we have

$$\begin{aligned}\sup_{z' \in \mathcal{Z}} \left(\mathcal{L}(\theta, z') - \gamma \|z - z'\|^q\right) &\geq \sup_{\|z - z'\| \leq \varepsilon} \left(\mathcal{L}(\theta, z') - \gamma \|z - z'\|^q\right) \\ &\geq \sup_{\|z - z'\| \leq \varepsilon} \mathcal{L}(\theta, z') - \gamma \varepsilon^q.\end{aligned}$$

Taking the expectation on both sides, we get

$$\mathbb{E}_{z \sim \rho_0}\left[\sup_{\|z - z'\| \leq \varepsilon} \mathcal{L}(\theta, z')\right] \leq \gamma \varepsilon^q + \mathbb{E}_{z \sim \rho_0}\left[\sup_{z' \in \mathcal{Z}} \left(\mathcal{L}(\theta, z') - \gamma \|z - z'\|^q\right)\right].$$

In turn, since $\gamma \geq 0$ was arbitrary, we take the infimum on the right-hand side and use the duality identity (6) with $W_{\mathsf{c}} = W_q^q$, which holds under assumptions (H.2) and (H.1), to get the lower bound.

Let us turn to the upper-bound. We use (6) and that $\mathcal{L}(\theta, \cdot)$ is $l_{\mathcal{Z}}$-Lipschitz continuous by (H.3) to get that for any $\gamma \geq 0$,

$$\begin{aligned}\sup_{W_q(\rho, \rho_0) \leq \varepsilon} \mathbb{E}_{z \sim \rho}[\mathcal{L}(\theta, z)] &\leq \gamma \varepsilon^q + \mathbb{E}_{z \sim \rho_0}\left[\sup_{z' \in \mathcal{Z}} \left(\mathcal{L}(\theta, z') - \gamma \|z' - z\|^q\right)\right] \\ &\leq \gamma \varepsilon^q + \mathbb{E}_{z \sim \rho_0}\left[\sup_{z' \in \mathcal{Z}} \left(l_{\mathcal{Z}} \|z' - z\| - \gamma \|z' - z\|^q\right)\right] + \mathbb{E}_{z \sim \rho_0}\left[\mathcal{L}(\theta, z)\right] \\ &= \gamma \varepsilon^q + \mathbb{E}_{z \sim \rho_0}\left[\sup_{t \geq 0} \sup_{\|z' - z\| = t} \left(l_{\mathcal{Z}} \|z' - z\| - \gamma \|z' - z\|^q\right)\right] + \mathbb{E}_{z \sim \rho_0}\left[\mathcal{L}(\theta, z)\right] \\ &= \gamma \varepsilon^q + \sup_{t \geq 0}\left(l_{\mathcal{Z}} t - \gamma t^q\right) + \mathbb{E}_{z \sim \rho_0}\left[\mathcal{L}(\theta, z)\right] \\ &\leq \gamma \varepsilon^q + \sup_{t \geq 0}\left(l_{\mathcal{Z}} t - \gamma t^q\right) + \mathbb{E}_{z \sim \rho_0}\left[\sup_{\|z - z'\| \leq \varepsilon^{1/q}} \mathcal{L}(\theta, z')\right].\end{aligned}$$

Optimizing for t and after basic algebra, we get

$$\sup_{W_q(\rho, \rho_0) \leq \varepsilon^{1/q}} \mathbb{E}_{z \sim \rho}[\mathcal{L}(\theta, z)] \leq \gamma \varepsilon^q + (q - 1)\left(\frac{l_{\mathcal{Z}}}{q}\right)^{\frac{q}{q-1}} \gamma^{-\frac{1}{q-1}} + \mathbb{E}_{z \sim \rho_0}\left[\sup_{\|z - z'\| \leq \varepsilon^{1/q}} \mathcal{L}(\theta, z')\right].$$

This upper-bound is minimal for $\gamma = c_{q, l_{\mathcal{Z}}} \varepsilon^{-(q-1)}$, for an explicit constant $c_{q, l_{\mathcal{Z}}}$. Plugging this value of γ in the upper-bound, and noting that by compactness ((H.1)) and continuity ((H.2)), the sup is actually a max, we get the claim. □

A.2 Proof of Proposition 4.1

We provide a concise self-contained proof as the arguments are standard. We equip $\mathcal{P}(\mathcal{C}_\varepsilon)$ with the weak–$*$ topology. Since $\mathcal{C}_\varepsilon$ is compact, $\mathcal{P}(\mathcal{C}_\varepsilon)$ is weak–$*$ compact by [54, Theorem 15.11]. It is also convex. In addition, recall that the weak–$*$ topology is the weakest topology which makes the integration against continuous bounded functions a continuous linear form. In then follows from continuity of $\mathcal{L}(\theta, \cdot)$ and compactness of $\mathcal{C}_\varepsilon$ that $\mu \mapsto \int_{\mathcal{C}_\varepsilon} \mathcal{L}(\theta, z')\mathrm{d}\mu(z')$ is weak–$*$ continuous. It is known that $\mathrm{KL}(\cdot, \mu_{\mathcal{U}})$ is convex and lower semicontinuous in the weak–$*$ topology on $\mathcal{P}(\mathcal{C}_\varepsilon)$. Thus, since $\tau > 0$, the objective in (11) is convex and upper semicontinuous. This together with convex and weak–$*$ compactness of $\mathcal{P}(\mathcal{C}_\varepsilon)$ entail that (11) has a nonempty convex and weak–$*$ compact set of solutions. Uniqueness of the minimizer then follows from strong convexity of the probability simplex $\mathrm{KL}(\cdot, \mu_{\mathcal{U}})$ on $\mathcal{P}(\mathcal{C}_\varepsilon)$ thanks to the celebrated Pinsker's inequality. The closed form solution follows from standard calculus of variations and Lagrangian duality; see e.g., [36, Lemma 6.6]. □

A.3 Proof of Theorem 4.2

(i) For any $z \in \mathcal{Z}$, define $\psi_\tau(\theta, \mu) \overset{\mathrm{def}}{=} \int_{\mathcal{C}_\varepsilon} \mathcal{L}(\theta, z + z')\mathrm{d}\mu(z') - \tau \mathrm{KL}(\mu, \mu_{\mathcal{U}})$. Since KL is non-negative, we have for $\tau' \geq \tau$ and any θ and $\mu \in \mathcal{P}(\mathcal{C}_\varepsilon)$, it holds that

$$\psi_{\tau'}(\theta, \mu) \leq \psi_\tau(\theta, \mu).$$

Thanks to Proposition 4.1, for any (θ, τ), there is a unique solution to the maximization problem in (11). Let $\bar{\mu}_{\theta,\tau'} \in \mathcal{P}(\mathcal{C}_\varepsilon)$ be this solution for θ and τ'. Thus we have

$$g_\tau(\theta) = \max_{\mu \in \mathcal{P}(\mathcal{C}_\varepsilon)} \psi_\tau(\theta, \mu) \geq \psi_\tau(\theta, \bar{\mu}_{\theta,\tau'}) \geq \psi_{\tau'}(\theta, \bar{\mu}_{\theta,\tau'}) = g_{\tau'}(\theta),$$

i.e., $\tau \in \mathbb{R}_+ \mapsto g_\tau(\theta)$ is nonincreasing. Moreover, it is easy to see that $g_\tau(\theta) \leq g(\theta)$. In turn $\tau \in \mathbb{R}_+ \mapsto \mathbb{E}_{z\sim\rho_0}[g_\tau(\theta)]$ is also nonincreasing and $\mathbb{E}_{z\sim\rho_0}[g_\tau(\theta)] \leq \mathbb{E}_{z\sim\rho_0}[g(\theta)]$.

Now, g_τ is bounded from below uniformly in z (by the same lower-bound as $\mathcal{L}$ thanks to (H.2)), and so is $\mathbb{E}_{z\sim\rho_0}[g_\tau]$. Thus, by the monotone convergence theorem, $\mathbb{E}_{z\sim\rho_0}[g_\tau]$ converges pointwise to $\mathbb{E}_{z\sim\rho_0}[\sup_\tau g_\tau]$. Continuity of $\mathcal{L}$ (see (H.2)), compactness of $\mathcal{C}_\varepsilon$ and Θ allow to apply the Lebesgue dominated convergence theorem to the inner integral in (12) to infer that g_τ is continuous on Θ. Therefore, by Fatou's lemma,

$$\liminf_{\theta_k \to \theta} \mathbb{E}_{z\sim\rho_0}[g_\tau(\theta_k)] \geq \mathbb{E}_{z\sim\rho_0}\left[\lim_{\theta_k \to \theta} g_\tau(\theta_k)\right] = \mathbb{E}_{z\sim\rho_0}[g_\tau(\theta)],$$

whence we have that $(\mathbb{E}_{z\sim\rho_0}[g_\tau])_{\tau>0}$ is a sequence of lower-semicontinuous functions on Θ. Since Θ is closed, the Γ-convergence claim then follows from [28, Proposition 5.4, Remark 5.5 and Example 6.24].

(ii) We focus on the lower bound as the upper bound was shown above. Let us denote $z^\star \in \mathrm{Argmax}_{z' \in \mathcal{C}_\varepsilon} \mathcal{L}(\theta, z + z')$, where the set of maximizers is a nonempty compact set thanks to continuity of $\mathcal{L}(\theta, \cdot)$ and compactness of $\mathcal{C}_\varepsilon$. We then have by (H.3)

$$\begin{aligned}
\tau\log\left(\int_{\mathcal{C}_\varepsilon} e^{\frac{\mathcal{L}(\theta,z+z')}{\tau}}\mathrm{d}z'\right) &= \mathcal{L}(\theta,z+z^\star)+\tau\log\int_{\mathcal{C}_\varepsilon} e^{\frac{\mathcal{L}(\theta,z+z')-\mathcal{L}(\theta,z+z^\star)}{\tau}}\mathrm{d}z' \\
&= g(\theta)+\tau\log\int_{\mathcal{C}_\varepsilon} e^{\frac{\mathcal{L}(\theta,z+z')-\mathcal{L}(\theta,z+z^\star)}{\tau}}\mathrm{d}z' \\
&\geq g(\theta)+\tau\log\int_{\mathcal{C}_\varepsilon} e^{\frac{-\phi_{\mathcal{Z}}(\|z'-z^\star\|)}{\tau}}\mathrm{d}z'.
\end{aligned}$$

Convexity of $\mathcal{C}_\varepsilon$ entails that

$$\tau(\mathcal{C}_\varepsilon - z^\star)+z^\star = (1-\tau)z^\star+\tau\mathcal{C}_\varepsilon \subset \mathcal{C}_\varepsilon,$$

and thus

$$\begin{aligned}
\tau\log\left(\int_{\mathcal{C}_\varepsilon} e^{\frac{\mathcal{L}(\theta,z+z')}{\tau}}\mathrm{d}z'\right) &\geq g(\theta)+\tau\log\int_{\tau(\mathcal{C}_\varepsilon-z^\star)+z^\star} e^{-\frac{\phi_{\mathcal{Z}}(\|z'-z^\star\|)}{\tau}}\mathrm{d}z' \\
&= g(\theta)+\tau\log\left(\tau^m\int_{\mathcal{C}_\varepsilon} e^{-\frac{\phi_{\mathcal{Z}}(\tau\|z'-z^\star\|)}{\tau}}\mathrm{d}z'\right) \\
&\geq g(\theta)+\tau\log\left(\tau^m\int_{\mathcal{C}_\varepsilon} e^{-\frac{\phi_{\mathcal{Z}}\left(\tau D_{\mathcal{C}_\varepsilon}\right)}{\tau}}\mathrm{d}z'\right) \\
&= g(\theta)-m\tau\log(\tau^{-1})+\tau\log(\mu_{\mathcal{L}}(\mathcal{C}_\varepsilon))-\phi_{\mathcal{Z}}\left(\tau D_{\mathcal{C}_\varepsilon}\right).
\end{aligned} \tag{35}$$

Inserting this into (12), we get the bound. Taking the expectation on both sides of (13) yields

$$\sup_{\theta\in\Xi}\left|\mathbb{E}_{z\sim\rho_0}\left[g_\tau(\theta)\right]-\mathbb{E}_{z\sim\rho_0}\left[g(\theta)\right]\right| \leq h(\tau).$$

By assumption on $\phi_{\mathcal{Z}}$, we have $h(\tau)\ \to\ 0$ as $\tau\ \to\ 0^+$ whence we get that $\mathbb{E}_{z\sim\rho_0}\left[g_\tau\right]$ converges uniformly on Ξ to $\mathbb{E}_{z\sim\rho_0}\left[g\right]$. Combining this with lower-semicontinuity of $\mathbb{E}_{z\sim\rho_0}\left[g_\tau\right]$ for each τ shown above and closedness of Ξ, the Γ-convergence claim is now a consequence of [28, Proposition 5.2, Remark 5.3 and Example 6.24]. □

A.4 Proof of Theorem 4.4

Taking the expectation on both sides of (13) gives

$$\mathbb{E}_{z\sim\rho_0}\left[g_\tau(\theta)\right] \geq \mathbb{E}_{z\sim\rho_0}\left[g(\theta)\right]-h(\tau) \geq \mathbb{E}_{z\sim\rho_0}\left[\mathcal{L}(\theta,z)\right]-h(1).$$

The right-hand side being bounded from below by (H.2) and Θ being compact by assumption, we deduce that the sequence $\left(\mathbb{E}_{z\sim\rho_0}\left[g_\tau\right]+\iota_\Theta\right)_{\tau\geq 0}$ is equi-coercive (see [28, Definition 7.6 and Proposition 7.7]). The first claim on convergence of the minimal values (resp. minimizers) follows by combining equi-coercivity, the Γ-convergence claim of Theorem 4.2, and [28, Theorem 7.4 or Theorem 7.8] (resp. [28, Corollary 7.20]). The last claim is immediate from the second as the cluster point is unique. □

A.5 Proof of Theorem 4.6

We have

$$|g_{\tau,N}(\theta) - g(\theta)| \leq |g_\tau(\theta) - g(\theta)| + |g_{\tau,N}(\theta) - g_\tau(\theta)| \leq h(\tau) + |g_{\tau,N}(\theta) - g_\tau(\theta)|$$

where we invoked Theorem 4.2, and more precisely (13), in the last inequality. It remains to bound the last term in the above bound. This is the subject of the following lemma.

Lemma A.1. *Under the assumptions (H.1)-(H.2), the following holds for any compact set $\Xi \subset \Theta$.*

(i) We have

$$\sup_{\theta \in \Xi} \mathbb{E}\left[|g_{\tau,N}(\theta) - g_\tau(\theta)|\right] \leq \frac{\tau e^{\frac{\Delta\mathcal{L}^\Xi}{\tau}}}{\sqrt{N}}.$$

(ii) If, moreover, (H.4) holds, then for any $t > 0$

$$\sup_{\theta \in \Xi} |g_{\tau,N}(\theta) - g_\tau(\theta)| \leq \tau e^{\frac{\Delta\mathcal{L}^\Xi}{\tau}} \sqrt{\frac{(p+t)\log N}{2N}} + e^{\frac{\Delta\mathcal{L}^\Xi}{\tau}} \frac{4 l_\Xi D_\Xi}{N},$$

with probability at least $1 - 2N^{-t}$.

(iii) Suppose that τ is a function of N, say τ_N, such that $\tau_N e^{\frac{\Delta\mathcal{L}^\Xi}{\tau_N}} \sqrt{\frac{\log N}{N}} + \frac{e^{\frac{\Delta\mathcal{L}^\Xi}{\tau_N}}}{N} \to 0$ as $N \to +\infty$, then

(a) $\sup_{\theta\in\Xi} \mathbb{E}\left[|g_{\tau_N,N}(\theta) - g_{\tau_N}(\theta)|\right] \underset{N\to+\infty}{\longrightarrow} 0.$

(b) If, moreover, (H.4) holds, then the event

$$g_{\tau_N,N}(\theta) - g_{\tau_N}(\theta) \underset{N\to+\infty}{\longrightarrow} 0 \quad \textit{for all} \quad \theta \in \Xi.$$

holds almost surely.

Proof. To lighten the notation, denote

$$S_N^\theta \stackrel{\text{def}}{=} \frac{1}{N} \sum_{k=1}^N e^{\frac{\mathcal{L}(\theta, z + z'_k)}{\tau}}.$$

(i) Since the z'_k's are independent samples from the uniform distribution supported on $\mathcal{C}_\varepsilon$, we get

$$\mathbb{E}[S_N^\theta] = \frac{1}{\mu_{\mathcal{L}}(\mathcal{C}_\varepsilon)} \int_{\mathcal{C}_\varepsilon} e^{\frac{\mathcal{L}(\theta, z+z')}{\tau}} \mathrm{d}z'.$$

We then have

$$g_{\tau,N}(\theta) - g_\tau(\theta) = \tau \log\left(\frac{S_N^\theta}{\mathbb{E}[S_N^\theta]}\right).$$

Using the standard inequality $\log(1+t) \leq t$ for $t \geq 0$, we have

$$\begin{aligned}
|g_{\tau,N}(\theta) - g_\tau(\theta)| &= \tau \log\left(\frac{S_N^\theta}{\mathbb{E}[S_N^\theta]}\right) \mathbb{1}\left(S_N^\theta \geq \mathbb{E}[S_N^\theta]\right) + \tau \log\left(\frac{\mathbb{E}[S_N^\theta]}{S_N^\theta}\right) \mathbb{1}\left(S_N^\theta \leq \mathbb{E}[S_N^\theta]\right) \\
&\leq \tau \frac{S_N^\theta - \mathbb{E}[S_N^\theta]}{\mathbb{E}[S_N^\theta]} \mathbb{1}\left(S_N^\theta \geq \mathbb{E}[S_N^\theta]\right) + \tau \frac{\mathbb{E}[S_N^\theta] - S_N^\theta}{S_N^\theta} \mathbb{1}\left(S_N^\theta \leq \mathbb{E}[S_N^\theta]\right) \\
&\leq \tau e^{-\frac{\underline{\mathcal{L}}^\Xi}{\tau}} \left(\left(S_N^\theta - \mathbb{E}[S_N^\theta]\right) \mathbb{1}\left(S_N^\theta \geq \mathbb{E}[S_N^\theta]\right) + \left(\mathbb{E}[S_N^\theta] - S_N^\theta\right) \mathbb{1}\left(S_N^\theta \leq \mathbb{E}[S_N^\theta]\right)\right) \\
&= \tau e^{-\frac{\underline{\mathcal{L}}^\Xi}{\tau}} |S_N^\theta - \mathbb{E}[S_N^\theta]|. \qquad (36)
\end{aligned}$$

Denote $y_i = e^{\frac{\mathcal{L}(\theta, z+z_i')}{\tau}}$. Jensen's inequality and and the fact that the random variables y_i are i.i.d. yield

$$\mathbb{E}[|S_N^\theta - \mathbb{E}[S_N^\theta]|] \leq \mathbb{E}[(S_N^\theta - \mathbb{E}[S_N^\theta])^2]^{1/2} = \frac{1}{\sqrt{N}} \mathbb{E}[(y - \mathbb{E}[y])^2]^{1/2} \leq \frac{1}{\sqrt{N}} \mathbb{E}[y^2]^{1/2} \leq \frac{e^{\frac{\overline{\mathcal{L}}^\Xi}{\tau}}}{\sqrt{N}}. \quad (37)$$

This shows the bound in expectation.

(ii) Let us turn to the bound in probability. Using again that y_i are i.i.d. and bounded (they live in the interval $[e^{\underline{\mathcal{L}}^\Xi/\tau}, e^{\overline{\mathcal{L}}^\Xi/\tau}]$), we are in position to invoke Hoeffding's inequality to obtain that for any $\epsilon > 0$,

$$\begin{aligned}
\mathbb{P}\left(|g_{\tau,N}(\theta) - g_\tau(\theta)| > \epsilon\right) &\leq \mathbb{P}\left(|S_N^\theta - \mathbb{E}[S_N^\theta]| > e^{\underline{\mathcal{L}}^\Xi/\tau} \epsilon/\tau\right) \\
&\leq 2\exp\left(-\frac{2N^2 e^{2\underline{\mathcal{L}}^\Xi/\tau} \epsilon^2}{N\left(e^{\overline{\mathcal{L}}^\Xi/\tau} - e^{\underline{\mathcal{L}}^\Xi/\tau}\right)^2 \tau^2}\right) \\
&\leq 2\exp\left(-\frac{2N e^{-2\Delta\mathcal{L}^\Xi/\tau} \epsilon^2}{\tau^2}\right).
\end{aligned}$$

Taking $\epsilon = \tau e^{\frac{\Delta\mathcal{L}^\Xi}{\tau}} \sqrt{\frac{\kappa \log N}{2N}}$, for any $\kappa > 0$, we get

$$\mathbb{P}\left(|g_{\tau,N}(\theta) - g_\tau(\theta)| \geq \tau e^{\frac{\Delta\mathcal{L}^\Xi}{\tau}} \sqrt{\frac{\kappa \log N}{2N}}\right) \leq 2N^{-\kappa}. \quad (38)$$

The rest of the proof uses a covering argument. Recall the covering number of Ξ in the norm $\|\cdot\|$ at resolution $\delta > 0$ is the smallest number, $N(\Xi, \delta)$, such that Ξ can be covered with balls $\mathbb{B}_\delta(\theta_i)$, $\theta_i \in \Xi$, $i \in [N(\Xi, \delta)]$, i.e., $\Xi \subseteq \bigcup_{i \in [N(\Xi,\delta)]} \mathbb{B}_\delta(\theta_i)$. The finite set of points $\Xi_\delta \overset{\text{def}}{=} \{\theta_i : i \in [N(\Xi, \delta)]\}$ is called a (proper) δ-net of Ξ.

For any $\theta \in \Xi$, there exists $\theta_i \in \Xi_\delta$ such that $\|\theta - \theta_i\| \leq \delta$. We then have

$$|g_{\tau,N}(\theta) - g_\tau(\theta)| \leq |g_{\tau,N}(\theta_i) - g_\tau(\theta_i)| + |g_{\tau,N}(\theta) - g_{\tau,N}(\theta_i)| + |g_\tau(\theta) - g_\tau(\theta_i)|.$$

Arguing as in (36), and by the mean value theorem, we have

$$\begin{aligned}
|g_{\tau,N}(\theta) - g_{\tau,N}(\theta_i)| &\leq \tau e^{-\frac{\underline{\mathcal{L}}^\Xi}{\tau}} |S_N^\theta - S_N^{\theta_i}| \\
&\leq \tau e^{-\frac{\underline{\mathcal{L}}^\Xi}{\tau}} \frac{1}{N} \sum_{k=1}^N \left| e^{\frac{\mathcal{L}(\theta, z+z_k')}{\tau}} - e^{\frac{\mathcal{L}(\theta_i, z+z_k')}{\tau}} \right| \\
&\leq \tau e^{-\frac{\underline{\mathcal{L}}^\Xi}{\tau}} \frac{1}{N} \sum_{k=1}^N \frac{e^{\frac{\overline{\mathcal{L}}^\Xi}{\tau}}}{\tau} \left| \mathcal{L}(\theta, z+z_k') - \mathcal{L}(\theta_i, z+z_k') \right| \\
&\leq e^{\frac{\Delta \mathcal{L}^\Xi}{\tau}} l_\Xi \|\theta - \theta_i\| \leq e^{\frac{\Delta \mathcal{L}^\Xi}{\tau}} l_\Xi \delta,
\end{aligned}$$

where we used (H.4) in the last inequality. The same bound is valid on $|g_\tau(\theta) - g_\tau(\theta_i)|$. Thus

$$\sup_{\theta \in \Xi} |g_{\tau,N}(\theta) - g_\tau(\theta)| \leq \sup_{i \in [N(\Xi,\delta)]} |g_{\tau,N}(\theta_i) - g_\tau(\theta_i)| + 2e^{\frac{\Delta \mathcal{L}^\Xi}{\tau}} l_\Xi \delta.$$

Taking $\kappa = p + t$ in (38) for any $t > 0$, and using a union bound, we have

$$\mathbb{P}\left(\sup_{i \in [N(\Xi,\delta)]} |g_{\tau,N}(\theta_i) - g_\tau(\theta_i)| \geq \tau e^{\frac{\Delta \mathcal{L}^\Xi}{\tau}} \sqrt{\frac{(p+t)\log N}{2N}} \right) \leq 2N(\Xi, \delta) N^{-(p+t)}. \tag{39}$$

Set $\delta = 2D_\Xi/N$. Since $\Xi \subset \mathbb{B}_{D_\Xi/2}(0)$, we have from standard estimates of the covering number that

$$N(\Xi, \delta) \leq \left(1 + \frac{D_\Xi}{\delta}\right)^p \leq N^p.$$

Plugging this into (39) gives the claim.

(iii) Let $\epsilon_N \overset{\text{def}}{=} \tau_N e^{\frac{\Delta \mathcal{L}^\Xi}{\tau_N}} \sqrt{\frac{(2+p)\log N}{2N}} + e^{\frac{\Delta \mathcal{L}^\Xi}{\tau_N}} \frac{4 l_\Xi D_\Xi}{N}$.

(a) The convergence in expectation is immediate.

(b) For the almost sure convergence, we start from claim (ii) to see that

$$\mathbb{P}\left(\sup_{\theta \in \Xi} |g_{\tau_N,N}(\theta) - g_{\tau_N}(\theta)| > \epsilon_N \right) \leq 2N^{-2}.$$

Since the right-hand side above is summable in N, we conclude by the (first) Borel-Cantelli lemma that with probability one

$$\limsup_{N \to +\infty} \sup_{\theta \in \Xi} |g_{\tau_N,N}(\theta) - g_{\tau_N}(\theta)| = 0,$$

whence almost sure convergence is immediate.

□

A.6 Proof of Theorem 4.8

By the triangle inequality

$$\operatorname{dist}\left(\nabla g_{\tau,N}(\theta), \partial^C g(\theta)\right) \leq \operatorname{dist}\left(\nabla g_\tau(\theta), \partial^C g(\theta)\right) + \|\nabla g_\tau(\theta) - \nabla g_{\tau,N}(\theta)\|\,.$$

We will bound each term separately.

Lemma A.2. *Assume that (H.1), (H.2), (H.3) and (H.5) hold. Then*

$$\sup_{\theta\in\Xi} \operatorname{dist}\left(\nabla g_\tau(\theta), \partial^C g(\theta)\right) \to 0 \quad \textit{as} \quad \tau \to 0^+. \tag{40}$$

To show Lemma A.2, one observes that under our assumptions,

$$\nabla g_\tau(\theta) = \int_{\mathcal{C}_\varepsilon} \nabla_\theta \mathcal{L}(\theta, z+z') \frac{\exp\left(\frac{\mathcal{L}(\theta,z+z')}{\tau}\right)}{\int_{\mathcal{C}_\varepsilon} \exp\left(\frac{\mathcal{L}(\theta,z+v)}{\tau}\right) \mathrm{d}v} \mathrm{d}z', \tag{41}$$

which is an expectation with respect to a Gibbs measure indexed by τ (and (z, θ)) and supported on $\mathcal{C}_\varepsilon$. In view of the rule (5), the proof will then amount to showing that as $\tau \to 0^+$, the family of such Gibbs measures has all its cluster points in the narrow topology (equivalent to the weak–$*$ topology) in $\mathcal{P}\Big(\operatorname{Argmax} \mathcal{L}(\theta, \mathcal{C}_\varepsilon)\Big)$, or that it even converges in the weak–$*$ topology to a measure supported on $\operatorname{Argmax}_{z'\in\mathcal{C}} \mathcal{L}(\theta, z+z')$[6].

Proof. To lighten notation in the proof, we drop the subscript in $\mathcal{C}_\varepsilon$. Let the (Gibbs) probability measure[7]

$$\mathrm{d}\mu_\tau(z') \overset{\text{def}}{=} \frac{e^{\frac{\mathcal{L}(\theta,z+z')}{\tau}}}{\int_{\mathcal{C}} e^{\frac{\mathcal{L}(\theta,z+v)}{\tau}} \mathrm{d}v} \mathrm{d}z'.$$

Denote $\mathcal{M} \overset{\text{def}}{=} \operatorname{Argmax} \mathcal{L}(\theta, z+\mathcal{C})$ (we drop the dependence of $\mathcal{M}$ on z and θ to lighten notation). Compactness of $\mathcal{C}$ and continuity of $\mathcal{L}(\theta, \cdot)$ imply that $\mathcal{M}$ is a nonempty compact set. Without loss of generality, we assume that $\max \mathcal{L}(\theta, z+\mathcal{C}) = \mathcal{L}(\theta, z+\mathcal{M}) = 0$ (otherwise, one can use a simple translation argument). The proof of this claim is inspired by standard arguments in the literature of simulated annealing and Markov chains (see e.g. [57, Proposition 1.2] or [55, Corollary 2.1 and Proposition 2.3])[8]. We provide a self-contained proof adapted to our setting.

Given $\epsilon > 0$, we define

$$\mathcal{U}^\epsilon = \{u \in \mathcal{C} :\ \mathcal{L}(\theta, z+u) \geq -\epsilon\}\,.$$

By assumption (H.3), ϕ_Z is strictly increasing and vanishes only at 0, and thus, for any u in the open tubular neighborhood of radius $\phi_{\mathcal{Z}}^{-1}(\epsilon) > 0$ around $\mathcal{M}$, we have

$$-\mathcal{L}(\theta, z+u) = \mathcal{L}(\theta, z+\mathcal{M}) - \mathcal{L}(\theta, z+u) \leq \phi_{\mathcal{Z}}(\|\bar{z} - u\|) < \epsilon,$$

[6] This is reminiscent of works on simulated annealing where τ is the temperature parameter; see e.g. [55, 56]. Our context is however different and in particular, $\mathcal{C}_\varepsilon$ is not the whole space nor it is a finite set nor a compact submanifold.

[7] Strictly speaking, we should also index it with (z, θ). In this proof, we will drop this to lighten notation.

[8] We thank the reviewer for raising similar arguments.

where $\bar{z}$ is the closest element of u in $\mathcal{M}$. We then deduce that $\mathcal{U}^\epsilon$ contains the open tubular neighborhood of radius $\phi_{\mathcal{Z}}^{-1}(\epsilon)$ around $\mathcal{M}$. This implies that $\mu_{\mathcal{L}}(\mathcal{U}^\epsilon) > 0$. We then have

$$\begin{aligned}\mu_\tau(\mathcal{C} \setminus \mathcal{U}^\epsilon) &= \frac{\int_{\mathcal{C}\setminus\mathcal{U}^\epsilon} e^{\frac{\mathcal{L}(\theta,z+z')}{\tau}} \mathrm{d}z'}{\int_{\mathcal{C}} e^{\frac{\mathcal{L}(\theta,z+v)}{\tau}} \mathrm{d}v} \\ &\leq \frac{e^{\frac{-\epsilon}{\tau}} \mu_{\mathcal{L}}(\mathcal{C} \setminus \mathcal{U}^\epsilon)}{\int_{\mathcal{U}^{\epsilon/2}} e^{\frac{\mathcal{L}(\theta,z+v)}{\tau}} \mathrm{d}v} \\ &\leq \frac{e^{\frac{-\epsilon}{\tau}} \mu_{\mathcal{L}}(\mathcal{C} \setminus \mathcal{U}^\epsilon)}{e^{\frac{-\epsilon}{2\tau}} \mu_{\mathcal{L}}(\mathcal{U}^\epsilon)} \leq e^{\frac{-\epsilon}{2\tau}} \frac{\mu_{\mathcal{L}}(\mathcal{C})}{\mu_{\mathcal{L}}(\mathcal{U}^\epsilon)}.\end{aligned}$$

Passing to the limit as $\tau \to 0^+$ we get

$$\mu_\tau(\mathcal{C} \setminus \mathcal{U}^\epsilon) \to 0.$$

Compactness of $\mathcal{C}$ implies that $\mathcal{P}(\mathcal{C})$ is weak–$*$ compact by [54, Theorem 15.11]. The family $(\mu_\tau)_{\tau\geq 0}$ is then sequentially precompact by Prokhorov's theorem. Let $(\mu_{\tau_k})_{k\in\mathbb{N}}$ be a subsequence with weak–$*$ cluster point $\bar{\mu}$. We then have $\bar{\mu}(\mathcal{C} \setminus \mathcal{U}^\epsilon) = 0$, and since ϵ is arbitrary, we get that $\bar{\mu}$ is supported on $\mathcal{U}^0 = \mathcal{M}$. We thus infer, in view of continuity of $\nabla_\theta \mathcal{L}(\theta, \cdot)$ (by (H.5)) that

$$\nabla g_{\tau_k}(\theta) = \int_{\mathcal{C}} \nabla_\theta \mathcal{L}(\theta, z + z') \mathrm{d}\mu_{\tau_k}(z') \xrightarrow[k\to+\infty]{} \int_{\mathcal{C}} \nabla_\theta \mathcal{L}(\theta, z + z') \mathrm{d}\bar{\mu}(z') \in \partial^C g(\theta),$$

where we used (5) in the last inclusion. This is being true for any subsequence $(\mu_{\tau_k})_{k\in\mathbb{N}}$, we conclude that all cluster points of $(\nabla g_{\tau_k}(\theta))_{k\in\mathbb{N}}$ belong to $\partial^C g(\theta)$ which is equivalent to (40). □

Lemma A.3. *Assume that (H.1), (H.2) and (H.5) hold. Then the following holds for any compact set $\Xi \subset \Theta$.*

(i) We have

$$\sup_{\theta\in\Xi} \mathbb{E}\left[\|\nabla g_{\tau,N}(\theta) - \nabla g_\tau(\theta)\|\right] \leq \frac{2L_{\Xi,\mathcal{Z}} e^{\frac{\Delta\mathcal{L}^\Xi}{\tau}}}{\sqrt{N}}. \tag{42}$$

(ii) If, moreover, (H.4) holds, then for any $t > 0$ and N large enough,

$$\sup_{\theta\in\Xi} \|\nabla g_\tau(\theta) - \nabla g_{\tau,N}(\theta)\| \leq 4L_{\Xi,\mathcal{Z}} e^{\frac{\Delta\mathcal{L}^\Xi}{\tau}} \sqrt{\frac{2(p+t)\log N}{N}} + e^{\frac{\Delta\mathcal{L}^\Xi}{\tau}} \frac{4\left(\tau^{-1} L_{\Xi,\mathcal{Z}}(l_\Xi + 1) + L_\Xi\right) D_\Xi}{N} \tag{43}$$

with probability at least $1 - 4N^{-t}$.

(iii) Suppose that τ is a function of N, say τ_N, such that $e^{\frac{\Delta\mathcal{L}^\Xi}{\tau_N}} \sqrt{\frac{\log N}{N}} + \frac{\tau_N^{-1} e^{\frac{\Delta\mathcal{L}^\Xi}{\tau_N}}}{N} \to 0$ as $N \to +\infty$, then

(a) $\sup_{\theta\in\Xi} \mathbb{E}\left[\|\nabla g_{\tau,N}(\theta) - \nabla g_\tau(\theta)\|\right] \xrightarrow[N\to+\infty]{} 0.$

(b) If, in addition, (H.4) holds, then the event

$$\sup_{\theta\in\Xi} \|\nabla g_\tau(\theta) - \nabla g_{\tau,N}(\theta)\| \xrightarrow[N\to+\infty]{} 0 \quad \textit{for all} \quad \theta \in \Xi$$

holds almost surely.

Proof. To lighten notation in the proof, we drop the super- and subscript in $\mathcal{C}_\varepsilon$. Denote the probability measures

$$\mathrm{d}\mu_\tau^\theta(z') \overset{\text{def}}{=} \frac{1}{\mu_{\mathcal{L}}(\mathcal{C})} \frac{e^{\frac{\mathcal{L}(\theta,z+z')}{\tau}}}{S_\tau^\theta} \mathrm{d}z' \text{ and } \mathrm{d}\mu_{\tau,N}^\theta(z') \overset{\text{def}}{=} \frac{1}{N} \sum_{k=1}^N \frac{e^{\frac{\mathcal{L}(\theta,z+z'_k)}{\tau}}}{S_{\tau,N}^\theta} \delta_{z'_k}$$

where

$$S_\tau^\theta \overset{\text{def}}{=} \frac{1}{\mu_{\mathcal{L}}(\mathcal{C})} \int_{\mathcal{C}} e^{\frac{\mathcal{L}(\theta,z+z')}{\tau}} \mathrm{d}z' \text{ and } S_{\tau,N}^\theta \overset{\text{def}}{=} \frac{1}{N} \sum_{k=1}^N e^{\frac{\mathcal{L}(\theta,z+z'_k)}{\tau}}.$$

Denote

$$G_\tau^\theta \overset{\text{def}}{=} \frac{1}{\mu_{\mathcal{L}}(\mathcal{C})} \int_{\mathcal{C}} \nabla_\theta \mathcal{L}(\theta, z+z') e^{\frac{\mathcal{L}(\theta,z+z')}{\tau}} \mathrm{d}z' \text{ and } G_{\tau,N}^\theta \overset{\text{def}}{=} \frac{1}{N} \sum_{k=1}^N \nabla_\theta \mathcal{L}(\theta, z+z'_k) e^{\frac{\mathcal{L}(\theta,z+z'_k)}{\tau}}.$$

We have made here the dependence on θ, τ and N explicit as it will make our reasoning clearer.

In the first part of the proof, θ is a fixed vector in Ξ. We have

$$\nabla g_\tau(\theta) - \nabla g_{\tau,N}(\theta) = \int_{\mathcal{C}} \nabla_\theta \mathcal{L}(\theta, z+z') \mathrm{d}\mu_\tau^\theta(z') - \int_{\mathcal{C}} \nabla_\theta \mathcal{L}(\theta, z+z') \mathrm{d}\mu_{\tau,N}^\theta(z')$$

and thus, by assumption (H.5),

$$\begin{aligned}
&\|\nabla g_{\tau,N}(\theta) - \nabla g_\tau(\theta)\| \\
&\leq \left\| \int_{\mathcal{C}} \nabla_\theta \mathcal{L}(\theta, z+z') \mathrm{d}\mu_\tau^\theta(z') \left(1 - \frac{S_\tau^\theta}{S_{\tau,N}^\theta}\right) \right\| + \left\| \int_{\mathcal{C}} \nabla_\theta \mathcal{L}(\theta, z+z') \left(\mathrm{d}\mu_\tau^\theta(z') \frac{S_\tau^\theta}{S_{\tau,N}^\theta} - \mathrm{d}\mu_{\tau,N}^\theta(z') \right) \right\| \\
&= \left\| \int_{\mathcal{C}} \nabla_\theta \mathcal{L}(\theta, z+z') \mathrm{d}\mu_\tau^\theta(z') \left(1 - \frac{S_\tau^\theta}{S_{\tau,N}^\theta}\right) \right\| + \frac{\left\| G_{\tau,N}^\theta - G_\tau^\theta \right\|}{S_{\tau,N}^\theta} \\
&\leq L_{\Xi,\mathcal{Z}} e^{\frac{-\underline{\mathcal{L}}^\Xi}{\tau}} \left| S_{\tau,N}^\theta - S_\tau^\theta \right| + e^{\frac{-\underline{\mathcal{L}}^\Xi}{\tau}} \left\| G_{\tau,N}^\theta - G_\tau^\theta \right\|. \qquad (44)
\end{aligned}$$

(i) Since $\mathbb{E}[S_{\tau,N}^\theta] = S_\tau^\theta$, the bound in expectation of the first term in (44) follows from (37). For the second term, we argue similarly. Indeed, let $Y_k = \nabla \mathcal{L}(\theta, z+z'_k) e^{\frac{\mathcal{L}(\theta,z+z'_k)}{\tau}}$. Since $\mathbb{E}[G_{\tau,N}^\theta] = G_\tau^\theta$ and the random vectors Y_k are i.i.d. , we have

$$\mathbb{E}\left[\left\| G_{\tau,N}^\theta - G_\tau^\theta \right\| \right] \leq \frac{1}{\sqrt{N}} \mathbb{E}[\|Y\|^2]^{1/2} \leq \frac{L_{\Xi,\mathcal{Z}} e^{\frac{\overline{\mathcal{L}}^\Xi}{\tau}}}{\sqrt{N}}.$$

Combining these two bounds and (44) yields (42).

(ii) Arguing as in (38), we get that for any $\kappa > 0$,

$$\mathbb{P}\left(\left| S_{\tau,N}^\theta - S_\tau^\theta \right| \geq e^{\frac{\overline{\mathcal{L}}^\Xi}{\tau}} \sqrt{\frac{\kappa \log N}{2N}} \right) \leq 2N^{-\kappa}.$$

Since $\mathbb{E}[G^{\theta}_{\tau,N}] = G^{\theta}_{\tau}$ and the random vectors $Y_k = \nabla\mathcal{L}(\theta, z + z'_k)e^{\frac{\mathcal{L}(\theta,z+z'_k)}{\tau}}$ are i.i.d. and bounded, we apply the vector Bernstein inequality [58, Lemma 18] (who refined slightly [59, Theorem 12]), to get that

$$\mathbb{P}\left(\left\|G^{\theta}_{\tau,N} - G^{\theta}_{\tau}\right\| \geq \epsilon\right) \leq e^{-\frac{n\epsilon^2}{8\varsigma^2}+\frac{1}{4}} \leq 2e^{-\frac{n\epsilon^2}{8\varsigma^2}}.$$

for any $0 < \epsilon < \varsigma^2/\nu$, where $\max_k \|Y_k - \mathbb{E}[Y_k]\| \leq \nu$ and $\max_k \mathbb{E}[\|Y_k - \mathbb{E}[Y_k]\|^2] \leq \varsigma^2$. But from above, we can take

$$\nu = 2L_{\Xi,\mathcal{Z}}e^{\frac{\mathcal{L}^{\Xi}}{\tau}} \quad \text{and} \quad \varsigma^2 = L^2_{\Xi,\mathcal{Z}}e^{2\frac{\mathcal{L}^{\Xi}}{\tau}}.$$

This entails that, choosing $\epsilon = 2L_{\Xi,\mathcal{Z}}e^{\frac{\mathcal{L}^{\Xi}}{\tau}}\sqrt{\frac{2\kappa\log N}{N}}$, we have

$$\mathbb{P}\left(\left\|G^{\theta}_{\tau,N} - G^{\theta}_{\tau}\right\| \geq 2L_{\Xi,\mathcal{Z}}e^{\frac{\mathcal{L}^{\Xi}}{\tau}}\sqrt{\frac{2\kappa\log N}{N}}\right) \leq 2N^{-\kappa},$$

for any $\kappa > 0$ and N large enough such that $\frac{N}{\log N} \geq 32\kappa$. Plugging the above bounds into (44), we obtain, for any fixed $\theta \in \Xi$, that

$$\mathbb{P}\left(\|\nabla g_{\tau}(\theta) - \nabla g_{\tau,N}(\theta)\| \geq 4L_{\Xi,\mathcal{Z}}e^{\frac{\Delta\mathcal{L}^{\Xi}}{\tau}}\sqrt{\frac{2\kappa\log N}{N}}\right) \leq 2N^{-\kappa} + 2N^{-\kappa} = 4N^{-\kappa}.$$

To make the above bound hold simultaneously for all $\theta \in \Xi$, rest of the proof uses a covering argument in the same vein as in the proof of Lemma A.1. Let Ξ_δ be a δ-net of Ξ. For any $\theta \in \Xi$, there exists $\theta_i \in \Xi_\delta$ such that $\|\theta - \theta_i\| \leq \delta$. Thus, we have

$$\|\nabla g_{\tau,N}(\theta) - \nabla g_{\tau}(\theta)\| \leq \|\nabla g_{\tau,N}(\theta_i) - \nabla g_{\tau}(\theta_i)\| + \|\nabla g_{\tau,N}(\theta_i) - \nabla g_{\tau,N}(\theta)\| + \|\nabla g_{\tau}(\theta_i) - \nabla g_{\tau}(\theta)\|. \tag{45}$$

For the first term, we use a union bound and take $\kappa = p + t$ for any $t > 0$ and N such that $\frac{N}{\log N} \geq 32(p+t)$, to obtain

$$\mathbb{P}\left(\sup_{i\in[N(\Xi,\delta)]} \|\nabla g_{\tau,N}(\theta_i) - \nabla g_{\tau}(\theta_i)\| \geq 4L_{\Xi,\mathcal{Z}}e^{\frac{\Delta\mathcal{L}^{\Xi}}{\tau}}\sqrt{\frac{2(p+t)\log N}{N}}\right) \leq 4N(\Xi,\delta)N^{-(p+t)}. \tag{46}$$

Let us turn to the two remaining terms in and bound the first (the second follows similarly). We have, arguing as in (44),

$$\|\nabla g_{\tau,N}(\theta) - \nabla g_{\tau,N}(\theta_i)\| \leq L_{\Xi,\mathcal{Z}}e^{\frac{-\mathcal{L}^{\Xi}}{\tau}}\left|S^{\theta}_{\tau,N} - S^{\theta_i}_{\tau}\right| + e^{\frac{-\mathcal{L}^{\Xi}}{\tau}}\left\|G^{\theta}_{\tau,N} - G^{\theta_i}_{\tau}\right\|$$

We have already shown in the proof of Lemma A.1, using (H.4), that

$$|S^{\theta}_{\tau,N} - S^{\theta_i}_{\tau,N}| \leq \tau^{-1}e^{\frac{\mathcal{L}^{\Xi}}{\tau}}l_{\Xi}\delta.$$

On the other hand,

$$\begin{aligned}\left\|G^{\theta}_{\tau,N} - G^{\theta_i}_{\tau,N}\right\| &\leq \frac{1}{N}\sum_{k=1}^{N} e^{\frac{\mathcal{L}(\theta_i, z+z'_k)}{\tau}} \left\|\nabla_\theta \mathcal{L}(\theta, z+z'_k) - \nabla_\theta \mathcal{L}(\theta_i, z+z'_k)\right\| \\ &\quad + \frac{1}{N}\sum_{k=1}^{N} \left\|\nabla_\theta \mathcal{L}(\theta, z+z'_k)\right\| \left| e^{\frac{\mathcal{L}(\theta, z+z'_k)}{\tau}} - e^{\frac{\mathcal{L}(\theta_i, z+z'_k)}{\tau}} \right| \\ &\leq e^{\frac{\overline{\mathcal{L}}^{\Xi}}{\tau}} L_{\Xi}\delta + \tau^{-1} e^{\frac{\overline{\mathcal{L}}^{\Xi}}{\tau}} L_{\Xi,\mathcal{Z}}\delta.\end{aligned}$$

Hence

$$\begin{aligned}\|\nabla g_{\tau,N}(\theta) - \nabla g_{\tau,N}(\theta_i)\| &\leq \tau^{-1} e^{\frac{\Delta\mathcal{L}^{\Xi}}{\tau}} L_{\Xi,\mathcal{Z}} l_{\Xi}\delta + e^{\frac{\Delta\mathcal{L}^{\Xi}}{\tau}} L_{\Xi}\delta + \tau^{-1} e^{\frac{\Delta\mathcal{L}^{\Xi}}{\tau}} L_{\Xi,\mathcal{Z}}\delta \\ &= e^{\frac{\Delta\mathcal{L}^{\Xi}}{\tau}} \left(\tau^{-1} L_{\Xi,\mathcal{Z}}(l_{\Xi}+1) + L_{\Xi}\right)\delta.\end{aligned}$$

Setting $\delta = 2D_{\Xi}/N$, we get $N(\Xi,\delta) \leq N^p$. Plugging this into (46) and combining with (45) proves the claim.

(iii) (a) The convergence in expectation is immediate.

(b) For the almost sure convergence, we argue as in the proof of Lemma A.1, setting $t = 2$ and invoking again the (first) Borel-Cantelli lemma.

□

A.7 Proof of Theorem 5.2

We first show that G is definable on an o-minimal structure. Indeed, o-minimal structures enjoy powerful stability results under many operations: for instance sublevel sets of definable functions are definable, finite sums of definable functions are definable, and functions of the type $\sup_{v\in\mathcal{S}} F(u,v)$ (resp. $\inf_{v\in\mathcal{S}} F(u,v)$) where F and $\mathcal{S}$ are definable, are definable. Definability of $\mathcal{C}_\varepsilon$ together with that of $\mathcal{L}$ imply that g_i is definable for each i. In turn, we get definability of G as a finite sum of definable functions.

Consider an absolutely continuous curve $\theta : \mathbb{R}_+ \to \mathbb{R}^p$. The function G being locally Lipschitz continuous, $t \mapsto G(\theta(t))$ is also absolutely continuous and thus

$$\frac{\mathrm{d}}{\mathrm{d}t}G(\theta(t)) = \frac{1}{M}\sum_{i=1}^{M}\frac{\mathrm{d}}{\mathrm{d}t}g_i(\theta(t)) = \langle \frac{1}{M}\sum_{i=1}^{M} u_i, \dot{\theta}(t)\rangle, \quad \text{for all } u_i \in \partial^C g_i(\theta(t)) \text{ and for a.e. } t \geq 0,$$

where we used that the functions g_i are path differentiable for the Clarke subdifferential by [60, Theorem 5.8]. Therefore, G is a Lyapunov function for the set crit–G. Moreover, by [39, Theorem 6], G(crit–G) has empty interior.

By assumption on $(\theta_k)_{k\in\mathbb{N}}$, there exists a $[0,+\infty[$-valued random variable C such that $\sup_{k\in\mathbb{N}}\|\theta_k\| \leq C$ almost surely. In turn, as the Clarke subdifferential is compact valued, there exists another $[0,+\infty[$-valued random variable $M(C)$, which only depends only on C, such that almost surely

$$\sup\left\{\|v\| : v \in \bigcup_{i\in[M],k\geq 1} \partial^C g_i(\theta_k)\right\} \leq M(C).$$

Moreover, the direction d_k is such that

$$d_k = v_k + \zeta_k,$$

with $v_k \in \frac{1}{M}\sum_{i=1}^{M} \partial^C g_i(\theta_k)$ and the random process ζ_k is a zero-mean bounded martingale difference noise. The boundedness properties and the choice of the sequence γ_k allows to apply [41, Remark 1.5(ii) and Proposition 1.4] to get by [41, Proposition 1.3] that the continuous-time affine interpolant of $(\theta_k)_{k\in\mathbb{N}}$ is almost surely an asymptotic pseudotrajectory of the flow (26). Combining this with [41, Theorem 3.6 and Proposition 3.27] gives the claimed results. □

A.8 Proof of Theorem 5.3

(i) Our definability assumption ensures that $G_{\tau,N}$ is definable. The first convergence statement then follows by arguing as in the proof of Theorem 5.2 (see also [61, Theorem 6.1.1] for a part of the claims since $G_{\tau,N}$ is smooth).

(ii) For any compact subset $\Xi \subset$ crit–$G_{\tau,N}$ and $\bar{\theta}_{\tau,N} \in \Xi$ we have

$$\begin{aligned}\mathrm{dist}\left(0, \frac{1}{M}\sum_{i=1}^{M} \partial^C g_i(\bar{\theta}_{\tau,N})\right) &= \mathrm{dist}\left(\nabla G_{\tau,N}(\bar{\theta}_{\tau,N}), \frac{1}{M}\sum_{i=1}^{M} \partial^C g_i(\bar{\theta}_{\tau,N})\right) \\ &\leq \frac{1}{M}\sum_{i=1}^{M} \mathrm{dist}\left(\nabla g^i_{\tau,N}(\bar{\theta}_{\tau,N}), \partial^C g_i(\bar{\theta}_{\tau,N})\right).\end{aligned}$$

Taking the supremum on both sides over Ξ, the bound (32) then follows from Theorem 4.8(ii).

Let us now turn to (33). Recall that the g_i's are locally Lipschitz continuous and definable under our assumption (see the proof of Theorem 5.2). It follows that $\frac{1}{M}\sum_{i=1}^{M} \partial^C g_i$ is a definable set-valued mapping with closed graph. We are then in position to apply the metric subregularity result as given in [62, Proposition 3.1][9] applied on the compact set crit_κ–G to see that there exists a nonnegative increasing definable function φ, continuous at 0 and vanishing only there, such that for all $\theta \in \mathrm{crit}_\kappa$–$G$

$$\mathrm{dist}(\theta, \text{crit–}G) \leq \varphi\left(\mathrm{dist}\left(0, \frac{1}{M}\sum_{i=1}^{M} \partial^C g_i(\theta)\right)\right).$$

Since $\varsigma(N, \tau, t, \Xi) \leq \kappa$, we have from (32) that $\Xi \subset \mathrm{crit}_\kappa$–$G$ with the same probability. Using again (32) and that φ is increasing, we conclude.

A.9 Proof of Theorem 5.5

Let us fix some notations. Equip $\mathbb{R}^p$ with the Borel σ-algebra. Let $\omega_k \overset{\mathrm{def}}{=} (\omega_k^B, \omega_k^S)$ be the event of sampling the mini-batch and integration points at iteration $k \in \mathbb{N}$ (the events ω_k^B and ω_k^S are mutually independent, and ω_k are i.i.d. from a probability measure $\mathbb{P}$). Let $\vartheta : \mathbb{R}^d \times \Omega \to \mathbb{R}^d$ be the stochastic subgradient oracle used in Algorithm 3, i.e., $\vartheta(\theta_k, \omega_k) = d_k$. A filtration is a sequence of non-decreasing sub-σ-algebras $(\mathcal{F}_k)_{k\in\mathbb{N}}$, i.e., $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all $k \in \mathbb{N}$. In our context, we take $\mathcal{F}_k = \sigma(\theta_i, \omega_{i-1} : i \leq k)$, i.e. the canonical filtration or history of Algorithm 3.

[9] Actually, [62, Proposition 3.1] is stated there for semialgebraic set-valued mappings, but can be extended to operators definable on an o-minimal structure (see also [63, Remark 3.2]).

The proof consists in invoking [60, Theorem 3.2]. To this end, we need to verify Assumption A and B therein. Our argument is in the same vein as in the proof of [60, Theorem 6.2] but we need to handle carefully the presence of the error term and the sampling of the integration points.

Observe that [60, Assumption B] is verified thanks to the definability assumption on $\|\cdot\|$, $\mathcal{L}$ and Ξ. We also have by construction that $(\theta_k)_{k\in\mathbb{N}} \subset \Xi$ which is a closed set and thus [60, Assumption A.1] is in force. Moreover, (34) is precisely [60, Assumption A.3].

To check the remaining assumptions, i.e., [60, Assumptions A.2, A.4 and A.5], we first rewrite the update of Algorithm 3 in the Robbins-Monro stochastic approximation form (see e.g., [61, Section 6.1.1])

$$\theta_{k+1} = \theta_k + \gamma_k(v_k + \zeta_k), \tag{47}$$

where

$$v_k = -\mathbb{E}\left[\vartheta(\theta_k,\omega) \mid \mathcal{F}_k\right] - \gamma_k^{-1}\mathbb{E}\left[\theta_k - \gamma_k\vartheta(\theta_k,\omega) - \mathrm{P}_\Xi\left(\theta_k - \gamma_k\vartheta(\theta_k,\omega)\right) \mid \mathcal{F}_k\right]$$

and

$$\zeta_k = \gamma_k^{-1}\big(\mathrm{P}_\Xi(\theta_k - \gamma_k\vartheta(\theta_k,\omega_k)) - \theta_k - \mathbb{E}\left[\mathrm{P}_\Xi\left(\theta_k - \gamma_k\vartheta(\theta_k,\omega)\right) - \theta_k \mid \mathcal{F}_k\right]\big).$$

It is understood that the conditional expectation is wrt to the event $\omega \sim \mathbb{P}$. Moreover, we will use the bound

$$\|\theta_{k+1} - \theta_k\| \leq \gamma_k\vartheta(\theta_k,\omega_k). \tag{48}$$

Indeed, since Ξ a closed convex set, θ_{k+1} is the unique solution to

$$\min_{\theta\in\Xi} \|\theta - (\theta_k - \gamma_k\vartheta(\theta_k,\omega_k))\|. \tag{49}$$

Thus, strong convexity of this objective yields

$$\|\theta_{k+1} - (\theta_k - \gamma_k\vartheta(\theta_k,\omega_k))\|^2 \leq \|\theta_k - (\theta_k - \gamma_k\vartheta(\theta_k,\omega_k))\|^2 - \frac{1}{2}\|\theta_{k+1} - \theta_k\|^2.$$

Simplifying yields

$$\|\theta_{k+1} - \theta_k\|^2 \leq \gamma_k\langle\vartheta(\theta_k,\omega_k), \theta_{k+1} - \theta_k\rangle.$$

Applying Cauchy-Schwarz inequality gives (48).

Checking [60, Assumptions A.2] corresponds to verifying that $(\theta_k)_{k\in\mathbb{N}}$ and $(v_k)_{k\in\mathbb{N}}$ are almost surely bounded. This is obviously true since $(\theta_k)_{k\in\mathbb{N}} \subset \Xi$ by construction (θ_k is even uniformly bounded). For v_k, we have using respectively Jensen's inequality, (48) and Assumption (H.5) that

$$\begin{aligned}\|v_k\| &= \left\|\gamma_k^{-1}\mathbb{E}\left[\theta_k - \mathrm{P}_\Xi\left(\theta_k - \gamma_k\vartheta(\theta_k,\omega)\right) \mid \mathcal{F}_k\right]\right\| \\ &\leq \mathbb{E}\left[\left\|\gamma_k^{-1}\left(\theta_k - \mathrm{P}_\Xi\left(\theta_k - \gamma_k\vartheta(\theta_k,\omega)\right)\right)\right\| \mid \mathcal{F}_k\right] \\ &\leq \mathbb{E}\left[\|\vartheta(\theta_k,\omega)\| \mid \mathcal{F}_k\right] \leq \max_{\theta\in\Xi, i\in[M], u\in\mathcal{C}_\varepsilon} \|\nabla_\theta\mathcal{L}(\theta, z_i + u)\| \leq L_{\Xi,\mathcal{Z}} < +\infty.\end{aligned} \tag{50}$$

Let us now turn to verifying [60, Assumptions A.4], i.e., the series $\sum_{i=1}^k \gamma_i\zeta_i$ converges almost surely. First observe that $(\gamma_k\zeta_k)_{k\in\mathbb{N}}$ is a martingale difference noise sequence. Indeed, $\mathbb{E}\left[\gamma_k\zeta_k \mid \mathcal{F}_k\right] = 0$ and using (48) and (50) we have

$$\begin{aligned}\mathbb{E}\left[\|\gamma_k\zeta_k\|^2 \mid \mathcal{F}_k\right] &= \mathrm{Var}\left[\mathrm{P}_\Xi(\theta_k - \gamma_k\vartheta(\theta_k,\omega_k)) - \theta_k \mid \mathcal{F}_k\right] \\ &\leq \mathbb{E}\left[\|\mathrm{P}_\Xi(\theta_k - \gamma_k\vartheta(\theta_k,\omega_k)) - \theta_k\|^2 \mid \mathcal{F}_k\right] \leq \gamma_k^2 L_{\Xi,\mathcal{Z}}^2,\end{aligned}$$

whence we get, thanks to (34) that

$$\sum_{k\in\mathbb{N}} \gamma_k^2 \mathbb{E}\left[\|\zeta_k\|^2 \mid \mathcal{F}_k\right] \leq L_{\Xi,\mathcal{Z}}^2 \sum_{k\in\mathbb{N}} \gamma_k^2 < +\infty \quad \text{almost surely.} \tag{51}$$

Thus the random sequence $(\sum_{i=1}^k \gamma_i \zeta_i)_{k\in\mathbb{N}}$ is a square-integrable martingale which satisfies (51) and it follows from [64, Theorem 5.4.9] that $\sum_{i=1}^k \gamma_i \zeta_i$ converges almost surely to a finite limit.

It remains to check [60, Assumptions A.5], i.e., for any converging subsequence θ_k (that we do not relabel for simplicity) in Ξ whose cluster point is some point $\bar{\theta} \in \Xi$, we have

$$\lim_{n\to+\infty} \operatorname{dist}\left(-\frac{1}{n}\sum_{k=1}^n v_k, A(\bar{\theta})\right) = 0 \quad \text{almost surely,}$$

where A is the set-valued operator on $\mathbb{R}^p$

$$A(\theta) \overset{\text{def}}{=} \frac{1}{M}\sum_{i=1}^M \partial^C g_i(\theta) + N_\Xi(\theta).$$

For this, we start by noticing that since A is convex-valued (both the Clarke subdifferential and the normal cone N_Ξ are), we use the Jensen's inequality to split this in two bounds according to

$$\begin{aligned}\operatorname{dist}\left(-\frac{1}{n}\sum_{k=1}^n v_k, A(\bar{\theta})\right) \leq \frac{1}{n}\sum_{k=1}^n \Bigg[&\operatorname{dist}\left(\mathbb{E}\left[\vartheta(\theta_k,\omega)\right) \mid \mathcal{F}_k\right], \frac{1}{M}\sum_{l=1}^M \partial^C g_l(\bar{\theta})\right) \\ &+ \operatorname{dist}\left(\mathbb{E}\left[\gamma_k^{-1}\left(\theta_k - \gamma_k\vartheta(\theta_k,\omega) - \mathrm{P}_\Xi\left(\theta_k - \gamma_k\vartheta(\theta_k,\omega)\right)\right) \mid \mathcal{F}_k\right], N_\Xi(\bar{\theta})\right)\Bigg]\end{aligned} \tag{52}$$

For the first term, using Jensen's inequality recalling that the Clarke subdifferential is convex-valued, we have

$$\begin{aligned}\operatorname{dist}\left(\mathbb{E}\left[\vartheta(\theta_k,\omega)\right) \mid \mathcal{F}_k\right], \frac{1}{M}\sum_{l=1}^M \partial^C g_l(\theta_k)\right) &= \operatorname{dist}\left(\mathbb{E}_{\omega^S}\left[\mathbb{E}_{\omega^B}\left[\vartheta(\theta_k,\omega^B,\omega^S) \mid \mathcal{F}_k\right] \mid \mathcal{F}_k\right], \frac{1}{M}\sum_{l=1}^M \partial^C g_l(\theta_k)\right) \\ &= \operatorname{dist}\left(\frac{1}{M}\sum_{i=1}^M \mathbb{E}_{\omega^S}\left[\nabla g^i_{\tau_k,N_k}(\theta_k,\omega^S) \mid \mathcal{F}_k\right], \frac{1}{M}\sum_{l=1}^M \partial^C g_l(\theta_k)\right) \\ &\leq \frac{1}{M}\mathbb{E}_{\omega^S}\left[\operatorname{dist}\left(\sum_{i=1}^M \nabla g^i_{\tau_k,N_k}(\theta_k,\omega^S), \sum_{l=1}^M \partial^C g_l(\theta_k)\right) \mid \mathcal{F}_k\right] \\ &\leq \frac{1}{M}\sum_{i=1}^M \mathbb{E}_{\omega^S}\left[\operatorname{dist}\left(\nabla g^i_{\tau_k,N_k}(\theta_k,\omega^S), \partial^C g_i(\theta_k)\right) \mid \mathcal{F}_k\right] \\ &\leq \frac{1}{M}\sum_{i=1}^M \mathbb{E}_{\omega^S}\left[\operatorname{dist}\left(\nabla g^i_{\tau_k,N_k}(\theta_k,\omega^S), \partial^C g_i(\theta_k)\right) \mid \mathcal{F}_k\right] \\ &\leq \max_{i\in[M]}\sup_{\theta\in\Xi} \mathbb{E}_{\omega^S}\left[\operatorname{dist}\left(\nabla g^i_{\tau_k,N_k}(\theta,\omega^S), \partial^C g_i(\theta)\right)\right].\end{aligned} \tag{53}$$

Denote κ_k the right hand side of the last inequality. (53) tells us that

$$\mathbb{E}\left[\vartheta(\theta_k,\omega)\right) \mid \mathcal{F}_k\right] + \epsilon_k e_k \in \frac{1}{M}\sum_{l=1}^M \partial^C g_l(\theta_k), \tag{54}$$

for some vector e_k on the unit sphere and $0 < \epsilon_k \leq \kappa_k$. Now $\kappa_k \to 0$, hence $\epsilon_k \to 0$, thanks to Theorem 4.8(iii)(a). Therefore, since the Clarke subdifferential is outer semicontinuous, hence the graph of $\partial^C g_l$ is sequentially closed for any $l \in [M]$, $\mathbb{E}\left[\vartheta(\theta_k, \omega) \mid \mathcal{F}_k\right]$ is (uniformly) bounded by (50) and $\theta_k \to \bar{\theta}$, we infer from (54) that every cluster point of $\mathbb{E}\left[\vartheta(\theta_k, \omega) \mid \mathcal{F}_k\right]$ belongs to $\frac{1}{M}\sum_{l=1}^{M} \partial^C g_l(\bar{\theta})$, i.e.,

$$\operatorname{dist}\left(\mathbb{E}\left[\vartheta(\theta_k, \omega)) \mid \mathcal{F}_k\right], \frac{1}{M}\sum_{l=1}^{M} \partial^C g_l(\bar{\theta})\right) \to 0. \tag{55}$$

Let us now turn to the second term of (52). The first order optimality condition of (49) reads

$$\gamma_k^{-1}\Big(\theta_k - \mathrm{P}_\Xi\left(\theta_k - \gamma_k \vartheta(\theta_k, \omega)\right)\Big) - \vartheta(\theta_k, \omega) \in N_\Xi\left(\mathrm{P}_\Xi\left(\theta_k - \gamma_k \vartheta(\theta_k, \omega)\right)\right). \tag{56}$$

$\vartheta(\theta_k, \omega)$ is bounded almost surely in ω and $\gamma_k \to 0$, and thus $\gamma_k \vartheta(\theta_k, \omega) \to 0$ almost surely in ω. As $\theta_k \to \bar{\theta} \in \Xi$ and P_Ξ is continuous (see [33, Theorem 2.26]), we have almost surely in ω

$$\mathrm{P}_\Xi\left(\theta_k - \gamma_k \vartheta(\theta_k, \omega)\right) \to \bar{\theta}.$$

Furthermore, similarly to (50),

$$\left\|\gamma_k^{-1}\Big(\theta_k - \mathrm{P}_\Xi\left(\theta_k - \gamma_k \vartheta(\theta_k, \omega)\right)\Big) - \vartheta(\theta_k, \omega)\right\| \leq 2L_{\Xi,\mathcal{Z}} < +\infty.$$

Thus using sequential closedness of N_Ξ in (56) implies that almost surely, each cluster of $\gamma_k^{-1}\Big(\theta_k - \mathrm{P}_\Xi\left(\theta_k - \gamma_k \vartheta(\theta_k, \omega)\right)\Big) - \vartheta(\theta_k, \omega)$ belongs to $N_\Xi(\bar{\theta})$, i.e.,

$$\operatorname{dist}\left(\gamma_k^{-1}\Big(\theta_k - \mathrm{P}_\Xi\left(\theta_k - \gamma_k \vartheta(\theta_k, \omega)\right)\Big) - \vartheta(\theta_k, \omega), N_\Xi(\bar{\theta})\right) \to 0 \quad \text{almost surely in } \omega.$$

Now since $0 \in N_\Xi(\bar{\theta})$, we have

$$\begin{aligned}\operatorname{dist}\left(\gamma_k^{-1}\Big(\theta_k - \mathrm{P}_\Xi\left(\theta_k - \gamma_k \vartheta(\theta_k, \omega)\right)\Big) - \vartheta(\theta_k, \omega), N_\Xi(\bar{\theta})\right) &\leq \left\|\gamma_k^{-1}\Big(\theta_k - \mathrm{P}_\Xi\left(\theta_k - \gamma_k \vartheta(\theta_k, \omega)\right)\Big) - \vartheta(\theta_k, \omega)\right\| \\ &\leq 2L_{\Xi,\mathcal{Z}}.\end{aligned}$$

We are then in position to apply the dominated convergence theorem to get

$$\begin{aligned}&\operatorname{dist}\left(\mathbb{E}\left[\gamma_k^{-1}\left(\theta_k - \gamma_k \vartheta(\theta_k, \omega) - \mathrm{P}_\Xi\left(\theta_k - \gamma_k \vartheta(\theta_k, \omega)\right)\right) \mid \mathcal{F}_k\right], N_\Xi(\bar{\theta})\right) \\ &\qquad \leq \mathbb{E}\left[\operatorname{dist}\left(\gamma_k^{-1}\left(\theta_k - \gamma_k \vartheta(\theta_k, \omega) - \mathrm{P}_\Xi\left(\theta_k - \gamma_k \vartheta(\theta_k, \omega)\right)\right), N_\Xi(\bar{\theta})\right) \mid \mathcal{F}_k\right] \to 0.\end{aligned} \tag{57}$$

Plugging (55) and (57) into (52) shows that [60, Assumptions A.5] is indeed verified. This completes the proof. □

A.10 Global convergence and convergence rate of the Gibbs measure

We will need some regularity properties on the set of maximizers $\mathcal{M}$. We say that a set $\mathcal{S} \subset \mathbb{R}^m$ is $\mathscr{C}^r$-stratifiable, for some integer $r \geq 1$, if there is a finite partition of $\mathcal{S}$ into disjoint $\mathscr{C}^r$ submanifolds $(\mathcal{M}_i)_{i \in I}$ of $\mathbb{R}^m$, called strata, with $m \geq \dim(\mathcal{M}_1) > \dim(\mathcal{M}_2) > \cdots > \dim(\mathcal{M}_{|I|}) \geq 0$.

Lemma A.4. *Suppose that (H.1), (H.2) and (H.5) hold. Assume also that:*

($\mathrm{H}_{\mathcal{M}}$.1) *$\mathcal{L}(\theta, .)$ is $\mathscr{C}^3$ on an open set containing $\mathcal{C}_\varepsilon$ with Hölder continuous third-order derivative;*

($\mathrm{H}_{\mathcal{M}}$.2) *for $r \geq 3$*

(a) $\mathcal{M}$ is $\mathscr{C}^r$-stratifiable with closed strata;

(b) for each $i \in I$, the Hessian $\nabla^2_{z'}\mathcal{L}(\theta, z+z')$ is negative semidefinite for any $z' \in \mathcal{M}_i$ with constant rank $m - \dim(\mathcal{M}_i)$.

Then

$$\nabla g_\tau(\theta) \underset{\tau \to 0^+}{\longrightarrow} \eta(\theta) \overset{\text{def}}{=} \int_{\mathcal{M}_1} \nabla_\theta \mathcal{L}(\theta, z+z')\mathrm{d}\mu(z') \subset \partial^C g(\theta),$$

where $\mu \in \mathcal{P}(\mathcal{M}_1)$.

Remark A.5. *The assumption that the partition of $\mathcal{M}$ is disjoint can be removed by assuming in addition that each intersecting pair of submanifolds $(\mathcal{M}_i, \mathcal{M}_j)$, $i \neq j$, do so transversely [65, Theorem 6.30]. Therefore, $\mathcal{M}_i \cap \mathcal{M}_j$ is also a submanifold whose dimension strictly smaller than that of $\mathcal{M}_i$ and $\mathcal{M}_j$. The main change in our proof will lie in subtracting the contribution of these intersections in (60), and then use that their dimensions are strictly smaller than that of the largest submanifold.*

Proof. For any Borel set $\mathcal{C} \subset \mathbb{R}^m$ and $k \in \mathbb{N}$, $\mathcal{H}^k(\mathcal{C})$ is the k-dimensional Hausdorff measure. It is normalized to coincide with the Lebesgue measure on $\mathbb{R}^k$. For a k-dimensional smooth submanifold of $\mathbb{R}^m$, its k-dimensional Hausdorff measure coincides with the Riemannian volume measure.

By ($\mathrm{H}_{\mathcal{M}}$.2)(a), for any $r \geq 3$, $\mathcal{M}$ is $\mathscr{C}^r$-stratifiable and thus the strata $(\mathcal{M}_i)_{i\in I}$ are $\mathscr{C}^r$-smooth compact submanifolds.

Given $\epsilon > 0$, for each $\mathcal{M}_i$, we define its open neighborhood

$$\mathcal{U}_i = \{u \in \mathbb{R}^m : \ \mathrm{dist}(u, \mathcal{M}_i) < \epsilon\}.$$

Let $\mathcal{U} \overset{\text{def}}{=} \bigcup_{i\in I} \mathcal{U}_i$. We then have for $f \in \mathscr{C}$

$$\int_{\mathcal{C}} f(z') e^{\frac{\mathcal{L}(\theta,z+z')}{\tau}} \mathrm{d}z' = \int_{\mathcal{C}\cap\mathcal{U}} f(z') e^{\frac{\mathcal{L}(\theta,z+z')}{\tau}} \mathrm{d}z' + \int_{\mathcal{C}\setminus(\mathcal{C}\cap\mathcal{U})} f(z) e^{\frac{\mathcal{L}(\theta,z+z')}{\tau}} \mathrm{d}z'. \tag{58}$$

Since $f(\mathcal{C})$ is compact by compactness of $\mathcal{C}$ and continuity of f, and $\exists \kappa > 0$ such that $\forall z' \in \mathcal{C} \setminus (\mathcal{C} \cap \mathcal{U})$, $\mathcal{L}(\theta, z+z') \leq -\kappa < \max \mathcal{L}(\theta, \mathcal{C}) = 0$, the second integral in (58) verifies, for any $s \geq 0$,

$$\tau^{-s} \left| \int_{\mathcal{C}\setminus(\mathcal{C}\cap\mathcal{U})} f(z) e^{\frac{\mathcal{L}(\theta,z+z')}{\tau}} \mathrm{d}z' \right| \leq (\mu_{\mathcal{L}}(\mathcal{C}) \sup |f(\mathcal{C})|)\, \tau^{-s} e^{-\kappa/\tau} \to 0 \quad \text{uniformly as } \tau \to 0^+. \tag{59}$$

Let us now turn to the first integral. We have, for ϵ sufficiently small

$$\int_{\mathcal{C}\cap\mathcal{U}} f(z') e^{\frac{\mathcal{L}(\theta,z+z')}{\tau}} \mathrm{d}z' = \sum_{i\in I} \int_{\mathcal{C}\cap\mathcal{U}_i} f(z) e^{\frac{\mathcal{L}(\theta,z+z')}{\tau}} \mathrm{d}z'. \tag{60}$$

Since, for any $i \in I$, $\mathcal{M}_i$ is a compact $\mathcal{C}^r$-smooth submanifold with $r \geq 2$, it is a set with positive reach thanks to [66, Theorem 4.12] (see [66, Definition 4.1] for definition of sets of positive reach). Thus, it follows from [66, Theorem 4.8] that $\mathrm{P}_{\mathcal{M}_i}$ is single-valued and Lipschitz continuous on $\mathcal{U}_i$, hence $\mathcal{C} \cap \mathcal{U}_i$, for some

$\epsilon > 0$ small enough. This together with rectifiability and measurability of the sets $\mathcal{C} \cap \mathcal{U}_i$ and $\mathcal{M}_i$ allows to apply the coarea change of variable formula [67, Theorem 3.2.22(3)] to get

$$\int_{\mathcal{C}\cap\mathcal{U}_i} f(z') e^{\frac{\mathcal{L}(\theta,z+z')}{\tau}} \mathrm{d}z' = \int_{\mathcal{M}_i} \left(\int_{\mathrm{P}^{-1}_{\mathcal{M}_i}(v)} f(u) e^{\frac{\mathcal{L}(\theta,u)}{\tau}} (\mathbf{J}_{m_i}(\mathrm{P}_{\mathcal{M}_i})(u))^{-1} d\mathcal{H}^{m-m_i}(u) \right) d\mathcal{H}^{m_i}(v),$$

where $m_i = \dim(\mathcal{M}_i)$, $\mathbf{J}_{m_i}(\mathrm{P}_{\mathcal{M}_i})$ is the m_i-dimensional Jacobian of $\mathrm{P}_{\mathcal{M}_i}$, i.e.,

$$\mathbf{J}_{m_i}(\mathrm{P}_{\mathcal{M}_i})(u) = \sqrt{\det\left(\mathbf{D}(\mathrm{P}_{\mathcal{M}_i})(u)\, \mathbf{D}(\mathrm{P}_{\mathcal{M}_i})(u)^{\top}\right)}$$

and $\mathbf{D}$ is the derivative operator. In addition, since $\mathcal{M}_i$ is $\mathscr{C}^r$-smooth, we have from [68, Proposition 5.1] that $\mathrm{P}_{\mathcal{M}_i}$ is $\mathscr{C}^{r-1}$-smooth with Lipschitz derivative on $\mathcal{U}_i$ (taking ϵ smaller if necessary). This entails that the key estimates of [56, Lemma 6.1] hold in our case. The rest of our argument follows then similar lines to those of [56, Theorem 3.1, starting from (9.3)]. This allows us to show that

$$\lim_{\tau\to 0^+} \tau^{-\frac{m-m_i}{2}} \int_{\mathcal{C}\cap\mathcal{U}_i} f(z) e^{\frac{\mathcal{L}(\theta,z)}{\tau}} \mathrm{d}z' \\ = 2^{\frac{m-m_i}{2}} (m-m_i) \alpha_{(m-m_i)} \beta_{(m-m_i)} \int_{\mathcal{M}_i} f(v) \left(\prod_{j=1}^{m-m_i} \lambda_j(v)^{-\frac{1}{2}} \right) d\mathcal{H}^{m_i}(v), \tag{61}$$

where $(-\lambda_j(v))_j$ are the $m - m_i$ eigenvalues of the Hessian $\nabla^2_{z'}\mathcal{L}(\theta, v)$ for $v \in \mathcal{M}_i$, which are negative by (H$_{\mathcal{M}}$.2)(b), α_k is the k-dimensional Lebesgue measure of the unit ball in $\mathbb{R}^k$, and

$$\beta_k \overset{\text{def}}{=} \begin{cases} 2^{-\frac{k}{2}}(k-2)(k-4)\cdot(2) & \text{for } k \text{ even} \\ 2^{-\frac{k}{2}}(k-2)(k-4)\cdot(3)\sqrt{\pi} & \text{for } k \text{ odd,} \end{cases}$$

Since the strata are ordered by strictly decreasing dimension, we have from (61) that for any $i > j$,

$$\lim_{\tau\to 0^+} \tau^{-\frac{m-m_j}{2}} \int_{\mathcal{C}\cap\mathcal{U}_i} f(z) e^{\frac{\mathcal{L}(\theta,z)}{\tau}} \mathrm{d}z' = \lim_{\tau\to 0^+} \tau^{\frac{m_j-m_i}{2}} \left(\tau^{-\frac{m-m_i}{2}} \int_{\mathcal{C}\cap\mathcal{U}_i} f(z) e^{\frac{\mathcal{L}(\theta,z)}{\tau}} \mathrm{d}z' \right) = 0. \tag{62}$$

Combining (62) (for $j = 1$) and (61) (for $i = 1$) with (58), (59) and (60), we get

$$\lim_{\tau\to 0^+} \tau^{-\frac{m-m_1}{2}} \int_{\mathcal{C}} f(z') e^{\frac{\mathcal{L}(\theta,z+z')}{\tau}} \mathrm{d}z' = \lim_{\tau\to 0^+} \tau^{-\frac{m-m_1}{2}} \int_{\mathcal{C}\cap\mathcal{U}_1} f(z) e^{\frac{\mathcal{L}(\theta,z)}{\tau}} \mathrm{d}z' \\ = 2^{\frac{m-m_1}{2}} (m-m_1) \alpha_{(m-m_1)} \beta_{(m-m_1)} \int_{\mathcal{M}_1} f(v) \left(\prod_{j=1}^{m-m_1} \lambda_1(v)^{-\frac{1}{2}} \right) d\mathcal{H}^{m_1}(v).$$

Applying this with $f \equiv 1$ and arbitrary $f \in \mathscr{C}$, we get that μ_τ converges in the narrow topology to the probability measure supported on $\mathcal{M}_1 \subset \operatorname{Argmax} \mathcal{L}(\theta, \mathcal{C})$

$$\mathrm{d}\mu(v) = \frac{1}{\int_{\mathcal{M}_1} \left(\prod_{j=1}^{m-m_1} \lambda_1(u)^{-\frac{1}{2}} \right) d\mathcal{H}^{m_1}(u)} \left(\prod_{j=1}^{m-m_1} \lambda_1(v)^{-\frac{1}{2}} \right) d\mathcal{H}^{m_1}(v).$$

By the continuity assumption (H.5) on $\nabla_\theta \mathcal{L}(\theta, z + z')$, we deduce that

$$\lim_{\tau\to 0^+} \nabla g_\tau(\theta) = \lim_{\tau\to 0^+} \int_{\mathcal{C}} \nabla_\theta \mathcal{L}(\theta, z + z') \mathrm{d}\mu_\tau(z') = \int_{\mathcal{M}_1} \nabla_\theta \mathcal{L}(\theta, v) \mathrm{d}\mu(v) \subset \partial^C g(\theta),$$

where we used (5) in the inclusion. This concludes the proof. □

B Additional information on the experiments

Following the recommendations of the paper that introduced the dataset [45], we removed the columns with the modular ratios and the data was centered and normalized. The neural network trained is a MLP with 2 layers of 200 neurons each and an output layer of 12 neurons for the 12 classes with ELU activation function.

Parameter	Value	Description
General parameters for all trainings		
Epochs/Iterations	1500	Number of epochs
Optimizer	SGD-type	SGD for vanilla training, Algorithm 2 for robust training
Initial step-size	0.01	Initial step-size/learning rate
Step-size decay	0.1 every 300 epochs	Multiplicative decay
Batch size	100	Batch size input data
Train set size	10430	
Test set size	10437	
Robustness radius	range between 0. and 0.3	Only relevant for adversarial and robust training
Loss function	Cross Entropy Loss	
Weight initialization	Xavier Glorot's [69]	Default initialization for Pytorch modules
Parameters for adversarial training		
Adversarial Loss	Cross Entropy	
Iteration number	40	Iterations for adversarial attack
Attack norm	ℓ_∞	Norm of the attack, taken accordingly to the Sampling ball
Parameters for robust training		
Monte-Carlo integration	150 000	Number of samples for computing LSE
Sampling ball	$\mathbb{B}_r^\infty$	Ball for uniform MC sampling, taken accordingly to the attack norm
τ	0.0001	Fixed τ for LSE computation

Table 1: Parameters for the trainings on Avila dataset