Regularized Robust Optimization with Application to Robust Learning

William Piat, Jalal Fadili, Frédéric Jurie, Sébastien da Veiga

Regularized Robust Optimization with Application to Robust Learning

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

December 16, 2022

Abstract

In this paper, we propose a computationally tractable and provably convergent algorithm for robust optimization, with application to robust learning. First, the distributionally robust optimization is 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 without any need of convexity/concavity assumptions. 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, Statistical learning, Smoothing, 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 statistical learning. Since there are growing applications in computer vision applied 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 our 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. Normandie Univ, ENSICAEN, CNRS, GREYC, Caen, 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@greyc.ensicaen.fr.

§Safran Tech, Digital Sciences and Technologies Department, Rue des Jeunes Bois, Châteaufort, 78114 Magny-Les-Hameaux, France, sebastien.da-veiga@safrangroup.com.

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 $\mathcal{L}(\cdot, \theta)$ 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]. 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., [3] the recent review paper on min-max problems and their applications from a signal processing and machine learning 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 $\mathsf{d}(z, z') = \|z - z'\|_q$ where $\|\cdot\|_q$ 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 points to a more general perspective in which it is not the points themselves that are perturbed, but rather their underlying distribution ρ_0. This approach is known as Distributionally Robust Optimization (DRO) which 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 $\mathcal{Z}$. Compared to PRO, DRO allows to consider a larger range of perturbations. The choice of D affects the richness of the uncertainty set and the tractability of the resulting optimization problem. Typical choices are the Wasserstein distances [4, 5, 6, 7, 8], the Maximum Mean Discrepancy (MMD) [9] or ϕ-divergences including the Kullback-Leibler divergence [10, 11, 12]. There are also other ways to parametrize the perturbation set in terms of the distribution moments, support, etc., [13, 14]. The Wasserstein distance has become very successful in this context, and unlike other distances/divergences, a Wasserstein distance enjoys the remarkable property that its ball around ρ_0 includes measures having a different support, which allows robustness to unseen data. In the rest of this paper, we focus on Wasserstein balls.

A useful observation at this stage is 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 is very challenging to solve.

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 adversarial training with neural networks [15, 16, 17, 18]. 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 [7]. 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/bias. The error/bias 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 (3) for which tractable algorithms can be designed. For D chosen as a ϕ-divergence, and under some assumptions on $\mathcal{L}$, [1, 11, 19] propose convex optimization approaches. For the Wasserstein distance, and a limited class of convex losses $\mathcal{L}$ and ground costs, some authors convert (3) into a regularized empirical risk minimization problem [20, 4, 5, 6]. For a larger class of losses and ground costs c, (3) is converted in [7] to a Lagrangian form of (2). Stochastic gradient descent is then applied 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, their algorithm resorts to an oracle corresponding to solving the inner supremum problem. This is again a

challenging problem and even NP-hard. When ρ_0 is the empirical measure and $\mathcal{L}$ is Lipschitz continuous in z uniformly in θ the authors in [21, 9] convert (3) into a finite dimensional saddle point problem different from (2) (see detailed discussion in Section 3). Stochastic coordinate descent was advocated in [9] to solve the latter problem 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 this paper, we became aware of the work of [22] 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.

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 Notations and 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 $r \geq 0$ centred at x. The subscript q will be omitted when $q = 2$. For $N \in \mathbb{N}$, $[N]$ is the set of integers $\{1, \dots, N\}$. $\mu_{\mathcal{L}}()$ is the Lebesgue measure/volume of a set. $\operatorname{dist}(x, \mathcal{C}) = \inf_{v \in \mathcal{C}} \|x - z\|$ is the distance function to the nonempty set $\mathcal{C}$. 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.

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 supported on $\mathcal{C}$

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

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

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

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

In the rest of the paper, we assume that $\mathcal{L}$ and the ground cost function c satisfy the standing assumption:

(H.1) $\mathcal{L}$ is continuous.

(H.2) $\mathsf{c} : \mathcal{Z}^2 \to \mathbb{R}_+$ is continuous, symmetric ($\mathsf{c}(z, z') = \mathsf{c}(z', z)$) and $\mathsf{c}(z, z) = 0$.

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 [23] 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 [24]. 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 [25, 26]. 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 [27].

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.*

[1] Functions f such that these assumptions are verified are known as lower-$\mathscr{C}^1$ functions; see [28].

3 Robustness bounds

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) hold.*

(i) If

(H.3) *for every* $\theta \in \Theta$, $\forall (z, z') \in \mathcal{Z}^2$, $|\mathcal{L}(\theta, z) - \mathcal{L}(\theta, z')| \leq L_{\mathcal{Z}}(\theta)\mathsf{c}(z, z')$ *with* $0 \leq L_{\mathcal{Z}} := \sup L_{\mathcal{Z}}(\Theta) < +\infty$.

Then

$$0 \leq \sup_{W_{\mathsf{c}}(\rho,\rho_0)\leq\varepsilon} \mathbb{E}_{\rho}[(\mathcal{L}(\theta, z)] - \mathbb{E}_{z\sim\rho_0}\left[\sup_{\mathsf{c}(z,z')\leq\varepsilon} \mathcal{L}(\theta, z')\right] \leq L_{\mathcal{Z}}\varepsilon.$$

(ii) If $\mathsf{c}(z, z') = \|z - z'\|^q$, $q \geq 1$, *where* $\|\cdot\|$ *is a norm on* $\mathbb{R}^m$ *(i.e.,* $W_{\mathsf{c}}^{1/q}$ *is the q-Wasserstein distance* W_q*), and* $\mathcal{L}(\theta, \cdot)$ *is* $L_{\mathcal{Z}}$*-Lipschitz continuous with respect to* $\|\cdot\|$ *uniformly in* θ*, then*

$$0 \leq \sup_{W_q(\rho,\rho_0)\leq\varepsilon^{1/q}} \mathbb{E}_{\rho}[(\mathcal{L}(\theta, z)] - \mathbb{E}_{z\sim\rho_0}\left[\sup_{\|z-z'\|\leq\varepsilon^{1/q}} \mathcal{L}(\theta, z')\right] \leq C_{q,L_{\mathcal{Z}}}\varepsilon^{1/q},$$

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 [7] (see also [5]), using Lagrangian duality arguments, it was shown that

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

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 [7] to use gradient descent to minimize over θ, but only for a fixed (large enough) parameter γ. But still, choosing γ is not easy.

In [21, 9], taking c as the ℓ_p cost, and ρ_0 the empirical measure on n points, the following bound was established

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

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 minmax problems remains a very challenging task unless stringent joint convexity/concavity assumptions are made. This is the motivation behind our smoothing hereafter.

From now on, we will denote $\mathcal{C}_z^\varepsilon := \{z' : \mathsf{c}(z', z) \leq \varepsilon\}$. The PRO problem now reads

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

We will use the shorthand notation[2]

$$g(\theta) := \max_{z'\in\mathcal{C}_z^\varepsilon} \mathcal{L}(\theta, z').$$

Provided that $\mathcal{C}_z^\varepsilon$ is bounded, hence compact since it is closed by assumption on c, and recalling the continuity assumption (H.1), the set of maximizers in g is a non-empty compact set. The function g can also be equivalently rewritten as

$$g(\theta) = \max_{\mu\in\mathcal{P}(\mathcal{C}_z^\varepsilon)} \int_{\mathcal{C}_z^\varepsilon} \mathcal{L}(\theta, z')\mathrm{d}\mu(z'), \tag{9}$$

and the integral is a duality pairing between $\mathcal{P}(\mathcal{C}_z^\varepsilon)$ and $\mathscr{C}(\mathcal{C}_z^\varepsilon)$. We will show that (9) 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 (9) is concave in μ, but operates in infinite-dimension and is thus hard to solve. Approximating (9) by an atomic measure supported on a finite set (i.e., replace $\mathcal{P}(\mathcal{C}_z^\varepsilon)$ by a finite-dimensional simplex by sampling N points at random in $\mathcal{C}_z^\varepsilon$), as done by some authors (see e.g., [29]), 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 (9), namely

$$g_\tau(\theta) := \max_{\mu\in\mathcal{P}(\mathcal{C}_z^\varepsilon)} \int_{\mathcal{C}_z^\varepsilon} \mathcal{L}(\theta, z')\mathrm{d}\mu(z') - \tau\mathrm{KL}(\mu, \nu), \tag{10}$$

where $\tau > 0$ is the regularization parameter, and ν is a reference measure supported on $\mathcal{C}_z^\varepsilon$. Entropic regularization has been used in several fields including optimal transport [30] and semi-infinite programming [31].

Observe that $\mathcal{C}_z^\varepsilon$ is Lebesgue measurable by continuity of c. In the sequel we also suppose that $\mathcal{C}_z^\varepsilon$ is of full dimension, and set ν as the uniform measure $\mu_{\mathcal{U}}$ on $\mathcal{C}_z^\varepsilon$, that is $\nu(z') = \mu_{\mathcal{U}}(z') := \mu_{\mathcal{L}}(\mathcal{C}_z^\varepsilon)^{-1} < +\infty$ for all $z' \in \mathcal{C}_z^\varepsilon$. 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, (10) is well-posed under mild conditions and has a unique solution taking an explicit form.

[2] g depends on z but we drop this in the notation to lighten the latter.

Proposition 4.1. *Assume that (H.1)-(H.2) hold and that $\mathcal{C}_z^\varepsilon$ is bounded and full dimensional. Then (10) has a unique solution and*

$$g_\tau(\theta) = \tau \log\left(\mathbb{E}_{z'\sim\mu_\mathcal{U}}\left[\exp\left(\frac{\mathcal{L}(\theta,z')}{\tau}\right)\right]\right) = \tau\log\left(\frac{\int_{\mathcal{C}_z^\varepsilon}\exp\left(\frac{\mathcal{L}(\theta,z')}{\tau}\right)\mathrm{d}z'}{\mu_\mathcal{L}(\mathcal{C}_z^\varepsilon)}\right). \tag{11}$$

The proof can be found in Appendix A.2. Note 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 $\mu_\mathcal{U}$.

Remark 4.2. *The boundedness assumption on $\mathcal{C}_z^\varepsilon$ is very mild and verified in most applications we have in mind, for instance in robust training in machine learning. For instance, when $\mathsf{c}(z,z') = \varphi(\|z-z'\|)$ where $\|\cdot\|$ is any norm on $\mathbb{R}^m$, and $\varphi:\mathbb{R}_+\to\mathbb{R}_+$ is a continuous increasing function.*

4.2 Consistency of the regularization

We now turn to describing how g_τ in (10) 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 4.3. *Under the assumptions of Proposition 4.1, the following statements hold:*

(i) $g_\tau(\theta) \nearrow g(\theta)$ *as* $\tau \searrow 0^+$. *In turn,* g_τ Γ*-converges to* g *as* $\tau\searrow 0^+$.

(ii) If, moreover, $\mathcal{C}_z^\varepsilon$ is convex and full dimensional, and (H.3) holds with $\mathsf{c}(z,z') = \|z-z'\|$, where $\|\cdot\|$ is a norm on $\mathbb{R}^m$. Then for any $\tau\in]0,1]$

$$g(\theta) - h(\tau) \le g_\tau(\theta) \le g(\theta), \tag{12}$$

where

$$h(\tau) = m\tau\log(\tau^{-1}) + \tau\log\left(\frac{\mu_\mathcal{L}(\mathcal{C}_z^\varepsilon)}{\mu_\mathcal{L}(\mathbb{B}^{R_{\mathcal{C}_z^\varepsilon}}(0))}\right) + \tau L_\mathcal{Z}(R_{\mathcal{C}_z^\varepsilon} + D_{\mathcal{C}_z^\varepsilon}), \tag{13}$$

and $R_{\mathcal{C}_z^\varepsilon}$ is the radius of the largest ball[3] contained in $\mathcal{C}_z^\varepsilon$ and $D_{\mathcal{C}_z^\varepsilon}$ is the diameter of $\mathcal{C}_z^\varepsilon$.

See Appendix A.3 for the proof.

Remark 4.4.

1. *The convexity of $\mathcal{C}_z^\varepsilon$ is again usual in robust training in machine learning. It holds if the robustness cost c on $\mathcal{Z}$ is convex.*
2. *The bound (12) yields uniform convergence. The Γ-convergence claim in this case also follows from [23, Proposition 5.2 and Remark 5.3].*
3. *The dependence of the convergence rate function $h(\tau)$ on τ is nearly linear (up to a logarithmic term).*

[3] In the norm $\|\cdot\|$ of course.

4. *For the dependence on the dimension of the convergence rate, the first term grows linearly. So does also the second term since it can be upper-bounded by* $m\tau \log\left(\bar{R}_{\mathcal{C}_z^\varepsilon}/R_{\mathcal{C}_z^\varepsilon}\right)$ *where* $\bar{R}_{\mathcal{C}_z^\varepsilon}$ *is the radius of the smallest ball containing* $\mathcal{C}_z^\varepsilon$*. For the last term, the Lipschitz constant* L_z *of the loss function may also depend on the dimension. This emphasizes the role of the Lipschitz constant of the loss in controlling the robustness of the model.*

In view of Theorem 4.3, we obtain the following key result, which relates the minimizers of the smoothed problem to those of the original PRO problem.

Theorem 4.5. *Suppose that the assumptions of Proposition 4.1 hold. Assume also that* Θ *is compact. Let* $\theta_\tau^\star \in \operatorname{Argmin}_\Theta(g_\tau)$*. Then the following holds:*

(i) $\lim_{\tau\to 0^+} \min_{\theta\in\Theta} g_\tau(\theta) = \min_{\theta\in\Theta} g(\theta)$.

(ii) Each cluster point of $\theta_\tau^\star$*, as* $\tau \to 0^+$*, lies in* $\operatorname{Argmin}_\Theta(g)$.

(iii) In particular, if $\operatorname{Argmin}_\Theta(g) = \{\theta^\star\}$*, then* $\lim_{\tau\to 0^+} \theta_\tau^\star = \theta^\star$.

See Appendix A.4 for the proof.

4.3 Monte Carlo approximation of the integral

Computing the values $g_\tau(\theta)$ in (11) 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 $\mathcal{C}_z^\varepsilon$. This gives the approximation

$$g_{\tau,N}(\theta) := \tau \log\left(\sum_{k=1}^N \frac{\exp\left(\frac{\mathcal{L}(\theta,z_k')}{\tau}\right)}{N}\right). \tag{14}$$

We now provide an error bound for such an approximation.

Theorem 4.6. *Suppose that the assumptions of Theorem 4.3(ii) hold. Then, the following holds.*

(i) For any $t > 0$ *and every* $\theta \in \Theta$*,*

$$|g_{\tau,N}(\theta) - g(\theta)| \le h(\tau) + \tau e^{\frac{\overline{\mathcal{L}}-\underline{\mathcal{L}}}{\tau}} \sqrt{\frac{t\log N}{2N}}, \tag{15}$$

with probability at least $1-2N^{-t}$*, where* h *is given in* (13)*,* $\underline{\mathcal{L}} := \inf \mathcal{L}(\Theta, \mathcal{C}_z^\varepsilon)$ *and* $\overline{\mathcal{L}} := \sup \mathcal{L}(\Theta, \mathcal{C}_z^\varepsilon)$.

(ii) Suppose that τ *is a function of* N*, say* τ_N*, with* $h(\tau_N) + \tau_N e^{\frac{\overline{\mathcal{L}}-\underline{\mathcal{L}}}{\tau_N}} \sqrt{\frac{\log N}{N}} \to 0$ *as* $N \to +\infty$*, then*

(a) for every $\theta \in \Theta$

$$g_{\tau_N,N}(\theta) \underset{N\to+\infty}{\longrightarrow} g(\theta) \quad \textit{almost surely.}$$

(b) If moreover,

(H.4) *for every* $z \in \mathcal{Z}$*,* $|\mathcal{L}(\theta, z) - \mathcal{L}(\theta', z)| \le L_\Theta(z)\,\|\theta - \theta'\|$*,* $\forall(\theta,\theta') \in \Theta^2$*, with* $0 \le L_\Theta := \sup L_\Theta(\mathcal{Z}) < +\infty$.

Then, almost surely,

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

The proof can be found in Appendix A.5.

Remark 4.7.

- *Observe that since $\mathcal{L}$ is continuous by (H.1) and $\mathcal{C}_z^\varepsilon$ is compact, $\underline{\mathcal{L}}$ and $\overline{\mathcal{L}}$ are well-defined as minimal and maximal values as soon as Θ is compact.*

- *If $\mathcal{L}$ also verifies assumption (H.5), the bound (15) can be extended to hold uniformly over θ on any convex compact subset $\mathcal{C} \subset \Theta$. Indeed, it can be shown, combining our proof in A.5 with a covering argument, that*

$$\sup_{\theta\in\mathcal{C}} |g_{\tau,N}(\theta) - g(\theta)| \leq h(\tau) + \tau e^{\frac{\overline{\mathcal{L}}-\underline{\mathcal{L}}}{\tau}} \sqrt{\frac{(p+t)\log N}{2N}} + 8\frac{L_{\Theta,\mathcal{Z}} D_{\mathcal{C}}}{N},$$

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

- *It is important to realize that the claim of Theorem 4.6(ii)(b) is different (and stronger) than that of Theorem 4.6(ii)(a); observe the order of quantifiers. In the latter, the set of events of probability one on which Theorem 4.6(ii)(a) holds actually depends on θ, while it is does not for claim (b). On the other hand getting this uniform claim requires some additional regularity. This discussion is very important when it will come to showing the convergence result of our SGD algorithm.*

For fixed τ, the convergence rate in (15) is nearly $O(N^{-1/2})$. But one has to keep in mind that we have not used any smoothness property of $\mathcal{L}(\theta,\cdot)$ and used a very simple (uniform) Monte Carlo integration. This could be possibly improved using the rich theory of Monte Carlo integration, see e.g., [32, 33], 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 (15) through the exponential term. One can alternatively use Bernstein's inequality instead of Hoeffding to show that

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

where

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

The error bound (15) 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 enforce the term $e^{\frac{\overline{\mathcal{L}}-\underline{\mathcal{L}}}{\tau}} \sqrt{\frac{t\log N}{2N}}$ to converge to 0. Taking $\tau = O\left(\frac{1}{\kappa \log N}\right)$, for any $\kappa > 0$ such that $\kappa(\overline{\mathcal{L}} - \underline{\mathcal{L}}) < 1/2$, the convergence rate in (15) is dominated by the first term $h(\tau)$ which scales as $O\left((\log N)^{-1}\right)$. This is obviously a slow convergence rate but reflects the difficulty of approximating the function g in (8).

4.4 Consistency of subgradient estimates

Equipped with the above results, a natural strategy now is to solve the PRO problem (8) by using $g_{\tau,N}$ in (14) 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

$$\nabla g_{\tau,N}(\theta) = \sum_{k=1}^{N} \nabla_\theta \mathcal{L}(\theta, z'_k) \frac{\exp\left(\frac{\mathcal{L}(\theta,z'_k)}{\tau}\right)}{\sum_{j=1}^{N} \exp\left(\frac{\mathcal{L}(\theta,z'_j)}{\tau}\right)}. \tag{16}$$

The natural question that arises is whether (16) 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 that are stronger than those required for consistency of the (zero-th order oracle) function values established in Theorem 4.6.

To show some of our results, in particular Theorem 4.8(ii), we will need some regularity properties on the set of maximizers $\mathcal{M} := \operatorname{Argmax} \mathcal{L}(\theta, \mathcal{C}) = \operatorname{Argmax}_{z\in\mathcal{C}} \mathcal{L}(\theta, z)$ (we drop the dependence of $\mathcal{M}$ on θ to lighten notation). 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 $m \geq \dim(\mathcal{M}_1) > \dim(\mathcal{M}_2) > \cdots > \dim(\mathcal{M}_{|I|}) \geq 0$.

Theorem 4.8. *Suppose that the assumptions of Proposition 4.1 hold and that*

(H.5) *$\mathcal{L}(\cdot, z')$ is differentiable with $\nabla_\theta \mathcal{L}(\cdot,\cdot)$ continuous in its both arguments (θ, z') and uniformly bounded by $L_{\Theta,\mathcal{Z}}$ on $\Theta \times \mathcal{Z}$.*

Then $\mathcal{L}$ and $g_{\tau,N}$ are continuously differentiable.

(i) If (H.3) also holds, then for every $\theta \in \Theta$ and any $t > 0$, with probability at least $1 - 2(p+1)N^{-t}$

$$\operatorname{dist}\left(\nabla g_{\tau,N}(\theta), \partial^C g(\theta)\right) \leq \max(1, 2L_{\Theta,\mathcal{Z}}\sqrt{p}) e^{\frac{\overline{\mathcal{L}}-\underline{\mathcal{L}}}{\tau}} \sqrt{\frac{t \log N}{2N}} + o_\tau(1). \tag{17}$$

where

$$\partial^C g(\theta) = \left\{ \int_{\mathcal{C}_z^\varepsilon} \nabla_\theta \mathcal{L}(\theta, z') \mathrm{d}\mu(z') : \ \mu \in \mathcal{P}\Big(\operatorname{Argmax} \mathcal{L}(\theta, \mathcal{C}_z^\varepsilon) \Big) \right\}. \tag{18}$$

(ii) Suppose in addition that

(H.6) *$\mathcal{L}(\theta, .)$ is $\mathscr{C}^3$ on an open set containing $\mathcal{C}_z^\varepsilon$ with Hölder continuous third-order derivative;*

(H.7) *for $r \geq 3$*

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

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

Then for every $\theta \in \Theta$ and any $t > 0$, with probability at least $1 - 2(p+1)N^{-t}$

$$\operatorname{dist}\left(\nabla g_{\tau,N}(\theta), \partial^C g(\theta)\right) \leq \|\nabla g_{\tau,N}(\theta) - \eta(\theta)\| \leq \max(1, 2L_{\Theta,\mathcal{Z}}\sqrt{p}) e^{\frac{\overline{\mathcal{L}}-\underline{\mathcal{L}}}{\tau}} \sqrt{\frac{t \log N}{2N}} + o_\tau(1), \tag{19}$$

where $\eta(\theta) := \int_{\mathcal{M}_1} \nabla_\theta \mathcal{L}(\theta, z') \mathrm{d}\mu(z') \subset \partial^C g(\theta)$ and $\mu \in \mathcal{P}(\mathcal{M}_1)$.

(iii) Under the assumptions of either statement (i) or (ii), if τ is a function of N, say τ_N, with $\tau_N \to 0$ and $e^{\frac{\overline{\mathcal{L}}-\underline{\mathcal{L}}}{\tau_N}}\sqrt{\frac{\log N}{N}} \to 0$ as $N \to +\infty$, then

(a) for every $\theta \in \Theta$

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

(b) Let Ξ be any closed convex subset of Θ. If moreover $\operatorname{Argmax} \mathcal{L}(\theta, \mathcal{C}_\varepsilon^z)$ is a singleton for each $\theta \in \Xi$, then almost surely,

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

The proof of this result follows from by a simple triangle inequality and using the following two lemmas whose proofs are deferred to Appendix A.6.

Lemma 4.9. *(i) Under the assumptions of Theorem 4.8(i), we have*

$$\operatorname{dist}\left(\nabla g_\tau(\theta), \partial^C g(\theta)\right) \to 0 \quad \textit{as} \quad \tau \to 0^+. \tag{20}$$

(ii) Under the assumptions of Theorem 4.8(ii),

$$\nabla g_\tau(\theta) \underset{\tau\to0^+}{\longrightarrow} \eta(\theta) := \int_{\mathcal{M}_1} \nabla_\theta \mathcal{L}(\theta, z) \mathrm{d}\mu(z) \subset \partial^C g(\theta),$$

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

Lemma 4.10. *Suppose that the assumptions of Proposition 4.1 hold, and that $\mathcal{L}(\cdot, z')$ is differentiable with $\nabla_\theta \mathcal{L}(\theta, z')$ uniformly bounded by $L_{\Theta,\mathcal{Z}}$ on $\Theta \times \mathcal{Z}$.*

(i) For any $t > 0$

$$\|\nabla g_\tau(\theta) - \nabla g_{\tau,N}(\theta)\| \leq \max(1, 2L_{\Theta,\mathcal{Z}}\sqrt{p}) e^{\frac{\overline{\mathcal{L}}-\underline{\mathcal{L}}}{\tau}} \sqrt{\frac{t \log N}{2N}} \tag{21}$$

with probability at least $1 - 2(p+1)N^{-t}$.

(ii) Suppose that τ is a function of N, say τ_N, with $\tau_N e^{\frac{\overline{\mathcal{L}}-\underline{\mathcal{L}}}{\tau_N}}\sqrt{\frac{\log N}{N}} \to 0$ as $N \to +\infty$, then

(a) for every $\theta \in \Theta$

$$\nabla g_{\tau_N}(\theta) - \nabla g_{\tau_N,N}(\theta) \underset{N\to+\infty}{\longrightarrow} 0 \quad \textit{almost surely.}$$

(b) Let Ξ be any closed convex subset of Θ. If moreover $\operatorname{Argmax} \mathcal{L}(\theta, \mathcal{C}_\varepsilon^z)$ is a singleton for each $\theta \in \Xi$, then almost surely,

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

To show Lemma 4.9, one observes that under our assumptions,

$$\nabla g_\tau(\theta) = \int_{\mathcal{C}_z^\varepsilon} \nabla_\theta \mathcal{L}(\theta, z') \frac{\exp\left(\frac{\mathcal{L}(\theta,z')}{\tau}\right)}{\int_{\mathcal{C}_z^\varepsilon} \exp\left(\frac{\mathcal{L}(\theta,v)}{\tau}\right) \mathrm{d}v} \mathrm{d}z', \tag{22}$$

which is an expectation with respect to a Gibbs measure indexed by τ (and θ) and supported on $\mathcal{C}_z^\varepsilon$. In view of the rule (5), the proof will then amount to showing that as $\tau \to 0^+$, the familty of such Gibbs measures has all its cluster points in the narrow topology (equivalent to the weak–$*$ topology) in $\mathcal{P}\Big(\operatorname{Argmax} \mathcal{L}(\theta, \mathcal{C}_z^\varepsilon)\Big)$ (first claim of Lemma 4.9), or that it even converges in the weak–$*$ topology to a measure supported on $\operatorname{Argmax}_{z\in\mathcal{C}} \mathcal{L}(\theta, z)$ (second claim of Lemma 4.9)[4].

Clearly, Theorem 4.8 tells us that, with high probability, $\nabla g_{\tau,N}(\theta)$ is at most within a ball around the Clarke subdifferential of g at θ. The result also quantifies its radius and how it vanishes with N and τ, and shows the influence of p, the dimension of Θ. Arguing as above, this radius vanishes as $N \to +\infty$ by taking $\tau = O\left(\frac{1}{\kappa \log N}\right)$, for any κ such that $\kappa(\overline{\mathcal{L}} - \underline{\mathcal{L}}) \in]0, 1/2 - \alpha]$, $\alpha \in]0, 1/2[$. The convergence rate of the first term in (17) or (19) is then nearly $O\left(N^{-\alpha}\right)$ (up to logarithmic factors). As far as the $o_\tau(1)$ term is concerned, we do not have any quantitative estimate for the corresponding rate in the case of (17). For (19), a close inspection of the proof Lemma 4.9(ii) reveals that the convergence rate in τ is at least

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

Remark 4.11. *It is worth noting that the statement of Lemma 4.9(i), hence Theorem 4.8(i), requires less stringent assumptions than claim (i). However, it only ensures subsequential convergence of the gradient whose cluster points are Clarke subgradients of g. On the other hand, Lemma 4.9(ii) not only shows global convergence of the gradient but also gives the precise form of the limit Clarke subgradient, and characterizes the corresponding measure. This in turn necessitates the extra regularity assumptions above.*

Remark 4.12. *Similarly to the discussion in Remark 4.7, we again need a little bit more to ensure almost sure convergence simultaneously for all θ. This observation is very important when it will come to using almost sure unbiasedness of the (sub)gradient estimate in our SGD algorithm, which in turn will be crucial to prove our convergence result in Theorem 5.2.*

Remark 4.13. *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 [36, 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 (35), and then use that their dimensions are strictly smaller than that of the largest submanifold.*

[4]This is reminiscent of works on simulated annealing where τ is the temperature parameter; see e.g. [34, 35]. Our context is however different and in particular, $\mathcal{C}_z^\varepsilon$ is not the whole space nor it is a finite set nor a compact submanifold.

5 Robust optimization algorithm via SGD

We are now ready to describe our algorithmic framework to solve the PRO problem (8) based on stochastic (sub)gradient descent. To make the presentation easier, we assume that $\Theta = \mathbb{R}^p$ though our algorithmic framework can be extended to the case where Θ is a convex closed set by including a projection step onto Θ (see e.g., [37, 38] in different settings). Moreover, as considered in most applications, we take in this section

$$\mathsf{c}(z, z') = \varphi(z - z'),$$

where $\varphi : \mathbb{R}^m \to \mathbb{R}_+$ is a continuous coercive function. It is also even-symmetric and $\varphi(0) = 0$ by assumption (H.2). 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

$$\min_{\theta\in\Theta} \left\{ G(\theta) := \frac{1}{M}\sum_{i=1}^{M} g_i(\theta) \right\} \quad \text{where} \quad g_i(\theta) := \max_{u\in\mathcal{C}^\varepsilon} \mathcal{L}(\theta, z_i + u), \tag{23}$$

where $\mathcal{C}^\varepsilon := \{u \in \mathbb{R}^m : \ \varphi(u) \le \varepsilon\}$ is obviously a nonempty compact and full dimensional set by the assumptions on φ. It is worth observing that convexity of φ is not needed in this section.

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.1), is $\mathcal{L}(\cdot, z_i+u)$ is assumed locally Lipschitz continuous for each (z_i, u), (4) gives us the inclusion

$$\partial^C G(\theta) = \partial^C \left(\frac{1}{M}\sum_{i=1}^{M} g_i(\theta) \right) \subset \frac{1}{M}\sum_{i=1}^{M} \partial^C g_i(\theta) \subset \frac{1}{M}\sum_{i=1}^{M} \mathcal{D}_i(\theta), \tag{24}$$

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. After a first version of this work was posted, we became aware of the recent work of [41] who established that conservative fields deriving from definable potentials also have a calculus rule under pointwise maximization involving again an inner maximization oracle. Thus in the rest of this subsection, we could just use the formula of that conservative field instead of $\mathcal{D}_i$ and our convergence result will remain true. The advantage of using conservative fields is their rich calculus, including the sum and chain rules. Nevertheless, we will not elaborate more on this to keep the presentation simpler and since anyway, this would necessitate to have the inner maximization oracle.*

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

$$\text{crit–}G := \left(\frac{1}{M}\sum_{i=1}^{M} \mathcal{D}_i \right)^{-1}(0). \tag{25}$$

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

$$\theta_{k+1} = \theta_k - \gamma_k d_k, \quad \text{where} \quad d_k \in \frac{1}{|B_k|}\sum_{i\in B_k} \mathcal{D}_i(\theta_k), \tag{26}$$

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 (26) 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 [42], and used recently in [39, 43]. 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

$$\begin{cases} 0 \in \dot\theta(t) + \frac{1}{M}\sum_{i=1}^M \mathcal{D}_i(\theta(t)) & \text{for almost every } t \in \mathbb{R}, \\ \theta(0) = \theta_0, \end{cases} \tag{27}$$

whose stationary solutions are the critical points in (25). A key argument to invoke the results of [42], is to build an appropriate Lyapunov function and show that the function G is path differentiable, that is, it obeys the chain rule

$$\frac{\mathrm{d}}{\mathrm{d}t} G(\theta(t)) = \langle \dot\theta(t), v\rangle, \quad \forall v \in \frac{1}{M}\sum_{i=1}^M \mathcal{D}_i(\theta(t)). \tag{28}$$

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 (18) applies and (24) becomes

$$\partial^C G(\theta) = \partial^C\left(\frac{1}{M}\sum_{i=1}^M g_i(\theta)\right) \subset \frac{1}{M}\sum_{i=1}^M \partial^C g_i(\theta), \tag{29}$$

where

$$\partial^C g_i(\theta) = \left\{ \int_{\mathcal{C}^\varepsilon} \nabla_\theta \mathcal{L}(\theta, z_i + u)\mathrm{d}\mu(u) : \ \mu \in \mathcal{P}\Big(\operatorname{Argmax} \mathcal{L}(\theta, z_i + \mathcal{C}^\varepsilon)\Big)\right\}. \tag{30}$$

This gives the scheme in Algorithm 1.

Algorithm 1: SGD for PRO without smoothing.

Input: Step-sizes $(\gamma_k)_{k\in\mathbb{N}}$;
Input: Initialization θ_0;
for $k = 0, \dots$ **do**
  Draw independently uniformly at random a mini-batch $B_k \subset [M]$;
  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) holds, 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 $\nabla_\theta \mathcal{L}(\theta, z)$ continuous in (θ, z). Suppose moreover that φ and $\mathcal{L}$ are definable, and that the step-sizes satisfy $\sum_{k\in\mathbb{N}} \gamma_k = +\infty$ and $\gamma_k = o\left(\frac{1}{\log k}\right)$. Consider the sequence $(\theta_k)_{k\in\mathbb{N}}$ generated by Algorithm 1 and suppose that there exists a constant $C > 0$ such that $\sup_{k\in\mathbb{N}} \|\theta_k\| \leq C$ almost surely. Then, almost surely, the set of cluster points of $(\theta_k)_{k\in\mathbb{N}}$ belong to* crit–$G = \left(\frac{1}{M}\sum_{i=1}^M \partial^C g_i\right)^{-1}(0)$. *Moreover $(G(\theta_k))_{k\in\mathbb{N}}$ converges and G is constant on* crit–G.

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 (30). 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

In this section, we work under the assumptions of Theorem 4.8 and capitalize on the results there. This gives the scheme summarized in Algorithm 2.

Algorithm 2: SGD for PRO with smoothing.

Input: Step-sizes $(\gamma_k)_{k\in\mathbb{N}}$; number of integration points $(N_k)_{k\in\mathbb{N}}$; smoothing parameters $(\tau_k)_{k\in\mathbb{N}}$;
Input: Initialization θ_0;
for $k = 0, \dots$ **do**
  Draw independently uniformly at random a mini-batch $B_k \subset [M]$;
  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 $\nabla g^i_{\tau_k, N_k}(\theta_k) := \sum_{j=1}^{N_k} \nabla_\theta \mathcal{L}(\theta, z_i + u_j) \frac{e^{\frac{\mathcal{L}(\theta, z_i+u_j)}{\tau_k}}}{\sum_{l=1}^{N_k} e^{\frac{\mathcal{L}(\theta, z_i+u_l)}{\tau_k}}}$;
  $\theta_{k+1} = \theta_k - \gamma_k d_k$.

The direction d_k in Algorithm 2 can also be written as

$$d_k = v_k + e_k + \zeta_k,$$

where

- $v_k = \frac{1}{M}\sum_{i=1}^M \mathrm{P}_{\partial^C g_i(\theta_k)}(\nabla g^i_{\tau_k, N_k}(\theta_k))$;
- $e_k = \frac{1}{|B_k|}\sum_{i\in B_k}\left(\nabla g^i_{\tau_k, N_k}(\theta_k) - \mathrm{P}_{\partial^C g_i(\theta_k)}(\nabla g^i_{\tau_k, N_k}(\theta_k))\right)$;
- $\zeta_k = \frac{1}{|B_k|}\sum_{i\in B_k} \mathrm{P}_{\partial^C g_i(\theta_k)}(\nabla g^i_{\tau_k, N_k}(\theta_k)) - \frac{1}{M}\sum_{l=1}^M \mathrm{P}_{\partial^C g_l(\theta_k)}(\nabla g^l_{\tau_k, N_k}(\theta_k))$;

where all projectors above are well-defined since the Clarke subdifferential is closed (in fact compact) and convex-valued. We then have the following convergence result.

Theorem 5.3. *Assume that the assumptions of Theorem 4.8(ii)(b) hold. Suppose moreover that φ and $\mathcal{L}$ are definable, that the step-sizes satisfy $\sum_{k\in\mathbb{N}} \gamma_k = +\infty$ and $\gamma_k = o\left(\frac{1}{\log k}\right)$, and that $(\tau_k, N_k)_{k\in\mathbb{N}}$ are such that $\tau_k \to 0$ and $e^{\frac{\overline{\mathcal{L}}-\underline{\mathcal{L}}}{\tau_k}}\sqrt{\frac{\log N_k}{N_k}} \to 0$ as $k \to +\infty$, and $\sum_{k\in\mathbb{N}} N_k^t < +\infty$ for some $t > 0$. Consider*

the sequence $(\theta_k)_{k\in\mathbb{N}}$ *generated by Algorithm 2 and suppose that there exists a constant* $C > 0$ *such that* $\sup_{k\in\mathbb{N}} \|\theta_k\| \leq C$ *almost surely. Then, almost surely, the set of cluster points of* $(\theta_k)_{k\in\mathbb{N}}$ *belong to* crit–$G = \left(\frac{1}{M}\sum_{i=1}^{M} \partial^C g_i\right)^{-1}(0)$. *Moreover* $(G(\theta_k))_{k\in\mathbb{N}}$ *converges and* G *is constant on* crit–G.

See Appendix A.8 for the proof.

From the discussion after Lemma 4.9, the vanishing assumption on the bias holds provided one chooses N_k an increasing function of k, and τ_k decreasing as $O\left(\frac{1}{\kappa \log N_k}\right)$ for $\kappa > 0$ small enough. The algorithm remain simple and effective at making the learning robust, the initial value of the parameters N and T do not matter as long as we increase and decrease them respectively. However in practice it is better to select their initial values considering the dimension of the problem. We suggest reading the following sources for considerations on Monte Carlo integration or Quasi Monte Carlo [32] [33]. Other algorithms can be used instead of SGD provided they are proven to converge with appropriate generalized subgradients (see [44] for instance).

6 Numerical results

6.1 Dataset, model and metrics

The following 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 centred 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 (14) reasonable, and because it has unevenly distributed labels which will help highlighting the compromise between generalization and robust learning. In these experiments, 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 model is the cross-entropy loss after a softmax step on the network output, i.e., 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

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

where the subscript here stands for the corresponding entry of a vector. Note that this loss also verifies our assumptions. All experiments were carried out with the same number of epochs, batch size and therefore the same number of updates (see Table 1 for details).

For comparison purposes, we trained the MLP network in 3 different ways: with a vanilla (non-robust) training, an adversarial training using PGD as a heuristic to solve the inner maximization problem [16], and robust training using Algorithm 2. 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 such as adversarial training, for different values of the perturbation radius ε. Throughout this section, we take $\mathcal{C}^\varepsilon = \mathbb{B}_\varepsilon^q(0)$ with typically $q = +\infty$ (see Table 1).

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

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

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

- **Adversarial Accuracy**: This metric is meant to represent the accuracy on an adversarial set given by applying a white-box PGD attack [16]. 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

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

- **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 samples, this metric is defined as

$$\text{Robust Accuracy} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \min_{(u_j)_{j=1}^N,\, u_j \sim_{\text{i.i.d.}} \mu_{\mathcal{U}}} S(F_\theta(x_i + u_j), y_i).$$

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

We added to these metrics estimates of an upper-bound of the Lipschitz constant of the 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.

These metrics are to be evaluated on the three kinds of training, 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 variance of the results would come only from differences in the initialization of the model and dynamics of the optimization. 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 intractable robust* 0-1 *gain. It is closely linked to the adversarial frequency [48]*

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

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, but the latter does not enjoy any convergence guarantee to the minimal value. The robust accuracy metric appears as a better representative metric of the robust behavior of a model with the proviso that N is large.

6.2 Dependence of smoothing factor and number of samples for robust optimization

Following Theorem 4.8, we advocated that N, the number of samples used by Monte Carlo integration must increase during the execution of Algorithm 2 and that $\tau = O(\frac{1}{\kappa \log N})$, the regularization parameter, goes to zero. However in practice, making the sampling infinite is impossible due to hardware memory limitations. 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 with different values of τ and N.

The sampling for the robust training is performed in $\mathbb{B}^{\infty}_{0.05}(0)$ for various values of τ and N. The inner maximization problem is assessed and averaged on the test base resorting to a fixed sampling of the perturbation set. This sampling is chosen larger (10^6) than the biggest value explored in the set so as to maximize the chance that the metric is evaluated accurately when testing.

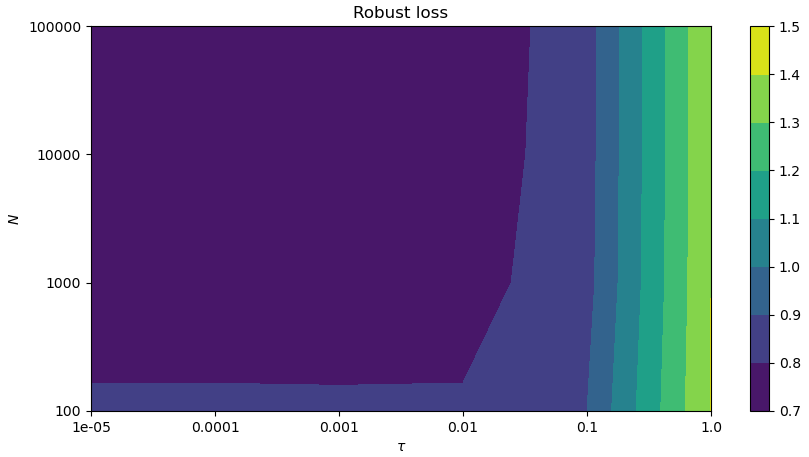

Figure 1: Influence of τ (x axis) and N (y axis) on the robust loss on test set (color scale). The darker the better.

Figure 1 shows the value of the loss for different values of τ and N, after a given number of iterations. The observed behavior is indeed the expected one with respect to the sampling: the more we shrink the temperature and increase the sampling the smaller the error on the robust problem we have.

6.3 Robustness to noise

In this section, we present a few experiments illustrating the robustness to noise a model trained with our method, and provide comparisons with alternative methods.

For these experiments, the adversarial training was performed in $\mathbb{B}^{\infty}_{\varepsilon}(0)$ and so was the sampling for the robust training: we kept the same fixed sampling of $N = 5.10^5$ points and a fixed temperature of $\tau = 10^{-4}$. Fixing the sampling to the highest possible value ensures the smallest error possible on the estimation of the robust loss. We give in Section B more information about the parameters for these experiments.

We plot the evolution of the test accuracy (Figure 2), adversarial accuracy (Figure 3) and worst case accuracy (Figure 4) against the radius of robustness with median and quantiles. On top of this we also

present the evolution of the Lipschitz constant of the learned network (Figure 5) estimated using LipSDP [47].

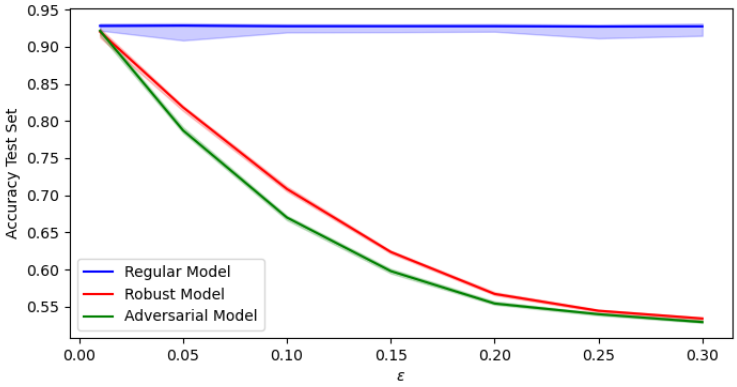

Figure 2: Test accuracy of the three trainings on Avila dataset: vanilla (blue), adversarial (green), robust (red) with median (plain line) and 10% and 90% quantiles for 50 trials.

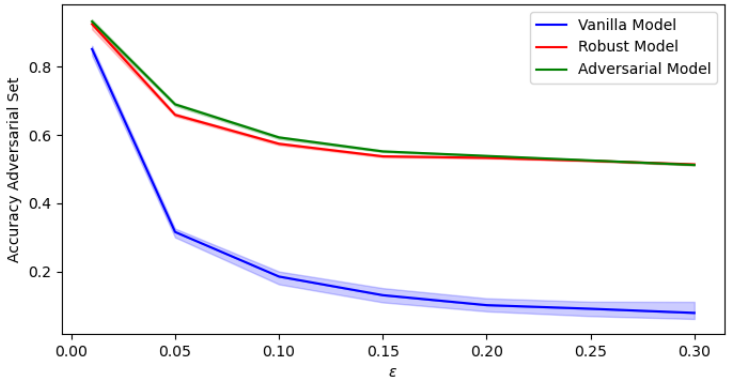

Figure 3: Adversarial accuracy of three trainings on Avila dataset: vanilla (blue), adversarial (green), robust (red) with median (plain line) and 10% and 90% quantiles for 50 trials.

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 the vanilla training, for the robust training with $\varepsilon = 0.3$ and the adversarial training with the same value of ε. The results are displayed in Figure 6.

On all the figures above, the behaviour of our robust training is competitive compared to the popular PGD-based adversarial training, and differences are in general small. We note that robust training is better on the test accuracy for moderate perturbations (Figure 2), while adversarial training appears to be slightly better in termes 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 model 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 vanilla training), confirming that the performance of all the trainings is reproducible from run to run and for different initializations.

The decrease in accuracy of the robust and adversarial training on the test dataset 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 $\mathbb{B}_\varepsilon^q(0)$ around those of label A. Class B has a very 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.

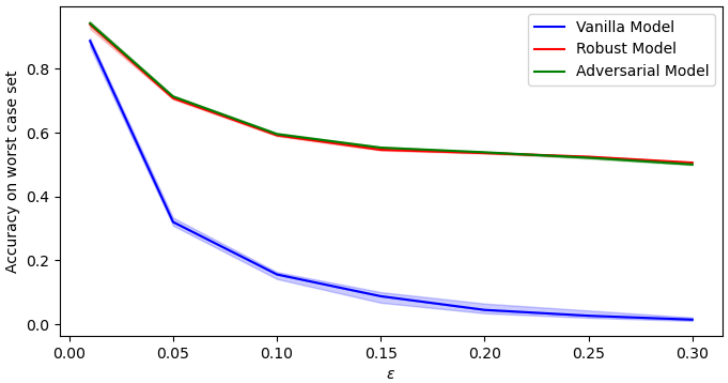

Figure 4: Worst-case robustness accuracy of three trainings on Avila dataset: vanilla (blue), adversarial (green), robust (red) with median (plain line) and 10% and 90% quantiles for 50 trials.

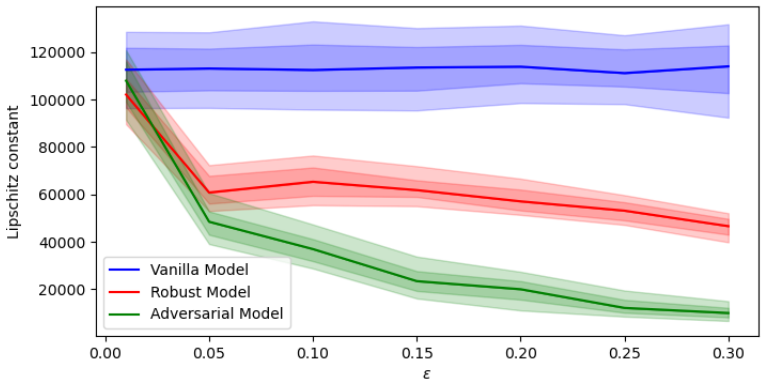

Figure 5: Lipschitz constant upper bound of the learned network by three training methods on Avila dataset: vanilla (blue), adversarial (green), robust (red) with median (plain line) and 10%, 25%, 75% and 90% quantiles for 50 trials.

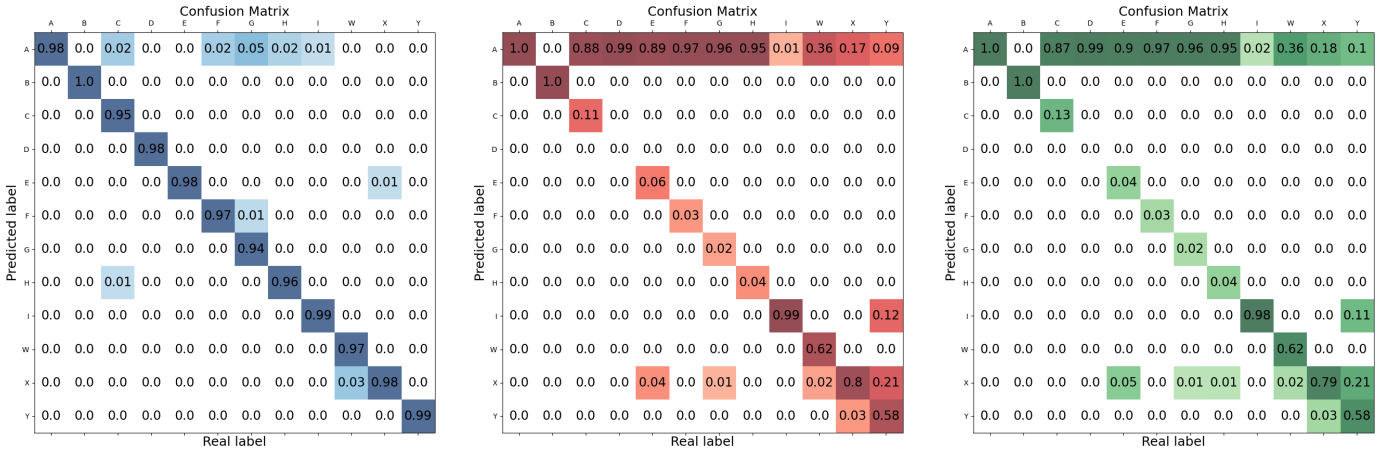

Figure 6: Confusion matrix for the vanilla training (left), for the robust training ($\varepsilon = 0.3$) (middle) and for the adversarial training ($\varepsilon = 0.3$) (right). All matrices are normalized column-wise to display the percentage of predicted labels per each class in the test set.

As far as Figure 5 is concerned, one can clearly (and unsurprisingly) see that the adversarial and robust training methods tend to monotonically reduce the Lipschitz constant of the learned networks as ε increases. The fact that, on an adversarial set, the vanilla model performs poorly compared to the adversarial one is expected (Figure 3), but the performance of the robust model is similar to the adversarial one on this dataset. Of course, adversarial attacks favor those models that have already been trained in an adversarial setting. Our experiments confirm that our approach via smoothing and Monte Carlo sampling on the perturbation set provably converges to a critical point of the PRO problem. This is in contrast to PGD-based adversarial training for which convergence guarantees are not available in such general setting. The main drawback of the robust training remains its computational cost compared to adversarial training. For instance, under the hyperparameters in Appendix B with $\varepsilon = 0.3$, both training methods were performed on the same GPU A100 with the same number of epochs and updates. Adversarial training took 47 min whereas robust training took 11 hours. Adversarial training remains a very effective method for solving empirically the PRO problem. Let us stress that, despite this increased cost, robust training can be more effectively parallelized as it only requires one expensive forward pass and one expensive backward pass whereas the adversarial training requires multiple iterations of both.

7 Conclusions

Solving numerically the DRO problem beyond stringent assumptions on the loss remains a challenging open problem. Here, we have shown that the DRO problem with sufficiently small error can be approached with a PRO problem. In order to solve the latter, we designed SGD-type algorithm hinging on smoothing of the inner maximization problem and Monte Carlo sampling. Our approach is one of the few that enjoys provable convergence guarantees at the expense of an overall higher computational cost. Our robust training has given performances similar to those of adversarial training on practical examples. This showcases the soundness of using robust training with an adequate sampling. However, in machine leaning applications with overparametrized models involving a very large number of parameters and high dimensional input space, scalability of our robust training framework is still a challenge due in particular to our simple Monte Carlo sampling step. More sophisticated sampling strategies, for instance those based on Langevin diffusion, is one direction that is worth investigating in a future work.

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] 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.

[4] 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.

[5] 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.

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

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

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

[22] 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.

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

[24] 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.

[25] 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.

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

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

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

[29] 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.

[30] 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.

[31] 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.

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

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

[34] 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.

[35] 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.

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

[37] 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.

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

[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] E. Pauwels. The ridge method for tame min-max problems. hal-03186676, February 2022.

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

[43] 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.

[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-Arné 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] 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.

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

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

[58] 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.

[59] Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Séminaire de probabilités XXXIII*, volume 1709 of *Lecture Notes in Mathematics*, pages 1–68. Springer, 1999.

[60] 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

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

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

Taking the expectation on both sides, we get

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

In turn, since $\gamma \geq 0$ was arbitrary, we take the infimum on the left-hand side and use the identity (6), which holds under assumptions (H.1) and (H.2), to get the lower bound.

Let us turn to the upper-bound. We embark from (6) and consider the case where $\gamma = L_{\mathcal{Z}}$:

$$\begin{aligned}\sup_{W_{\mathsf{c}}(\rho, \rho_0) \leq \varepsilon} \mathbb{E}_{z \sim \rho}[\mathcal{L}(\theta, z)] &\leq L_{\mathcal{Z}} \varepsilon + \mathbb{E}_{z \sim \rho_0} \left[\sup_{z' \in \mathcal{Z}} \left(\mathcal{L}(\theta, z') - L_{\mathcal{Z}} \mathsf{c}(z', z) \right) \right] \\ &\leq L_{\mathcal{Z}} \varepsilon + \mathbb{E}_{z \sim \rho_0} \left[\mathcal{L}(\theta, z) \right] \\ &\leq L_{\mathcal{Z}} \varepsilon + \mathbb{E}_{z \sim \rho_0} \left[\sup_{\mathsf{c}(z,z') \leq \varepsilon} \mathcal{L}(\theta, z') \right]\end{aligned}$$

where we used the uniform Lipschitz continuity assumption (H.3) in the second inequality: $\mathcal{L}(\theta, z') - \mathcal{L}(\theta, z) \leq |\mathcal{L}(\theta, z') - \mathcal{L}(\theta, z)| \leq L_{\mathcal{Z}} \mathsf{c}(z', z)$ and that z is a feasible solution in the last constrained supremum since $\mathsf{c}(z, z') = 0$ when $z = z'$.

(ii) The proof of the lower-bound part is the same as in the first claim above. Let us turn to the upper-bound. We use again (6) and Lipschitz continuity of $\mathcal{L}(\theta, \cdot)$ to get that for any $\gamma \geq 0$,

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

Optimizing for t and after basic algebra, we get

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

This upper-bound is minimal for $\gamma = c_{q,L_{\mathcal{Z}}} \varepsilon^{-\frac{q-1}{q}}$, for an explicit constant $c_{q,L_{\mathcal{Z}}}$. Plugging this value of γ in the upper-bound, 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 $\mathcal{P}(\mathcal{C}_z^\varepsilon)$ with the weak–$*$ topology. Continuity of c implies closedness of $\mathcal{C}_z^\varepsilon$. This together with its boundedness assumption imply compactness of $\mathcal{C}_z^\varepsilon$ as it is finite dimensional. Thus $\mathcal{P}(\mathcal{C}_z^\varepsilon)$ 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 $\mathcal{C}_z^\varepsilon$ that $\mu \mapsto \int_{\mathcal{C}_z^\varepsilon} \mathcal{L}(\theta, z') \mathrm{d}\mu(z')$ is weak–$*$ continuous. It is known that $\mathrm{KL}(\cdot, \mu_{\mathcal{U}})$ is convex and lower semicontinuous in the weak–$*$ topology on $\mathcal{P}(\mathcal{C}_z^\varepsilon)$. Thus, since $\tau > 0$, the objective in (10) is convex and upper semicontinuous. This together with convex and weak–$*$ compactness of $\mathcal{P}(\mathcal{C}_z^\varepsilon)$ entail that (10) has a non-empty convex and weak–$*$ compact set of solutions. Uniqueness of the minimizer then follows from strong convexity of $\mathrm{KL}(\cdot, \mu_{\mathcal{U}})$ on $\mathcal{P}(\mathcal{C}_z^\varepsilon)$ thanks to the celebrated Pinsker's inequality. The closed form solution follows from standard calculus of variations and Lagrangian duality; see e.g., [31, Lemma 6.6]. □

A.3 Proof of Theorem 4.3

(i) Define $\psi_\tau(\theta) := \int_{\mathcal{C}_z^\varepsilon} \mathcal{L}(\theta, z') \mathrm{d}\mu(z') - \tau \mathrm{KL}(\mu, \mu_{\mathcal{U}})$. The function $\tau \mapsto \psi_\tau(\theta)$ obviously increases as τ decreases and so is g_τ. Continuity of $\mathcal{L}$, compactness of $\mathcal{C}_z^\varepsilon$ and Proposition 4.1 entail that g_τ is continuous and converges pointwise to g. The Γ-convergence claim in this case then follows from [23, Proposition 5.4 and Remark 5.5].

(ii) The upper bound in (12) is immediate by definition of g_τ. Let us turn to the lower bound. Let us denote $z^\star \in \mathrm{Argmax}_{z' \in \mathcal{C}_z^\varepsilon} \mathcal{L}(\theta, z')$, where the latter is a non-empty compact set thanks to continuity of $\mathcal{L}(\theta, \cdot)$ and compactness of $\mathcal{C}_z^\varepsilon$. We then have

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

By definition of $R_{\mathcal{C}_z^\varepsilon}$, there exists $\bar{z}$ such that $\mathbb{B}^{R_{\mathcal{C}_z^\varepsilon}}(\bar{z}) \subset \mathcal{C}_z^\varepsilon$. Convexity of $\mathcal{C}_z^\varepsilon$ entails that:

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

and thus:

$$\begin{aligned}
\tau \log\left(\int_{\mathcal{C}_z^\varepsilon} e^{\frac{\mathcal{L}(\theta,z')}{\tau}} \mathrm{d}z'\right) &\geq g(\theta) + \tau \log \int_{\tau(\mathbb{B}_{R_{\mathcal{C}_z^\varepsilon}}(\bar{z})-z^\star)+z^\star} e^{\frac{-L_{\mathcal{Z}}\|z'-z^\star\|}{\tau}} \mathrm{d}z' \\
&= g(\theta) + \tau \log\left(\tau^m \int_{\mathbb{B}_{R_{\mathcal{C}_z^\varepsilon}}(\bar{z})} e^{-L_{\mathcal{Z}}\|z'-z^\star\|} \mathrm{d}z'\right) \\
&\geq g(\theta) + \tau \log\left(\tau^m \int_{\mathbb{B}_{R_{\mathcal{C}_z^\varepsilon}}(\bar{z})} e^{-L_{\mathcal{Z}}(\|z'-\bar{z}\|+\|z^\star-\bar{z}\|)} \mathrm{d}z'\right) \\
&\geq g(\theta) + \tau \log\left(\tau^m \int_{\mathbb{B}_{R_{\mathcal{C}_z^\varepsilon}}(0)} e^{-L_{\mathcal{Z}}(R_{\mathcal{C}_z^\varepsilon}+D_{\mathcal{C}_z^\varepsilon})} \mathrm{d}z'\right) \\
&= g(\theta) - m\tau \log(\tau^{-1}) + \tau \log(\mu_{\mathcal{L}}(\mathbb{B}_{R_{\mathcal{C}_z^\varepsilon}}(0))) - \tau L_{\mathcal{Z}}(R_{\mathcal{C}_z^\varepsilon} + D_{\mathcal{C}_z^\varepsilon}).
\end{aligned} \tag{31}$$

Inserting this into the expression of g_τ (see (11)), we get the upper-bound. □

A.4 Proof of Theorem 4.5

Compactness of Θ entails that g_τ and g are equi-coercive (see [23, Definition 7.6 and Proposition 7.7]. The first claim on convergence of the minimal values follows by combining the first claim in Theorem 4.3 and [23, Theorem 7.8]. The second claim is a consequence of Γ-convergence of g_τ (Theorem 4.3), compactness of Θ and [23, 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 used (12). It remains to bound the last term. This is the subject of the following lemma.

Lemma A.1. *Under the assumptions of Theorem 4.3, the following holds.*

(i) For any $t > 0$ and fixed $\theta \in \Theta$,

$$|g_{\tau,N}(\theta) - g_\tau(\theta)| \leq \tau e^{\frac{\overline{\mathcal{L}}-\underline{\mathcal{L}}}{\tau}} \sqrt{\frac{t \log N}{2N}},$$

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

(ii) Suppose that τ is a function of N, say τ_N, with $\tau_N e^{\frac{\overline{\mathcal{L}}-\underline{\mathcal{L}}}{\tau_N}} \sqrt{\frac{\log N}{N}} \to 0$ as $N \to +\infty$, then

(a) for every $\theta \in \Theta$

$$g_{\tau_N,N}(\theta) - g_{\tau_N}(\theta) \underset{N\to+\infty}{\longrightarrow} 0 \quad \textit{almost surely.}$$

(b) If moreover (H.4) holds then, almost surely,

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

Proof. To lighten the notation, denote

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

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

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

We then have

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

Using the standard inequality $\log(1+t)\le t$ for $t\ge 0$, we can write, for any $\epsilon\ge 0$

$$\begin{aligned}
&\Pr\left(|g_{\tau,N}(\theta) - g_\tau(\theta)| \ge \epsilon\right) = \Pr\left(\left|\log\left(\frac{S_N}{\mathbb{E}[S_N]}\right)\right| \ge \epsilon/\tau\right)\\
&= \Pr\left(\log\left(\frac{S_N}{\mathbb{E}[S_N]}\right) > \epsilon/\tau\right)\mathbb{1}\left(S_N \ge \mathbb{E}[S_N]\right) + \Pr\left(\log\left(\frac{\mathbb{E}[S_N]}{S_N}\right) > \epsilon/\tau\right)\mathbb{1}\left(S_N \le \mathbb{E}[S_N]\right)\\
&\le \Pr\left(\frac{S_N - \mathbb{E}[S_N]}{\mathbb{E}[S_N]} > \epsilon/\tau\right)\mathbb{1}\left(S_N \ge \mathbb{E}[S_N]\right) + \Pr\left(\frac{\mathbb{E}[S_N] - S_N}{S_N} > \epsilon/\tau\right)\mathbb{1}\left(S_N \le \mathbb{E}[S_N]\right)\\
&\le \Pr\left(S_N - \mathbb{E}[S_N] > e^{\underline{\mathcal{L}}/\tau}\epsilon/\tau\right)\mathbb{1}\left(S_N \ge \mathbb{E}[S_N]\right) + \Pr\left(S_N - \mathbb{E}[S_N] < -e^{\underline{\mathcal{L}}/\tau}\epsilon/\tau\right)\mathbb{1}\left(S_N \le \mathbb{E}[S_N]\right)\\
&= \Pr\left(|S_N - \mathbb{E}[S_N]| > e^{\underline{\mathcal{L}}/\tau}\epsilon/\tau\right)\mathbb{1}\left(S_N \ge \mathbb{E}[S_N]\right) + \Pr\left(|S_N - \mathbb{E}[S_N]| > e^{\underline{\mathcal{L}}/\tau}\epsilon/\tau\right)\mathbb{1}\left(S_N \le \mathbb{E}[S_N]\right)\\
&= \Pr\left(|S_N - \mathbb{E}[S_N]| > e^{\underline{\mathcal{L}}/\tau}\epsilon/\tau\right).
\end{aligned}$$

Since the random variables $e^{\frac{\mathcal{L}(\theta,z'_i)}{\tau}}$ are independent and bounded (they live in the interval $[e^{\underline{\mathcal{L}}/\tau}, e^{\overline{\mathcal{L}}/\tau}]$), we are in position to invoke Hoeffding's inequality to obtain

$$\begin{aligned}
\Pr\left(|g_{\tau,N}(\theta) - g_\tau(\theta)| \ge \epsilon\right) &\le 2\exp\left(-\frac{2N^2 e^{2\underline{\mathcal{L}}/\tau}\epsilon^2}{N\left(e^{\overline{\mathcal{L}}/\tau} - e^{\underline{\mathcal{L}}/\tau}\right)^2\tau^2}\right)\\
&\le 2\exp\left(-\frac{2Ne^{-2(\overline{\mathcal{L}}-\underline{\mathcal{L}})/\tau}\epsilon^2}{\tau^2}\right). \qquad (32)
\end{aligned}$$

Taking

$$\epsilon = \tau e^{(\overline{\mathcal{L}}-\underline{\mathcal{L}})/\tau}\sqrt{\frac{t\log N}{2N}},$$

we get

$$\Pr\left(|g_{\tau,N}(\theta) - g_\tau(\theta)| > \epsilon\right) \le 2e^{-t\log N} = 2N^{-t}$$

(ii) Let $\epsilon_N := \tau_N e^{(\overline{\mathcal{L}}-\underline{\mathcal{L}})/\tau_N}\sqrt{\frac{\log N}{N}}$.

(a) We have from the first claim above that

$$\Pr\left(|g_{\tau_N,N}(\theta) - g_{\tau_N}(\theta)| > \epsilon_N\right) \leq 2N^{-2}.$$

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

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

whence almost sure convergence is immediate.

(b) Since $\mathbb{R}^p$ is separable, there exists a countable set T whose closure is Θ. According to claim (a), for every $\theta \in \Theta$ there exists a set of events Ω_θ of probability one and, for every $\omega \in \Omega_\theta$, $g_{\tau_N,N}(\theta,\omega) - g_{\tau_N}(\theta,\omega) \to 0$. Set $\widetilde{\Omega} = \bigcap_{\theta\in\mathsf{T}} \Omega_\theta$. Since T is countable, a union bound immediately shows that $\widetilde{\Omega}$ is also of probability one. For fixed $\theta \in \Theta$, there exists a sequence $(\theta_k)_{k\in\mathbb{N}}$ in T such that $\theta_k \to \theta$. Let $\omega \in \widetilde{\Omega}$. We have

$$\begin{aligned}|g_{\tau_N,N}(\theta,\omega) - g_{\tau_N}(\theta)| \leq |g_{\tau_N,N}(\theta_k,\omega) - g_{\tau_N,N}(\theta,\omega)| + |g_{\tau_N,N}(\theta_k,\omega) - g_{\tau_N}(\theta_k)| \\ + |g_{\tau_N}(\theta_k) - g_{\tau_N}(\theta)|.\end{aligned}$$

Since $\theta_k \in \mathsf{T}$, $\widetilde{\Omega} \subset \Omega_{\theta_k}$, and as just seen hereabove, the second term in the last inequality vanishes as $N \to +\infty$. Let us turn to the first term. Let

$$I^\theta_{\max}(\omega) = \left\{ i \in [N] : \mathcal{L}(\theta, z_i(\omega)) = \max_{j\in[N]} \mathcal{L}(\theta, z_j(\omega)) \right\}.$$

We have

$$g_{\tau_N,N}(\theta,\omega) = \max_{i\in[N]} \mathcal{L}(\theta, z_i(\omega)) + \tau_N \log\left(|I^\theta_{\max}(\omega)| - 1 + \sum_{i\notin I^\theta_{\max}(\omega)} e^{\frac{\mathcal{L}(\theta,z_i(\omega)) - \max_{i\in[N]} \mathcal{L}(\theta,z_i(\omega))}{\tau_N}} \right)$$

It then follows that

$$\begin{aligned}|g_{\tau_N,N}(\theta_k,\omega) - g_{\tau_N,N}(\theta,\omega)| &\leq |\max_{i\in[N]} \mathcal{L}(\theta_k, z_i(\omega)) - \max_{i\in[N]} \mathcal{L}(\theta, z_i(\omega))| + 2\tau_N \log(p-1) \\ &\leq L_\Theta \|\theta_k - \theta\| + 2\tau_N \log(p-1).\end{aligned}$$

A similar reasoning also yields

$$|g_{\tau_N}(\theta_k) - g_{\tau_N}(\theta)| \leq L_\Theta \|\theta_k - \theta\| + 2\tau_N \log(\mu_{\mathcal{L}}(\mathcal{C}^\varepsilon_z)).$$

Collecting the above we get

$$\limsup_{N\to+\infty} |g_{\tau_N,N}(\theta,\omega) - g_{\tau_N}(\theta)| \leq 2L_\Theta \|\theta_k - \theta\|.$$

Taking the limit as $k \to +\infty$, we obtain $g_{\tau_N,N}(\theta,\omega) - g_{\tau_N}(\theta) \to 0$. This completes the proof.

□

A.6 Proof of Theorem 4.8

A.6.1 Proof of Lemma 4.9

To lighten notation in the proof, we drop the super- and subscript in $\mathcal{C}_z^\varepsilon$. Let the (Gibbs) probability measure[5]

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

Compactness of $\mathcal{C}$ and continuity of $\mathcal{L}(\theta,\cdot)$ imply that $\mathcal{M}$ is a non-empty compact set. Without loss of generality, we assume that $\max \mathcal{L}(\theta,\mathcal{C}) = \mathcal{L}(\theta,\mathcal{M}) = 0$ (otherwise, one can use a simple translation argument).

(i) The proof of this claim is inspired by standard arguments in the literature of simulated annealing and Markov chains (see e.g. [55, Proposition 1.2] or [34, Corollary 2.1 and Proposition 2.3])[6]. We provide a self-contained proof adapted to our setting.

Given $\epsilon > 0$, we define

$$\mathcal{U}^\epsilon = \{u \in \mathbb{R}^m : \ \mathcal{L}(\theta,u) \geq -\epsilon\}.$$

By assumption (H.3), $\mathcal{L}(\theta,\cdot)$ is $L_{\mathcal{Z}}$-Lipschitz continuous, and thus $\mathcal{U}^\epsilon$ is contained in the open tubular neighborhood of radius $\epsilon/L_{\mathcal{Z}}$ around $\operatorname{Argmax}\mathcal{L}(\theta,\mathcal{C})$. This implies that $\mu_{\mathcal{L}}(\mathcal{U}^\epsilon) > 0$. We then have

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

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 $\mathcal{U}^0 = \operatorname{Argmax}\mathcal{L}(\theta,\mathcal{C})$. We thus infer, in view of continuity of $\nabla_\theta\mathcal{L}(\theta,\cdot)$ (by (H.5)) that

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

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 (20).

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

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

(ii) 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 (H.7)(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 $\mathcal{U} := \bigcup_{i\in I} \mathcal{U}_i$. We then have for $f \in \mathscr{C}$

$$\int_{\mathcal{C}} f(z')e^{\frac{\mathcal{L}(\theta,z')}{\tau}}\mathrm{d}z' = \int_{\mathcal{C}\cap\mathcal{U}} f(z')e^{\frac{\mathcal{L}(\theta,z')}{\tau}}\mathrm{d}z' + \int_{\mathcal{C}\setminus(\mathcal{C}\cap\mathcal{U})} f(z)e^{\frac{\mathcal{L}(\theta,z')}{\tau}}\mathrm{d}z'. \tag{33}$$

Since $f(\mathcal{C})$ is compact by compactness of $\mathcal{C}$ and continuity of f, and $\exists \kappa > 0$ such that $\forall z' \in \mathcal{C}\setminus(\mathcal{C}\cap\mathcal{U})$, $\mathcal{L}(\theta, z') \leq -\kappa < \max \mathcal{L}(\theta, \mathcal{C}) = 0$, the second integral in (33) verifies, for any $s \geq 0$,

$$\tau^{-s}\left|\int_{\mathcal{C}\setminus(\mathcal{C}\cap\mathcal{U})} f(z)e^{\frac{\mathcal{L}(\theta,z')}{\tau}}\mathrm{d}z'\right| \leq (\mu_{\mathcal{L}}(\mathcal{C})\sup|f(\mathcal{C})|)\,\tau^{-s}e^{-\kappa/\tau} \to 0 \quad \text{uniformly as } \tau \to 0^+. \tag{34}$$

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

$$\int_{\mathcal{C}\cap\mathcal{U}} f(z')e^{\frac{\mathcal{L}(\theta,z')}{\tau}}\mathrm{d}z' = \sum_{i\in I}\int_{\mathcal{C}\cap\mathcal{U}_i} f(z)e^{\frac{\mathcal{L}(\theta,z')}{\tau}}\mathrm{d}z'. \tag{35}$$

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 [56, Theorem 4.12] (see [56, Definition 4.1] for definition of sets of positive reach). Thus, it follows from [56, 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 [57, Theorem 3.2.22(3)] to get

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

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 [58, 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 [35, Lemma 6.1] hold in our case. The rest of our argument follows then similar lines to those of [35, Theorem 3.1, starting from (9.3)]. This allows us to show that

$$\begin{aligned}\lim_{\tau\to 0^+} \tau^{-\frac{m-m_i}{2}} &\int_{\mathcal{C}\cap\mathcal{U}_i} f(z)e^{\frac{\mathcal{L}(\theta,z)}{\tau}}\mathrm{d}z' \\ &= 2^{\frac{m-m_i}{2}}(m-m_i)\alpha_{(m-m_i)}\beta_{(m-m_i)}\int_{\mathcal{M}_i} f(v)\left(\prod_{j=1}^{m-m_i}\lambda_j(v)^{-\frac{1}{2}}\right)d\mathcal{H}^{m_i}(v),\end{aligned} \tag{36}$$

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.7)(b), α_k is the k-dimensional Lebesgue measure of the unit ball in $\mathbb{R}^k$, and

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

Since the strata are ordered by strictly decreasing dimension, we have from (36) that for any $i > j$,

$$\lim_{\tau\to 0^+} \tau^{-\frac{m-m_j}{2}} \int_{\mathcal{C}\cap\mathcal{U}_i} f(z) e^{\frac{\mathcal{L}(\theta,z)}{\tau}} \mathrm{d}z' = \lim_{\tau\to 0^+} \tau^{\frac{m_j-m_i}{2}} \left(\tau^{-\frac{m-m_i}{2}} \int_{\mathcal{C}\cap\mathcal{U}_i} f(z) e^{\frac{\mathcal{L}(\theta,z)}{\tau}} \mathrm{d}z' \right) = 0. \quad (37)$$

Combining (37) (for $j = 1$) and (36) (for $i = 1$) with (33), (34) and (35), we get

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

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 $\nabla_\theta \mathcal{L}(\theta, z')$, we deduce that

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

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

□

A.6.2 Proof of Lemma 4.10

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

$$\mathrm{d}\mu^\theta_\tau(z') := \frac{1}{\mu_{\mathcal{L}}(\mathcal{C})} \frac{e^{\frac{\mathcal{L}(\theta,z')}{\tau}}}{S^\theta_\tau} \mathrm{d}z' \quad \text{and} \quad \mathrm{d}\mu^\theta_{\tau,N}(z') := \frac{1}{N} \sum_{k=1}^N \frac{e^{\frac{\mathcal{L}(\theta,z'_k)}{\tau}}}{S^\theta_{\tau,N}} \delta_{z'_k}$$

where

$$S^\theta_\tau := \frac{1}{\mu_{\mathcal{L}}(\mathcal{C})} \int_{\mathcal{C}} e^{\frac{\mathcal{L}(\theta,z')}{\tau}} \mathrm{d}z' \quad \text{and} \quad S^\theta_{\tau,N} := \frac{1}{N} \sum_{k=1}^N e^{\frac{\mathcal{L}(\theta,z'_k)}{\tau}}.$$

We have made here the dependence on θ, τ and N explicit as it will make our reasoning clearer especially for proving the last claim of the lemma.

(i) It follows from (16) and (22) that

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

and thus, by assumption (H.5), we get

$$\|\nabla g_\tau(\theta) - \nabla g_{\tau,N}(\theta)\| \leq L_{\Theta,\mathcal{Z}} \left| 1 - \frac{S_\tau^\theta}{S_{\tau,N}^\theta} \right| + \left\| \int_{\mathcal{C}} \nabla_\theta \mathcal{L}(\theta, z') \left(\mathrm{d}\mu_\tau^\theta(z') \frac{S_\tau^\theta}{S_{\tau,N}^\theta} - \mathrm{d}\mu_{\tau,N}^\theta(z') \right) \right\|. \tag{38}$$

For the first term, since $S_\tau^\theta = \mathbb{E}[S_{\tau,N}^\theta]$, we get from the proof of Lemma A.1 that for any $t > 0$,

$$\left| 1 - \frac{S_\tau^\theta}{S_{\tau,N}^\theta} \right| \leq e^{\frac{\overline{\mathcal{L}} - \underline{\mathcal{L}}}{\tau}} \sqrt{\frac{t \log N}{2N}}$$

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

Let us now turn to the second term in (38). Denote

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

We then have

$$\left\| \int_{\mathcal{C}} \nabla_\theta \mathcal{L}(\theta, z') \left(\mathrm{d}\mu_\tau^\theta(z') \frac{S_\tau^\theta}{S_{\tau,N}^\theta} - \mathrm{d}\mu_{\tau,N}^\theta(z') \right) \right\| = \frac{\left\| G_{\tau,N}^\theta - G_\tau^\theta \right\|}{S_{\tau,N}^\theta} \leq e^{\frac{-\underline{\mathcal{L}}}{\tau}} \left\| G_{\tau,N}^\theta - G_\tau^\theta \right\|.$$

Since $\mathbb{E}[G_{\tau,N}^\theta] = G_\tau^\theta$ and the random vectors $\nabla \mathcal{L}(\theta, z'_k) e^{\frac{\mathcal{L}(\theta, z'_k)}{\tau}}$ are independent and bounded, we apply Hoeffding's inequality and the union bound to obtain

$$\begin{aligned} \Pr\left(\left\| G_{\tau,N}^\theta - G_\tau^\theta \right\| > \epsilon \right) &\leq \Pr\left(\max_j |(G_{\tau,N}^\theta)_j - (G_\tau^\theta)_j| > \epsilon/\sqrt{p} \right) \\ &\leq p \max_j \Pr\left(|(G_{\tau,N}^\theta)_j - (G_\tau^\theta)_j| > \epsilon/\sqrt{p} \right). \end{aligned}$$

Taking $\epsilon = L_{\Theta,\mathcal{Z}} e^{\overline{\mathcal{L}}/\tau} \sqrt{\frac{2tp \log N}{N}}$, we infer that

$$\left\| \int_{\mathcal{C}} \nabla_\theta \mathcal{L}(\theta, z') \left(\mathrm{d}\mu_\tau^\theta(z') \frac{S_\tau^\theta}{S_{\tau,N}^\theta} - \mathrm{d}\mu_{\tau,N}^\theta(z') \right) \right\| \leq L_{\Theta,\mathcal{Z}} e^{\frac{\overline{\mathcal{L}} - \underline{\mathcal{L}}}{\tau}} \sqrt{\frac{2tp \log N}{N}}$$

with probability larger than $1 - 2pN^{-t}$. Combining the above bounds with the union bound, we get the claim.

(ii) Let $\epsilon_N := \max(1, 2L_{\Theta,\mathcal{Z}}\sqrt{p}) e^{\frac{\overline{\mathcal{L}} - \underline{\mathcal{L}}}{\tau_N}} \sqrt{\frac{\log N}{N}}$.

(a) We argue as in the proof of Lemma A.1(ii)(a). We have from claim (i) that

$$\Pr\left(\|\nabla g_{\tau_N}(\theta) - \nabla g_{\tau_N,N}(\theta)\| > \epsilon_N\right) \leq 2(p+1)N^{-2}.$$

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

$$\limsup_{N\to+\infty} \|\nabla g_{\tau_N}(\theta) - \nabla g_{\tau,N}(\theta)\| = 0.$$

(b) We will follow a reasoning similar to the proof of Lemma A.1(ii)(b) using separability of $\mathbb{R}^p$ and a density argument. There exists a countable set T whose closure is Ξ. According to claim (a), for every $\theta \in \Xi \subset \Theta$, there exists a set of events Ω_θ of probability one and, for every $\omega \in \Omega_\theta$, $\nabla g_{\tau_N,N}(\theta,\omega) - \nabla g_{\tau_N}(\theta,\omega) \to 0$. Set $\widetilde{\Omega} = \bigcap_{\theta\in\mathsf{T}} \Omega_\theta$. Countability of T shows that $\widetilde{\Omega}$ is also of probability one. Moreover, for fixed $\theta \in \Xi$, there exists a sequence $(\theta_k)_{k\in\mathbb{N}}$ in T such that $\theta_k \to \theta$. Let $\omega \in \widetilde{\Omega}$. We have

$$\begin{aligned}\|\nabla g_{\tau_N,N}(\theta,\omega) - \nabla g_{\tau_N}(\theta)\| \leq \|\nabla g_{\tau_N}(\theta_k) - \nabla g_{\tau_N}(\theta)\| + \|\nabla g_{\tau_N,N}(\theta_k,\omega) - \nabla g_{\tau_N}(\theta_k)\| \\ + \|\nabla g_{\tau_N,N}(\theta_k,\omega) - \nabla g_{\tau_N,N}(\theta,\omega)\|. \quad (39)\end{aligned}$$

Since $\theta_k \in \mathsf{T}$, $\widetilde{\Omega} \subset \Omega_{\theta_k}$, claim (a) gives us that the second term in the last inequality vanishes as $N \to +\infty$.

Let us turn to the first term. We have

$$\nabla g_{\tau_N}(\theta_k) - \nabla g_{\tau_N}(\theta) = \int_{\mathcal{C}} \nabla\mathcal{L}(\theta,z')\mathrm{d}\mu^{\theta_k}_{\tau_N}(z') - \int_{\mathcal{C}} \nabla\mathcal{L}(\theta,z')\mathrm{d}\mu^{\theta}_{\tau_N}(z').$$

By Lemma 4.9(i), each weak–$*$ cluster point of $(\mu^{\theta_k}_{\tau_N})_{N\in\mathbb{N}}$ belongs to $\operatorname{Argmin}\mathcal{L}(\theta_k,\mathcal{Z})$. Since the latter reduces to a single element $\widehat{z}^{\theta_k}$ by uniqueness of the maximizer, we get by Prokhorov's theorem that $(\mu^{\theta_k}_{\tau_N})_{N\in\mathbb{N}}$ converges in the weak–$*$ topology to the Dirac measure supported on $\widehat{z}^{\theta_k}$. Similarly, $(\mu^{\theta}_{\tau_N})_{N\in\mathbb{N}}$ $\widehat{z}^{\theta}$) converges in the weak–$*$ topology to the Dirac measure supported on the unique maximizer $\widehat{z}^{\theta}$ of $\operatorname{Argmin}\mathcal{L}(\theta,\mathcal{Z})$. Consequently,

$$\nabla g_{\tau_N}(\theta_k) - \nabla g_{\tau_N}(\theta) \xrightarrow[N\to+\infty]{} \nabla_\theta\mathcal{L}(\theta_k,\widehat{z}^{\theta_k}) - \nabla_\theta\mathcal{L}(\theta,\widehat{z}^{\theta}). \quad (40)$$

Now, observe that by (H.5) and convexity of Ξ, the mean value theorem yields

$$\sup_{z'\in\mathcal{C}} |\mathcal{L}(\theta_k,z') - \mathcal{L}(\theta,z')| \leq \sup_{z'\in\mathcal{C},\xi\in\Xi} \left\|\nabla_\theta\mathcal{L}(\xi,z')\right\| \|\theta_k - \theta\| \leq L_{\Theta,\mathcal{Z}} \|\theta_k - \theta\|,$$

and taking the limit as $k \to +\infty$, we see that $\mathcal{L}(\theta_k,\cdot)$ converges uniformly to $\mathcal{L}(\theta,\cdot)$, and thus by [23, Proposition 5.2 and Remark 5.3] $\mathcal{L}(\theta_k,\cdot)$ Γ-converges to $\mathcal{L}(\theta,\cdot)$. Compactness of $\mathcal{C}$ also entails equi-coercivity of $-\mathcal{L}(\theta_k,\cdot)$ on $\mathcal{C}$. This together with Γ-convergence of $\mathcal{L}(\theta_k,\cdot)$ seen just above allows to apply [23, Corollary 7.20] to infer that $\widehat{z}^{\theta_k} \to \widehat{z}^{\theta}$ as $k \to +\infty$. In view of this, taking the limit as $k \to +\infty$ in (40), using continuity of $\nabla_\theta\mathcal{L}$ in both arguments (see (H.5)), we get that

$$\lim_{k\to+\infty} \lim_{N\to+\infty} \|\nabla g_{\tau_N}(\theta_k) - \nabla g_{\tau_N}(\theta)\| = 0.$$

A similar reasoning can be applied to arrive at the same conclusion for the third term in (39). Passing to the limit in N and then in k in (39), we have proved that for every $\omega \in \widetilde{\Omega}$

$$\lim_{N\to+\infty} \|\nabla g_{\tau_N,N}(\theta,\omega) - \nabla g_{\tau_N}(\theta)\| = 0, \quad \text{for all} \quad \theta \in \Theta.$$

This completes the proof.

□

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. Thus since φ is definable, so is $\mathcal{C}^\varepsilon$. This together with definability of $\mathcal{L}$ implies 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

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

where we used that the functions g_i are path differentiable for the Clarke subdifferential by [37, Theorem 5.8]. Therefore, G is a Lyapunov function for the set crit–G. Moreover, by [39, Theorem 6], $G(\text{crit–}G)$ has empty interior.

By the almost sure boundedness assumption, $\max_i \left\|\partial^C g_i(\theta_k)\right\|$ is also uniformly bounded almost surely. 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 uniformly bounded martingale difference noise. These uniform boundedness properties and the choice of the sequence γ_k allows to apply [42, Remark 1.5(ii) and Proposition 1.4] to get by [42, 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 (27). Combining this with [42, Theorem 3.6 and Proposition 3.27] gives the claimed results. □

A.8 Proof of Theorem 5.3

We obviously have $v_k \in \frac{1}{M}\sum_{i=1}^M \partial^C g_i(\theta_k)$. Moreover, the bias term e_k obeys

$$\|e_k\| \leq \frac{1}{|B_k|}\sum_{i\in B_k} \operatorname{dist}\left(\nabla g^i_{\tau_k,N_k}(\theta_k), \partial^C g_i(\theta_k)\right) \leq \max_{\theta\in\mathbb{B}_C(0)} \operatorname{dist}\left(\nabla g^i_{\tau_k,N_k}(\theta), \partial^C g_i(\theta)\right). \tag{41}$$

In view of Theorem 4.8(ii)(b), if the sequence $(\tau_k, N_k)_{k\in\mathbb{N}}$ is as devised, then almost surely, $\lim_{k\to+\infty}\|e_k\| = 0$ for all $(\theta_k)_{k\in\mathbb{N}} \subset \mathbb{B}_C(0)$. By independent and uniform sampling of the mini-batches, ζ_k is a zero-mean martingale difference noise. We are then in position to invoke [59, Remark 4.5] to get that the conclusions of [42, Remark 1.5(ii) and Proposition 1.4] still hold provided that γ_k decays as devised. The rest of the proof is then same as that of Theorem 5.2. □

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 model trained is a Multi layer Perceptron 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	1500	Number of epochs
Optimizer	SGD-type	SGD for vanilla training, Algorithm 2 for robust training
Learning rate	0.01	Initial learning rate
Learning rate decay	0.1 every 300 epochs	Multiplicative decay for learning rate
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 [60]	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 sampling	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
Temperature	0.0001	Fixed temperature for LSE computation

Table 1: Parameters for the trainings on Avila dataset