



**HAL**  
open science

# Regularized Robust Optimization with Application to Robust Learning

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

► **To cite this version:**

William Piat, Jalal Fadili, Frédéric Jurie, Sébastien da Veiga. Regularized Robust Optimization with Application to Robust Learning. 2022. hal-03689825v1

**HAL Id: hal-03689825**

**<https://hal.science/hal-03689825v1>**

Preprint submitted on 7 Jun 2022 (v1), last revised 1 Feb 2023 (v4)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# Regularized Robust Optimization with Application to Robust Learning

William Piat\*      Jalal Fadili<sup>†</sup>      Frédéric Jurie<sup>‡</sup>      Sébastien Da Veiga<sup>§</sup>

June 7, 2022

## Abstract

In this paper, we propose a computationally tractable and provably convergent algorithm for robust optimization, with application to robust learning. First, the distributional 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 parameters and the number of integration samples decay 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.

## 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. As 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](mailto:william.piat@safrangroup.com).

<sup>†</sup>Normandie Univ, ENSICAEN, CNRS, GREYC, Caen, France, [Jalal.Fadili@greyc.ensicaen.fr](mailto:Jalal.Fadili@greyc.ensicaen.fr).

<sup>‡</sup>Normandie Univ, ENSICAEN, CNRS, GREYC, Caen, France, [frederic.jurie@greyc.ensicaen.fr](mailto:frederic.jurie@greyc.ensicaen.fr).

<sup>§</sup>Safran Tech, Digital Sciences and Technologies Department, Rue des Jeunes Bois, Châteaufort, 78114 Magny-Les-Hameaux, France, [sebastien.da-veiga@safrangroup.com](mailto:sebastien.da-veiga@safrangroup.com).

## 1.1 Problem statement

Let  $(\mathcal{Z}, d)$  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  $\rho_0$  be a probability measure on  $\mathcal{Z}$ ,  $\Theta \subset \mathbb{R}^p$  the parameter/action space,  $\mathcal{L} : \Theta \times \mathcal{Z} \rightarrow \mathbb{R}_+$  a loss function such that  $\mathcal{L}(\cdot, \theta)$  is  $\rho_0$ -measurable for all  $\theta \in \Theta$ . Consider the optimization problem

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

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  $\theta$ ), whose goal is to minimize an objective function under action constraints  $\Theta$ , 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

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

where  $\varepsilon > 0$  is the size of uncertainty/perturbation/attack, and the perturbation acts pointwise on  $z$  in the adversarial risk, hence the terminology Pointwise Robust Optimization (PRO) for problem (2). A common choice is  $d(z, z') = \|z - z'\|_q$  where  $\|\cdot\|_q$  is the  $\ell_q$  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, existence and pervasiveness of adversarial examples point to the limitations of the usual independent and identically distributed 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  $\rho_0$ . This approach is known as Distributional Robust Optimization (DRO) which takes the form

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

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  $\phi$ -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  $\rho_0$  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  $\rho$ . 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  $(\theta, z)$  and smoothness which are key to design efficient and provably convergent algorithms [1]. These assumptions are however very 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  $\Gamma$ -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 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  $D$  an  $\phi$ -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  $\rho_0$  is the empirical measure and  $\mathcal{L}$  is Lipschitz continuous in  $z$  uniformly in  $\theta$  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 applied 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  $\ell_q$  norm,  $\mathbb{B}_r^q(x)$  is the  $\ell_q$  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  $\{1, \dots, N\}$ .  $\text{vol}(\cdot)$  is the volume of a set.  $\text{dist}(x, \mathcal{C}) = \inf_{z \in \mathcal{C}} \|x - z\|$  is the distance function to the nonempty set  $\mathcal{C}$ . The set of nearest points of  $x$  in  $\mathcal{C}$  are denoted by  $P_{\mathcal{C}}(x)$ .  $\mathcal{C}^s$  is the class of  $s$ -continuously differentiable functions and  $\mathcal{C}$  is the space of continuous functions.

**Probability measures** For a subset  $\mathcal{C} \subset \mathbb{R}^m$ , let  $\mathcal{B}$  be the Borel sigma 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  $\mathcal{C}$

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

$\delta_x$  is the Dirac measure at  $x$ .

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

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

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

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

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

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

**$\Gamma$ - or epi-convergence** We will invoke the notion of  $\Gamma$ -convergence, which plays a fundamental role in convergence of optimization problems (values and extrema points). In finite dimension,  $\Gamma$ -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 semi-algebraic 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  $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$  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  $t$ ,  $f_t : \mathbb{R}^n \rightarrow \mathbb{R}$  is locally Lipschitz continuous. Let  $f(x) = \max_{t \in T} f_t(x)$ . Let  $S$  be a subset of full Lebesgue measure. Then*

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

*If moreover, the functions  $f_t$  are of class  $\mathcal{C}^1$  such that  $f_t(x)$  and  $\nabla f_t(x)$  depend continuously on  $(t, x)$ <sup>1</sup>, then*

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

**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.*

<sup>1</sup>Functions  $f$  such that these assumptions are verified are known as lower- $\mathcal{C}^1$  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  $\varepsilon$  is not too large.

**Proposition 3.1.** *Assume that (H.1)-(H.2) hold, and that*

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

Then

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

See Appendix A.1 for the proof.

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

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

One recognizes a penalized form of (2) in the infimum problem over the Lagrange multiplier  $\gamma$ . While this is an identity rather than a bound, one has to optimize over the multiplier  $\gamma$  which is far from obvious. In [21, 9], taking  $c$  as the  $\ell_p$  cost, and  $\rho_0$  the empirical measure on  $n$  points, the following bound was established

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

As in our case, this gives access to a uniform bound, but which depends now on  $n$  rather than  $\varepsilon$ . However, the price to pay is that the inner maximization in the surrogate problem is even less tractable since it does not have the finite sum form along with the constraint/optimization problem being on a variable in  $\mathbb{R}^{mn}$  rather than  $\mathbb{R}^m$ .

### 4 Entropic regularization

In all the surrogate formulations above, the inner maximization problem remains a very challenging task unless stringent joint convexity/concavity assumptions are made. This is the motivation behind our smoothing hereafter.

#### 4.1 Regularized objective

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

$$\min_{\theta \in \Theta} \mathbb{E}_{z \sim \rho_0} \left\{ \max_{z' \in \mathcal{C}_z^\varepsilon} \mathcal{L}(\theta, z') \right\}. \quad (7)$$

We will use the shorthand notation<sup>2</sup>

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

Provided that  $\mathcal{C}_z^\varepsilon$  is bounded, hence compact since it is closed by assumption on  $\mathfrak{c}$ , 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

$$g(\theta) = \max_{\mu \in \mathcal{P}(\mathcal{C}_z^\varepsilon)} \int_{\mathcal{C}_z^\varepsilon} \mathcal{L}(\theta, z') d\mu(z'), \quad (8)$$

and the integral is a duality pairing between  $\mathcal{P}(\mathcal{C}_z^\varepsilon)$  and  $\mathcal{C}(\mathcal{C}_z^\varepsilon)$ . Problem (8) is concave in  $\mu$ , but operates in infinite-dimension and is thus hard to solve. Approximating (8) 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 a two-step strategy, first starting with the following regularized version of (8)

$$g_\tau(\theta) := \max_{\mu \in \mathcal{P}(\mathcal{C}_z^\varepsilon)} \int_{\mathcal{C}_z^\varepsilon} \mathcal{L}(\theta, z') d\mu(z') - \tau \text{KL}(\mu, \nu), \quad (9)$$

where  $\tau > 0$  is the regularization parameter, and  $\nu$  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].

In the sequel we suppose that  $\mathcal{C}_z^\varepsilon$  is of full dimension, and set  $\nu$  as the uniform measure  $\mu_u$  on  $\mathcal{C}_z^\varepsilon$ , that is  $\nu(z') = \mu_u(z') := \text{vol}(\mathcal{C}_z^\varepsilon)^{-1}$  for all  $z' \in \mathcal{C}_z^\varepsilon$ . The KL regularization term then prevents solutions to be sparse measures as such measures are not absolutely continuous with respect to the uniform one.

Remarkably, (9) is well-posed under mild conditions and has a unique solution taking an explicit form.

**Proposition 4.1.** *Assume that (H.1)-(H.2) hold and that  $\mathcal{C}_z^\varepsilon$  is bounded and full dimensional. Then (9) has a unique solution and*

$$g_\tau(\theta) = \tau \log \left( \mathbb{E}_{z' \sim \mu_u} \left[ \exp \left( \frac{\mathcal{L}(\theta, z')}{\tau} \right) \right] \right) = \tau \log \left( \frac{\int_{\mathcal{C}_z^\varepsilon} \exp \left( \frac{\mathcal{L}(\theta, z')}{\tau} \right) dz'}{\text{vol}(\mathcal{C}_z^\varepsilon)} \right). \quad (10)$$

The proof can be found in Appendix A.2. Note that this is an extension to infinite dimension of the standard log-sum-exp formula for softmax smoothing of the maximum of a finite number of functions.

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

## 4.2 Consistency of the regularization

We now turn to describing how  $g_\tau$  in (9) is a good surrogate for  $g$  in (7). We provide both a qualitative result as the regularization parameter  $\tau$  vanishes, and a quantitative convergence rate.

---

<sup>2</sup> $g$  depends on  $z$  but we drop this in the notation to lighten the latter.



**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,  $g_\tau$   $\Gamma$ -converges to  $g$  as  $\tau \searrow 0^+$ .
- (ii) *If, moreover,  $C_z^\varepsilon$  is convex and full dimensional, and (H.3) is reinforced to Lipschitz continuity w.r.t. some norm  $\|\cdot\|$  on  $\mathbb{R}^m$ , then for any  $\tau \in ]0, 1]$*

$$g(\theta) - h(\tau) \leq g_\tau(\theta) \leq g(\theta), \quad (11)$$

where

$$h(\tau) = m\tau \log(\tau^{-1}) + \tau \log \left( \frac{\text{vol}(C_z^\varepsilon)}{\text{vol}(\mathbb{B}^{R_{C_z^\varepsilon}}(0))} \right) + \tau L_{\mathcal{Z}}(R_{C_z^\varepsilon} + D_{C_z^\varepsilon}),$$

and  $R_{C_z^\varepsilon}$  is the radius of the largest ball<sup>3</sup> contained in  $C_z^\varepsilon$  and  $D_{C_z^\varepsilon}$  is the diameter of  $C_z^\varepsilon$ .

See Appendix A.3 for the proof.

Let us pause to make some remarks.

**Remark 4.4.**

1. *The convexity of  $C_z^\varepsilon$  is again mild and verified in all applications we have in mind. It holds if  $c$  is convex, e.g., a convex metric on  $\mathcal{Z}$ .*
2. *The bound (11) yields uniform convergence. The  $\Gamma$ -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  $\tau$  is nearly linear (up to a logarithmic term).*
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  $m\tau \log(\bar{R}_{C_z^\varepsilon}/R_{C_z^\varepsilon})$  where  $\bar{R}_{C_z^\varepsilon}$  is the radius of the smallest ball containing  $C_z^\varepsilon$ . For the last term, the Lipschitz constant  $L_{\mathcal{Z}}$  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  $\Theta$  is compact. Let  $\theta_\tau^* \in \text{Argmin}_\Theta(g_\tau)$ . Then the following holds:*

- (i)  $\lim_{\tau \rightarrow 0^+} \min_{\theta \in \Theta} g_\tau(\theta) = \min_{\theta \in \Theta} g(\theta)$ .
- (ii) *Each cluster point of  $\theta_\tau^*$ , as  $\tau \rightarrow 0^+$ , lies in  $\text{Argmin}_\Theta(g)$ .*
- (iii) *In particular, if  $\text{Argmin}_\Theta(g) = \{\theta^*\}$ , then  $\lim_{\tau \rightarrow 0^+} \theta_\tau^* = \theta^*$ .*

See Appendix A.4 for the proof.

---

<sup>3</sup>In the norm  $\|\cdot\|$  of course.

### 4.3 Monte Carlo approximation of the integral

Computing the values  $g_\tau(\theta)$  in (10) 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

$$g_{\tau,N}(\theta) := \tau \log \left( \sum_{k=1}^N \frac{\exp\left(\frac{\mathcal{L}(\theta, z'_k)}{\tau}\right)}{N} \right). \quad (12)$$

We now provide an error bound for such an approximation.

**Theorem 4.6.** *Suppose that the assumptions of Theorem 4.3(ii) hold. Then, for any  $t > 0$ ,*

$$|g_{\tau,N}(\theta) - g(\theta)| \leq h(\tau) + \tau e^{\frac{\bar{\mathcal{L}} - \underline{\mathcal{L}}}{\tau}} \sqrt{\frac{t \log N}{N}}, \quad (13)$$

*with probability at least  $1 - 2N^{-t}$ , where  $\underline{\mathcal{L}} := \inf \mathcal{L}(\Theta, \mathcal{C}_z^\varepsilon)$  and  $\bar{\mathcal{L}} := \sup \mathcal{L}(\Theta, \mathcal{C}_z^\varepsilon)$ . In turn, (13) holds almost surely for any  $t > 1$ .*

The proof can be found in Appendix A.5.

Since  $\mathcal{L}$  is continuous by (H.1) and  $\mathcal{C}_z^\varepsilon$  is compact,  $\underline{\mathcal{L}}$  and  $\bar{\mathcal{L}}$  are well-defined as minimal and maximal values as soon as  $\Theta$  is compact. For fixed  $\tau$ , the convergence rate 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 (13) 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

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

The error bound (13) reveals an exponential dependence in  $\tau$ , and thus for the right-hand side to vanish as  $\tau \rightarrow 0^+$  and  $N \rightarrow +\infty$ , there is trade-off between  $N$  and  $\tau$  to enforce the term  $e^{\frac{\bar{\mathcal{L}} - \underline{\mathcal{L}}}{\tau}} \sqrt{\frac{t \log N}{N}}$  to converge to 0. Taking  $\tau = O\left(\frac{1}{\kappa \log N}\right)$ , for any  $\kappa > 0$  such that  $\kappa(\bar{\mathcal{L}} - \underline{\mathcal{L}}) < 1/2$ , the convergence rate in (13) is dominated by the first term  $h(\tau)$  which scales as  $O((\log N)^{-1})$ . This is obviously a slow convergence rate but reflects the difficulty of approximating the function  $g$  in (7).

#### 4.4 Consistency of subgradient estimates

Equipped with the above results, a natural strategy now is to solve the PRO problem (7) by using  $g_{\tau,N}$  in (12) 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

$$\nabla g_{\tau,N}(\theta) = \sum_{k=1}^N \nabla_{\theta} \mathcal{L}(\theta, z'_k) \frac{\exp\left(\frac{\mathcal{L}(\theta, z'_k)}{\tau}\right)}{\sum_{j=1}^N \exp\left(\frac{\mathcal{L}(\theta, z'_j)}{\tau}\right)}. \quad (14)$$

The natural question that arises is whether (14) behaves well as  $\tau$  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 4.7 shows that this is indeed the case under stronger assumptions than those required for consistency of the (zero-th order oracle) function values established above.

To show our result, we will need some regularity properties on the set of maximizers  $\mathcal{M} := \text{Argmax } \mathcal{L}(\theta, \mathcal{C})$  (we drop the dependence of  $\mathcal{M}$  on  $\theta$  to lighten notation). We say that a set  $\mathcal{S} \subset \mathbb{R}^m$  is  $\mathcal{C}^r$ -stratifiable, for some integer  $r \geq 1$ , if there is a finite partition of  $\mathcal{S}$  into disjoint  $\mathcal{C}^r$  submanifolds  $(\mathcal{M}_i)_{i \in I}$  of  $\mathbb{R}^m$ , called strata, with  $m \geq \dim(\mathcal{M}_1) > \dim(\mathcal{M}_2) > \dots > \dim(\mathcal{M}_{|I|}) \geq 0$ .

**Theorem 4.7.** *In addition to the assumptions of Proposition 4.1, suppose that*

(H.4)  $\mathcal{L}(\cdot, z')$  is differentiable with  $\nabla_{\theta} \mathcal{L}(\theta, z')$  continuous in  $(\theta, z')$  and uniformly bounded by  $L_{\Theta, \mathcal{Z}}$  on  $\Theta \times \mathcal{Z}$ ;

(H.5)  $\mathcal{L}(\theta, \cdot)$  is  $\mathcal{C}^3$  on an open set containing  $\mathcal{C}_z^{\varepsilon}$  with Hölder continuous third-order derivative;

(H.6) for  $r \geq 3$

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

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

Then  $\mathcal{L}$  and  $g_{\tau,N}$  are continuously differentiable on  $\mathcal{C}_z^{\varepsilon}$ , and for any  $t > 0$ , with probability at least  $1 - 2(p+1)N^{-t}$

$$\text{dist}(\nabla g_{\tau,N}(\theta), \partial^{\mathcal{C}} g(\theta)) \leq L_{\Theta, \mathcal{Z}} e^{2\frac{\bar{\mathcal{L}} - \underline{\mathcal{L}}}{\tau}} \sqrt{\frac{2tp \log N}{N}} + o_{\tau}(1). \quad (15)$$

where

$$\partial^{\mathcal{C}} g(\theta) = \left\{ \int_{\mathcal{C}_z^{\varepsilon}} \nabla_{\theta} \mathcal{L}(\theta, z') d\mu(z') : \mu \in \mathcal{P}\left(\text{Argmax } \mathcal{L}(\theta, \mathcal{C}_z^{\varepsilon})\right) \right\}. \quad (16)$$

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 4.8.** *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  $\Theta \times \mathcal{Z}$ . Then for any  $t > 0$*

$$\|\nabla g_{\tau}(\theta) - \nabla g_{\tau,N}(\theta)\| \leq L_{\Theta, \mathcal{Z}} e^{2\frac{\bar{\mathcal{L}} - \underline{\mathcal{L}}}{\tau}} \sqrt{\frac{2tp \log N}{N}} \quad (17)$$

with probability at least  $1 - 2(p+1)N^{-t}$ . The bound holds almost surely for any  $t > 1$ .

**Lemma 4.9.** *Under the assumptions of Theorem 4.7, we have*

$$\text{dist}(\nabla g_\tau(\theta), \partial^C g(\theta)) \rightarrow 0 \quad \text{as } \tau \rightarrow 0^+. \quad (18)$$

To prove the last lemma, one observes that under our assumptions,

$$\nabla g_\tau(\theta) = \int_{\mathcal{C}_z^\varepsilon} \nabla_\theta \mathcal{L}(\theta, z') \frac{\exp\left(\frac{\mathcal{L}(\theta, z')}{\tau}\right)}{\int_{\mathcal{C}_z^\varepsilon} \exp\left(\frac{\mathcal{L}(\theta, v)}{\tau}\right) dv} dz', \quad (19)$$

which is an expectation with respect to a Gibbs measure. The proof will then amount to showing that this measure converges in the narrow topology (equivalent to the weak-\* topology) to a measure supported on  $\text{Argmax } \mathcal{L}(\theta, \mathcal{C}_z^\varepsilon)$ <sup>4</sup>, and then apply the rule (5).

Clearly, Theorem 4.7 tells us that  $\nabla g_{\tau, N}(\theta)$  is at most within a ball around the Clarke subdifferential of  $g$  at  $\theta$ . The result also quantifies its radius and how it vanishes with  $N$  and  $\tau$ , and shows the influence of  $p$ , the dimension of  $\Theta$ . Arguing as above, this radius vanishes as  $N \rightarrow +\infty$  by taking  $\tau = O\left(\frac{1}{\kappa \log N}\right)$ , for any  $\kappa$  such that  $2\kappa(\bar{\mathcal{L}} - \underline{\mathcal{L}}) \in ]0, 1/2 - \alpha]$ ,  $\alpha \in ]0, 1/2[$ . The convergence rate of the first term (15) is then nearly  $O(N^{-\alpha})$  (up to logarithmic factors). However, we do not have any quantitative estimate for the rate of the second term.

**Remark 4.10.** *The requirement in Theorem 4.7 that the strata  $(\mathcal{M}_i)_{i \in I}$  of  $\mathcal{M}$  are smooth submanifolds is not stringent, and can be ensured for instance if the function  $\mathcal{L}$  and the cost  $c$  are semialgebraic (and more generally tame functions [26]). However closedness of the submanifolds is more demanding and it appears challenging at this stage to remove such an assumption. One of the key reasons we need closedness is to ensure that each stratum is a set with positive reach; see [36] and [37] for the closely related concept of proximal smoothness. These are sets where the corresponding projector mapping is single-valued and Lipschitz continuous on an open tubular neighbourhood of uniform thickness around the set.*

**Remark 4.11.** *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 [38, 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 (34), and then use that their dimensions are strictly smaller than that of the largest submanifold.*

## 5 Robust optimization algorithm via SGD

We are now ready to describe our algorithmic framework to solve the PRO problem (7) 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  $\Theta$  is a convex closed set by including a projection step onto  $\Theta$  (see e.g., [39, 40] in different settings). Moreover, without loss of generality, we take in this section

$$c(z, z') = \varphi(z - z'),$$

<sup>4</sup>This is reminiscent of works on simulated annealing where  $\tau$  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..

where  $\varphi : \mathbb{R}^m \rightarrow \mathbb{R}_+$  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  $\rho_0$  in (7) is the empirical measure on  $\mathcal{Z}$ , hence leading to the finite sum minimization problem

$$\min_{\theta \in \Theta} \left\{ G(\theta) := \frac{1}{M} \sum_{i=1}^M g_i(\theta) \right\} \quad \text{where} \quad g_i(\theta) := \max_{u \in \mathcal{C}^\varepsilon} \mathcal{L}(\theta, z_i + u), \quad (20)$$

where  $\mathcal{C}^\varepsilon := \{u \in \mathbb{R}^m : \varphi(u) \leq \varepsilon\}$  is obviously a nonempty compact and full dimensional set by the assumptions on  $\varphi$ . It is worth observing that convexity of  $\varphi$  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),  $\mathcal{L}(\cdot, z_i + u)$  is assumed locally Lipschitz continuous for each  $(z_i, u)$ , (4) gives us the inclusion

$$\partial^C G(\theta) = \partial^C \left( \frac{1}{M} \sum_{i=1}^M g_i(\theta) \right) \subset \frac{1}{M} \sum_{i=1}^M \partial^C g_i(\theta) \subset \frac{1}{M} \sum_{i=1}^M \mathcal{D}_i(\theta), \quad (21)$$

where

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

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 [41, 42]. These fields enjoy nice sum and chain rules and coincide with the Clarke subdifferential almost everywhere. However, we are not aware of any calcul rule of such fields under pointwise maximization. We are then naturally led to consider the set of critical points

$$\text{crit-}G := \left( \frac{1}{M} \sum_{i=1}^M \mathcal{D}_i \right)^{-1} (0). \quad (22)$$

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

$$\theta_{k+1} = \theta_k - \gamma_k d_k, \quad \text{where} \quad d_k \in \frac{1}{|B_k|} \sum_{i \in B_k} \mathcal{D}_i(\theta_k), \quad (23)$$

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 (23) enjoys some convergence guarantees to the set of critical points in  $\text{crit-}G$ . For this, we will rely on the stochastic approximation method for differential inclusions with compact and convex-valued operators developed in [43], and used recently in [41, 44]. 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

$$\begin{cases} 0 \in \dot{\theta}(t) + \frac{1}{M} \sum_{i=1}^M \mathcal{D}_i(\theta(t)) & \text{for almost every } t \in \mathbb{R}, \\ \theta(0) = \theta_0, \end{cases} \quad (24)$$

whose stationary solutions are the critical points in (22). A key argument to invoke the results of [43], is to build an appropriate Lyapunov function and show that the function  $G$  is path differentiable, that is, it obeys the chain rule

$$\frac{d}{dt}G(\theta(t)) = \langle \dot{\theta}(t), v \rangle, \quad \forall v \in \frac{1}{M} \sum_{i=1}^M \mathcal{D}_i(\theta(t)). \quad (25)$$

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.4), in which case (16) applies and (21) becomes

$$\partial^C G(\theta) = \partial^C \left( \frac{1}{M} \sum_{i=1}^M g_i(\theta) \right) \subset \frac{1}{M} \sum_{i=1}^M \partial^C g_i(\theta), \quad (26)$$

where

$$\partial^C g_i(\theta) = \left\{ \int_{\mathcal{C}^\varepsilon} \nabla_\theta \mathcal{L}(\theta, z_i + u) d\mu(u) : \mu \in \mathcal{P} \left( \text{Argmax}_{\mathcal{C}^\varepsilon} \mathcal{L}(\theta, z_i + \cdot) \right) \right\}. \quad (27)$$

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  $\theta_0$ ;

**for**  $k = 0, \dots$  **do**

Draw independently uniformly at random a mini-batch  $B_k \subset [M]$  ;

**for**  $i \in B_k$  **do**

└ Solve  $\bar{u}_i \in \text{Argmax}_{u \in \mathcal{C}^\varepsilon} \mathcal{L}(\theta_k, z_i + u)$ .

$d_k = \frac{1}{|B_k|} \sum_{i \in B_k} \nabla_\theta \mathcal{L}(\theta_k, z_i + \bar{u}_i)$  ;

$\theta_{k+1} = \theta_k - \gamma_k d_k$ .

---

This algorithm enjoys the the following guarantees.

**Theorem 5.1.** *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  $(\theta, z)$ . Suppose moreover that  $\varphi$  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  $\sup_{k \in \mathbb{N}} \|\theta_k\| = C < +\infty$  almost surely. Then, almost surely, the set of cluster points of  $(\theta_k)_{k \in \mathbb{N}}$  belong to  $\text{crit-}G = \left( \frac{1}{M} \sum_{i=1}^M \partial^C g_i \right)^{-1}(0)$ . Moreover  $(G(\theta_k))_{k \in \mathbb{N}}$  converges and  $G$  is constant on  $\text{crit-}G$ .*

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 (27). 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 4.7 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  $\theta_0$ ;  
**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  $\mathcal{C}^\varepsilon$  ;  
     $d_k = \frac{1}{|B_k|} \sum_{i \in B_k} \nabla g_{\tau_k, N_k}^i(\theta_k)$ , with  $\nabla g_{\tau_k, N_k}^i(\theta_k) := \sum_{j=1}^{N_k} \nabla_{\theta} \mathcal{L}(\theta, z_i + u_j) \frac{e^{-\frac{\mathcal{L}(\theta, z_i + u_j)}{\tau_k}}}{\sum_{l=1}^{N_k} e^{-\frac{\mathcal{L}(\theta, z_i + u_l)}{\tau_k}}}$  ;  
     $\theta_{k+1} = \theta_k - \gamma_k d_k$ .

---

The direction  $d_k$  in Algorithm 2 can also be written as

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

where

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

where all projectors above are well-defined since the Clarke subdifferential is closed (in fact compact) and convex-valued. We obviously have that  $v_k \in \frac{1}{M} \sum_{i=1}^M \partial^C g_i(\theta_k)$ . Moreover, in view of (15), the bias term  $e_k$  obeys, for any  $t > 1$ ,

$$\|e_k\| \leq \frac{1}{|B_k|} \sum_{i \in B_k} \text{dist}(\nabla g_{\tau_k, N_k}^i(\theta_k), \partial^C g_i(\theta_k)) \leq L_{\Theta} e^{2\frac{\bar{\mathcal{L}} - \underline{\mathcal{L}}}{\tau_k}} \sqrt{\frac{2tp \log N_k}{N_k}} + o_{\tau_k}(1) \quad \text{almost surely.} \quad (28)$$

By independent and uniform sampling of the mini-batches,  $\zeta_k$  is a zero-mean martingale difference noise. We then have the following convergence result.

**Theorem 5.2.** *Assume that the assumptions of Theorem 4.7 hold. Suppose moreover that  $\varphi$  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 the right hand side of (28) vanishes as  $k \rightarrow +\infty$ . Consider the sequence  $(\theta_k)_{k \in \mathbb{N}}$  generated by Algorithm 2 and suppose that  $\sup_{k \in \mathbb{N}} \|\theta_k\| = C < +\infty$  almost surely. Then, almost surely, the set of cluster points of  $(\theta_k)_{k \in \mathbb{N}}$  belong to  $\text{crit-}G = \left(\frac{1}{M} \sum_{i=1}^M \partial^C g_i\right)^{-1}(0)$ . Moreover  $(G(\theta_k))_{k \in \mathbb{N}}$  converges and  $G$  is constant on  $\text{crit-}G$ .*

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  $\tau_k$  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]. SGD is of course not the most efficient gradient descent algorithm and other algorithms with accelerated gradients can be used provided they are proven to converge with subgradients (see [45] for instance). It remains to be seen whether this training provides robustness on some use cases, indeed a convergence to a critical point does not necessarily mean that it would perform well given robust metrics.

## 6 Numerical results

### 6.1 Dataset, model and metrics

The following experiments were all performed on the Avila dataset, introduced in [46]. This dataset is representative of a classification task – writer identification in medieval manuscripts through page layout features – with  $q = 8$  input features and 12 classes. We centered and normalized the attributes in a preprocessing step.

As regard as the model used in these experiments, we build a 3 layer MLP with two hidden layers of 200 neurons each and an output layer, resulting in  $p = 44000$  parameters. We used the ELU activation function [47] and with the SGD algorithm respectively the only activation and optimizer that comply with our hypothesis. The loss used for training the model is the cross entropy loss. On a bounded set, it remains within our hypothesis but might augment the overall Lipschitz constant.

All the runs were made with the same number of epochs, batch size and therefore the same number of updates.

As regard to comparisons, we trained the same model in 3 different ways: with a vanilla training, an Adversarial training and a Robust training. Our aim is to show that we can train a robust model using algorithm 2, which is competitive with current state of the art procedures for robustifying neural networks such as adversarial training, for different values of the robustness radius. By considering the one-zero gain  $G(p, y)$  that give 1 if the prediction of the model  $p = F_\theta(x)$  is equal to the label  $y$  and 0 if they are different, we defined 3 different metrics:

**Test Accuracy** It is the accuracy of the models on a test base of size  $N_{test}$  of tuples  $(x_i, y_i)$ , defined as:

$$\text{Test Accuracy} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} G(F(x_i), y_i)$$

**Adversarial Accuracy** This is 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  $\mathbb{B}_\varepsilon^n$ ,  $n \in \mathbb{N}$  in which the attack is constrained, and a tuple  $(x_i, y_i)$  of datapoints to be attacked.



$$\text{Adversarial Accuracy} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} G(F(\hat{x}_i), y_i)$$

where  $\hat{x}_i = \text{PGD}(\mathcal{L}, \mathbb{B}_\varepsilon^q, x_i, y_i)$

**Robust Accuracy** The robust sampled accuracy is the worst accuracy after sampling random values around the data points. Enough points must be sampled to approach the worst case. We note  $\mu_u(x)$  the uniform measure on  $\mathbb{B}_\varepsilon^n$ ,  $n \in \mathbb{N}$  (same ball than for the PGD attack) and  $N$  the number of samples used for evaluating the robust loss:

$$\text{Robust Accuracy} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \min_{(x_{i,j})_{j=1}^N, x_{i,j} \sim \mu_u(x_i)} (G(F(x_{i,j}), y_i))$$

If  $\mathcal{C}_x^\varepsilon$  is sampled enough (i.e. for  $N$  big enough), the third metric is the most penalizing one, and, as such, should show if the PRO problem was solved during the training.

We added to these metrics the evaluation of the Lipschitz constant of the network. To this end, we chose the LipSDP method [48] which is efficient, accurate and adequate for the size and structure of the models used.

These metrics are to be evaluated for the three kinds of training, for different values of  $\varepsilon$ . We made 100 runs for each configuration with different initializations to have some statistical relevance. 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 charts present the median value and the quantile at 0.1, 0.25, 0.75 and 0.9.

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

Following Theorem 4.7, 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  $\tau$  and  $N$ .

The sampling for the robust training is performed in  $\mathbb{B}_{0,05}^\infty$  for various values of  $\tau$  and  $N$ . The inner maximisation 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 shows the value of the loss for different values of  $\tau$  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.

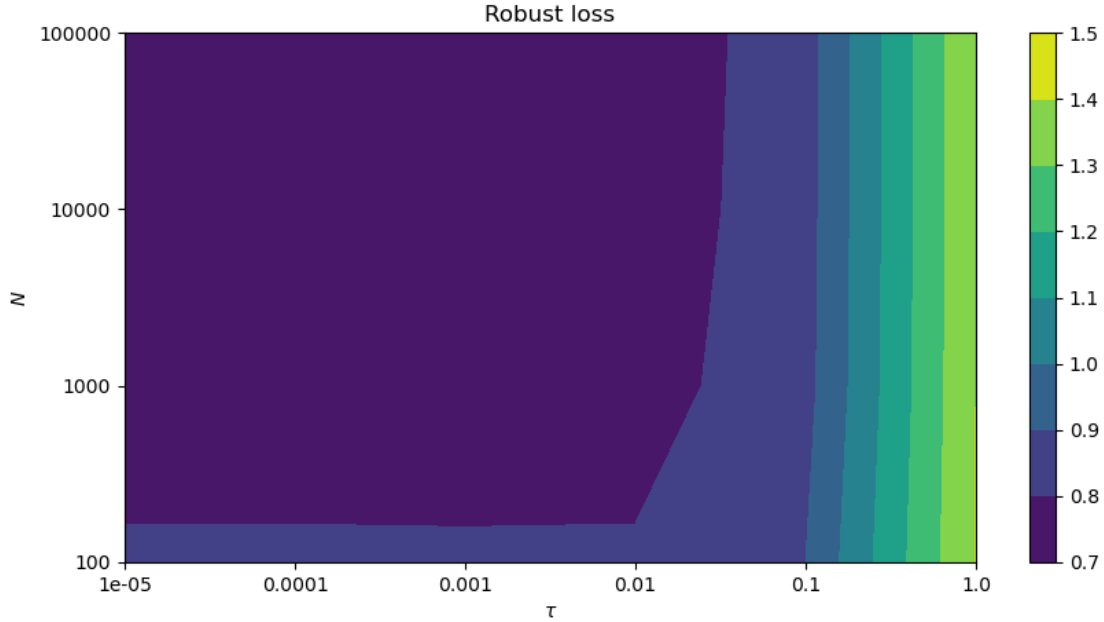


Figure 1: Influence of  $\tau$  (x axis) and  $N$  (y axis) on the Robust loss on test set (color scale). The darker the better.

### 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  $\mathbb{B}_\epsilon^\infty$  and so was the sampling for the robust training: we kept the same fixed sampling of  $N = 500000$  points and a fixed temperature of  $\tau = 0.0001$ . Fixing the sampling to the highest possible value insures the smallest error possible on the estimation of the robust loss. We give in Section B.1 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 (Figure 5) estimated using LipSDP [48].

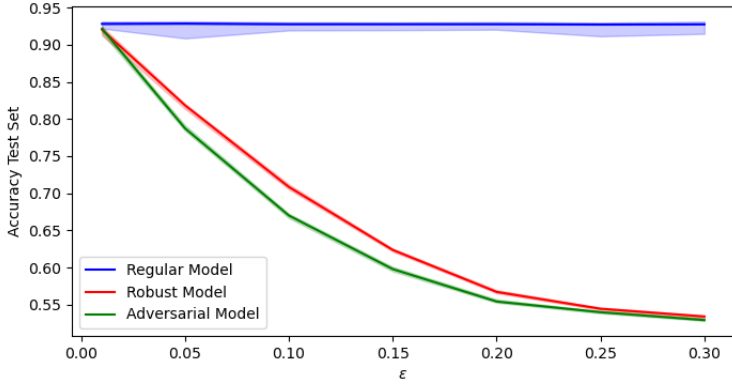


Figure 2: Accuracy of three trainings on Avila base: Vanilla (Blue), Adversarial (Green), Robust (Red) with median (plain line) and 10% and 90% quantiles for 50 trials

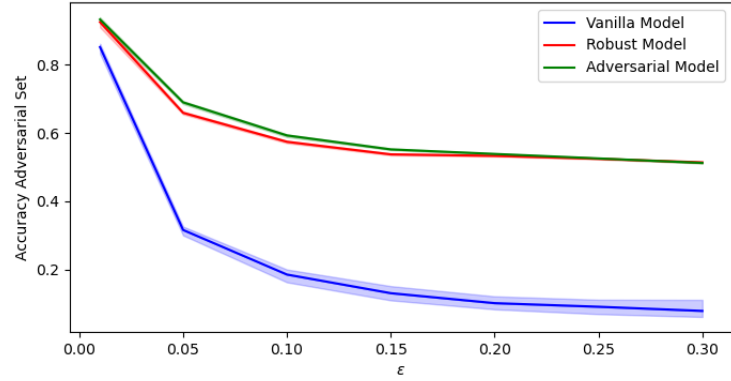


Figure 3: Adversarial Accuracy of three trainings on Avila base: Vanilla (Blue), Adversarial (Green), Robust (Red) with median (plain line) and 10% and 90% quantiles for 50 trials

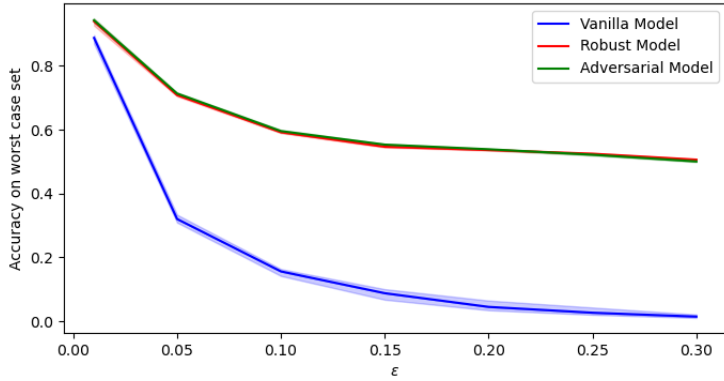


Figure 4: Worst-case Accuracy of three trainings on Avila base: Vanilla (Blue), Adversarial (Green), Robust (Red) with median (plain line) and 10% and 90% quantiles for 50 trials

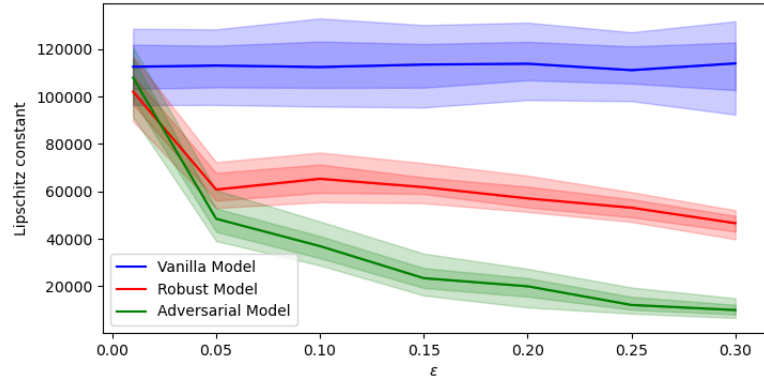


Figure 5: Lipschitz constant of three trainings on Avila base: Vanilla (Blue), Adversarial (Green), Robust (Red) with median (plain line) and 10%, 25%, 75% and 90% quantiles for 50 trials

For these experiments, the adversarial training was performed with  $\mathbb{B}_\varepsilon^\infty(0)$  and so was the sampling for the robust training. For the latter, we kept the same fixed sampling of  $N = 5.10^6$  points and a fixed temperature of  $\tau = 10^{-4}$  (see again Table 1 for more details about the experimental setting).

We plot the evolution of the test accuracy (Figure 2), adversarial accuracy (Figure 3) and worst case robust accuracy (Figure 4) versus the perturbation radius. On top of this we also present the evolution of the Lipschitz constant (Figure 5) estimated using LipSDP.

On all the above figures, the behaviour of our robust training is competitive compared to the popular PGD-based adversarial training, and practical differences see small. However we note that the robust training is

better on the test accuracy for moderate perturbations. We note, as expected, a decrease in accuracy on the test dataset but a better tolerance to attacks. This shows that the models were successfully robustified. The variance is small (except for the Vanilla training) suggesting that there is a good stability from run to run, for different initializations. Regarding Figure 5, it is notable how the adversarial training and robust training tend to reduce the Lipschitz constant. This results in smoother decision boundaries in both cases, although there is no explicit additional regularization during training. 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 base. Usually, adversarial attacks favor the corresponding models that have already being trained on an adversarial setting. On these examples we see that smoothing approximates the inner maximization problem but comes with guarantees unlike PGD-based adversarial training. Both approximations of the robust objective lead to smoother decision boundaries, hinting that it implicitly regularizes the network contrary to the Vanilla training. The sampling on the perturbation set is sufficient for approximating the PRO problem and we are closer to solving the DRO problem as the Lipschitz constant shrinks during the Robust training. The main drawback of the robust training is its computational cost compared to adversarial training, which remains a very effective method for solving empirically the PRO problem.

## 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 guaranties at the expense of an overall higher computational cost. Our robust training has given performances similar to those of adversarial training. 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 is one direction that is worth investigating in a future work.

## 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 Waegenare, 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  $\Gamma$ -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] H. Federer. Curvature measures. *Trans. Amer. Math. Soc.*, 93:418–491, 1959.
- [37] F. H. Clarke, R. J. Stern, and P. R. Wolenski. Proximal smoothness and the lower- $c^2$  property. *J. Convex Analysis*, 2:117–144, 1995.
- [38] J. M. Lee. *Introduction to smooth manifolds*. Springer, 2003.
- [39] 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.
- [40] H. J. Kushner and G. G. Yin. *Stochastic Approximation Algorithms and Applications*, volume 35 of *Applications of Mathematics*. Springer, 1997.
- [41] Jérôme Bolte and Edouard Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Math. Program.*, 188(1):19–51, 2021.
- [42] 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.
- [43] Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. Stochastic Approximations and Differential Inclusions. *SIAM J. Control. Optim.*, 44(1):328–348, 2005.
- [44] 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.

- [45] Andrzej Ruszczyński. Convergence of a stochastic subgradient method with averaging for nonsmooth nonconvex constrained optimization. *Optim. Lett.*, 14(7):1615–1625, 2020.
- [46] 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.
- [47] 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.
- [48] 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.
- [49] Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: a Hitchhiker’s Guide*. Springer, Berlin; London, 2006.
- [50] Herbert Federer. *Geometric Measure Theory*. Classics in Mathematics. Springer Berlin Heidelberg, 1996.
- [51] 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.
- [52] Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Séminaire de probabilités XXXIII*, volume 1709 of *Lecture Notes in Mathematics*, pages 1–68. Springer, 1999.
- [53] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and D. Mike Titterton, 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

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

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

Taking the expectation on both sides, we get

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

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 upper bound. We embark from (6) and use the uniform Lipschitz continuity assumption (H.3) to infer that

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

where we used that  $c(z, z') = 0$ . □

### 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  $c$  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 [49, 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  $\mu \mapsto \int_{\mathcal{C}_z^\varepsilon} \mathcal{L}(\theta, z') d\mu(z')$  is weak-\* continuous. It is known that  $\text{KL}(\cdot, \mu_u)$  is convex and lower semicontinuous in the weak-\* topology on  $\mathcal{P}(\mathcal{C}_z^\varepsilon)$ . Thus, since  $\tau > 0$ , the objective in (9) is convex and upper semicontinuous. This together with convex and weak-\* compactness of  $\mathcal{P}(\mathcal{C}_z^\varepsilon)$  entail that (9) has a non-empty convex and weak-\* compact set of solutions. Uniqueness of the minimizer then follows from strong convexity of  $\text{KL}(\cdot, \mu_u)$  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  $\psi_\tau(\theta) := \int_{\mathcal{C}_z^\varepsilon} \mathcal{L}(\theta, z') d\mu(z') - \tau \text{KL}(\mu, \mu_u)$ . The function  $\tau \mapsto \psi_\tau(\theta)$  obviously increases as  $\tau$  decreases and so is  $g_\tau$ . Continuity of  $\mathcal{L}$ , compactness of  $\mathcal{C}_z^\varepsilon$  and Proposition 4.1 entail that  $g_\tau$  is continuous and converges pointwise to  $g$ . The  $\Gamma$ -convergence claim in this case then follows from [23, Proposition 5.4 and Remark 5.5].

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

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

By definition of  $R_{C_z^\varepsilon}$ , there exists  $\bar{z}$  such that  $\mathbb{B}^{R_{C_z^\varepsilon}}(\bar{z}) \subset C_z^\varepsilon$ . Convexity of  $C_z^\varepsilon$  entails that  $\tau(\mathbb{B}^{R_{C_z^\varepsilon}}(\bar{z}) - z^*) + z^* = (1 - \tau)z^* + \tau\mathbb{B}^{R_{C_z^\varepsilon}}(\bar{z}) \subset C_z^\varepsilon$ , and thus

$$\begin{aligned} \tau \log \left( \int_{C_z^\varepsilon} e^{\frac{\mathcal{L}(\theta, z')}{\tau}} dz' \right) &\geq g(\theta) + \tau \log \int_{\tau(\mathbb{B}^{R_{C_z^\varepsilon}}(\bar{z}) - z^*) + z^*} e^{\frac{-L_{\mathcal{Z}} \|z' - z^*\|}{\tau}} dz' \\ &= g(\theta) + \tau \log \left( \tau^m \int_{\mathbb{B}^{R_{C_z^\varepsilon}}(\bar{z})} e^{-L_{\mathcal{Z}} \|z' - z^*\|} dz' \right) \\ &\geq g(\theta) + \tau \log \left( \tau^m \int_{\mathbb{B}^{R_{C_z^\varepsilon}}(\bar{z})} e^{-L_{\mathcal{Z}} (\|z' - \bar{z}\| + \|z^* - \bar{z}\|)} dz' \right) \tag{29} \\ &\geq g(\theta) + \tau \log \left( \tau^m \int_{\mathbb{B}^{R_{C_z^\varepsilon}}(0)} e^{-L_{\mathcal{Z}} (R_{C_z^\varepsilon} + D_{C_z^\varepsilon})} dz' \right) \\ &= g(\theta) - m\tau \log(\tau^{-1}) + \tau \log(\text{vol}(\mathbb{B}^{R_{C_z^\varepsilon}}(0))) - \tau L_{\mathcal{Z}} (R_{C_z^\varepsilon} + D_{C_z^\varepsilon}). \end{aligned}$$

Inserting this into the expression of  $g_\tau$  (see (10)), we get the upper-bound.

#### A.4 Proof of Theorem 4.5

Compactness of  $\Theta$  entails that  $g_\tau$  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 Theorem 4.3 and [23, Theorem 7.8]. The second claim is a consequence of  $\Gamma$ -convergence of  $g_\tau$  (Theorem 4.3), compactness of  $\Theta$  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 (11). 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, for any  $t > 0$ ,*

$$|g_{\tau, N}(\theta) - g_\tau(\theta)| \leq \tau e^{\frac{\bar{\varepsilon} - \underline{\varepsilon}}{\tau}} \sqrt{\frac{t \log N}{N}}, \tag{30}$$

*with probability at least  $1 - 2N^{-t}$ . The bound holds almost surely for any  $t > 1$ .*

*Proof.* To lighten the notation, denote

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

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

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

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  $\log(1+t) \leq t$ ,  $t \geq 0$ , we can write, for any  $\epsilon \geq 0$

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

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

$$\begin{aligned} \Pr(|g_{\tau, N}(\theta) - g_\tau(\theta)| \geq \epsilon) &\leq 2 \exp\left(-\frac{2N^2 e^{2\mathcal{L}/\tau} \epsilon^2}{N \left(e^{\bar{\mathcal{L}}/\tau} - e^{\underline{\mathcal{L}}/\tau}\right)^2 \tau^2}\right) \\ &\leq 2 \exp\left(-\frac{2N e^{-2(\bar{\mathcal{L}} - \underline{\mathcal{L}})/\tau} \epsilon^2}{\tau^2}\right). \end{aligned}$$

Taking

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

we get

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

as claimed. The almost sure statement follows from the Borel-Cantelli lemma.  $\square$

## A.6 Proof of Theorem 4.7

### A.6.1 Proof of Lemma 4.8

To lighten notation, denote the probability measures

$$d\mu(z') := \frac{1}{\text{vol}(\mathcal{C}_z^\varepsilon)} \frac{e^{\frac{\mathcal{L}(\theta, z')}{\tau}}}{S} dz' \quad \text{and} \quad d\mu_N(z') := \frac{1}{N} \sum_{k=1}^N \frac{e^{\frac{\mathcal{L}(\theta, z'_k)}{\tau}}}{S_N} \delta_{z'_k}$$

where

$$S := \frac{1}{\text{vol}(\mathcal{C}_z^\varepsilon)} \int_{\mathcal{C}_z^\varepsilon} e^{\frac{\mathcal{L}(\theta, z')}{\tau}} dz' := \quad \text{and} \quad S_N := \frac{1}{N} \sum_{k=1}^N e^{\frac{\mathcal{L}(\theta, z'_k)}{\tau}}$$

It follows from (14) and (19) that

$$\nabla g_\tau(\theta) - \nabla g_{\tau, N}(\theta) = \int_{\mathcal{C}_z^\varepsilon} \nabla_\theta \mathcal{L}(\theta, z') d\mu(z') - \int_{\mathcal{C}_z^\varepsilon} \nabla_\theta \mathcal{L}(\theta, z') d\mu_N(z')$$

and thus, by  $L_{\Theta, \mathcal{Z}}$ -Lipschitz continuity of  $\mathcal{L}(\cdot, z')$  in  $\theta$ , uniformly in  $\mathcal{Z}$ , we get

$$\|\nabla g_\tau(\theta) - \nabla g_{\tau, N}(\theta)\| \leq L_{\Theta, \mathcal{Z}} \int_{\mathcal{C}_z^\varepsilon} \left| 1 - \frac{S}{S_N} \right| d\mu(z') + \left\| \int_{\mathcal{C}_z^\varepsilon} \nabla_\theta \mathcal{L}(\theta, z') \left( d\mu(z') \frac{S}{S_N} - d\mu_N(z') \right) \right\|. \quad (31)$$

For the first term, we have

$$\begin{aligned} \int_{\mathcal{C}_z^\varepsilon} \left| 1 - \frac{S}{S_N} \right| d\mu(z') &\leq \left( \int_{\mathcal{C}_z^\varepsilon} \frac{e^{\frac{\mathcal{L}(\theta, z')}{\tau}}}{\text{vol}(\mathcal{C}_z^\varepsilon)} dz' \right) \left| \frac{S_N - S}{S S_N} \right| \\ &\leq e^{\frac{\bar{\mathcal{L}} - 2\underline{\mathcal{L}}}{\tau}} |S_N - S|. \end{aligned}$$

Since  $S = \mathbb{E}[S_N]$ , we get from the proof of Lemma A.1 that for any  $t > 0$ ,

$$\int_{\mathcal{C}_z^\varepsilon} \left| 1 - \frac{S}{S_N} \right| d\mu(z') \leq e^{2\frac{\bar{\mathcal{L}} - \underline{\mathcal{L}}}{\tau}} \sqrt{\frac{t \log N}{2N}}$$

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

Let us now turn to the second term in (31). Denote

$$G := \frac{1}{\text{vol}(\mathcal{C}_z^\varepsilon)} \int_{\mathcal{C}_z^\varepsilon} \nabla_\theta \mathcal{L}(\theta, z') e^{\frac{\mathcal{L}(\theta, z')}{\tau}} dz' := \quad \text{and} \quad G_N := \frac{1}{N} \sum_{k=1}^N \nabla_\theta \mathcal{L}(\theta, z'_k) e^{\frac{\mathcal{L}(\theta, z'_k)}{\tau}}.$$

We then have

$$\left\| \int_{\mathcal{C}_z^\varepsilon} \nabla_\theta \mathcal{L}(\theta, z') \left( d\mu(z') \frac{S}{S_N} - d\mu_N(z') \right) \right\| = \frac{\|G_N - G\|}{S_N} \leq e^{\frac{-\underline{\mathcal{L}}}{\tau}} \|G_N - G\|.$$

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

$$\Pr(\|G_N - G\| > \epsilon) \leq \Pr\left(\max_j |(G_N)_j - G_j| > \epsilon/\sqrt{p}\right) \leq p \max_j \Pr(|(G_N)_j - G_j| > \epsilon/\sqrt{p}).$$

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

$$\left\| \int_{\mathcal{C}_z^\epsilon} \nabla_\theta \mathcal{L}(\theta, z') \left( d\mu(z') \frac{S}{S_N} - d\mu_N(z') \right) \right\| \leq L_{\Theta, Z} e^{\frac{\bar{\mathcal{L}} - \underline{\mathcal{L}}}{\tau}} \sqrt{\frac{2tp \log N}{N}}$$

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

### A.6.2 Proof of Lemma 4.9

For any Borel set  $\mathcal{C} \subset \mathbb{R}^m$  and  $k \geq 0$  and any  $k \in \mathbb{N}$ ,  $\mathcal{H}^k(\mathcal{C})$  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.

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

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

We will show the stronger statement that our assumptions,  $\mu_\tau$  converges in the narrow topology to a measure  $\mu \in \mathcal{P}(\text{Argmax } \mathcal{L}(\theta, \mathcal{C}))$  as  $\tau \rightarrow 0^+$ , i.e., for all  $f \in \mathcal{C}$

$$\int_{\mathcal{C}} f(z') d\mu_\tau(z') \xrightarrow{\tau \rightarrow 0^+} \int_{\mathcal{C}} f(z') d\mu(z').$$

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). Moreover, by (H.6)(a), for any  $r \geq 3$ ,  $\mathcal{M}$  is  $\mathcal{C}^r$ -stratifiable and thus the strata  $(\mathcal{M}_i)_{i \in I}$  are  $\mathcal{C}^r$ -smooth compact submanifolds.

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

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

Let  $\mathcal{U} := \bigcup_{i \in I} \mathcal{U}_i$ . We then have

$$\int_{\mathcal{C}} f(z') e^{\frac{\mathcal{L}(\theta, z')}{\tau}} dz' = \int_{\mathcal{C} \cap \mathcal{U}} f(z') e^{\frac{\mathcal{L}(\theta, z')}{\tau}} dz' + \int_{\mathcal{C} \setminus (\mathcal{C} \cap \mathcal{U})} f(z') e^{\frac{\mathcal{L}(\theta, z')}{\tau}} dz'. \quad (32)$$

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 (32) verifies, for any  $s \geq 0$ ,

$$\tau^{-s} \int_{\mathcal{C} \setminus (\mathcal{C} \cap \mathcal{U})} f(z') e^{\frac{\mathcal{L}(\theta, z')}{\tau}} dz' \leq (\text{vol}(\mathcal{C}) \sup |f(\mathcal{C})|) \tau^{-s} e^{-\kappa/\tau} \rightarrow 0 \quad \text{uniformly as } \tau \rightarrow 0^+. \quad (33)$$

Let us now turn to the first integral. We have, for  $\epsilon$  sufficiently small

$$\int_{\mathcal{C} \cap \mathcal{U}} f(z') e^{\frac{\mathcal{L}(\theta, z')}{\tau}} dz' = \sum_{i \in I} \int_{\mathcal{C} \cap \mathcal{U}_i} f(z) e^{\frac{\mathcal{L}(\theta, z')}{\tau}} dz'. \quad (34)$$

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 [36, Theorem 4.12] (see [36, Definition 4.1] for definition of sets of positive reach). Thus, it follows from [36, Theorem 4.8] that  $P_{\mathcal{M}_i}$  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 [50, Theorem 3.2.22(3)] to get

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

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

$$\mathbf{J}_{m_i}(P_{\mathcal{M}_i})(u) = \sqrt{\det(\mathbf{D}(P_{\mathcal{M}_i})(u) \mathbf{D}(P_{\mathcal{M}_i})(u)^\top)}$$

and  $\mathbf{D}$  is the derivative operator. In addition, since  $\mathcal{M}_i$  is  $\mathcal{C}^r$ -smooth, we have from [51, Proposition 5.1] that  $P_{\mathcal{M}_i}$  is  $\mathcal{C}^{r-1}$ -smooth with Lipschitz derivative on  $\mathcal{U}_i$  (taking  $\epsilon$  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

$$\begin{aligned} \lim_{\tau \rightarrow 0^+} \tau^{-\frac{m-m_i}{2}} \int_{\mathcal{C} \cap \mathcal{U}_i} f(z) e^{\frac{\mathcal{L}(\theta, z)}{\tau}} dz' \\ = 2^{\frac{m-m_i}{2}} (m-m_i) \alpha_{(m-m_i)} \beta_{(m-m_i)} \int_{\mathcal{M}_i} f(v) \left( \prod_{j=1}^{m-m_i} \lambda_j(v)^{-\frac{1}{2}} \right) d\mathcal{H}^{m_i}(v), \end{aligned} \quad (35)$$

where  $(-\lambda_j(v))_j$  are the  $m-m_i$  eigenvalues of the Hessian  $\nabla_z^2 \mathcal{L}(\theta, v)$  for  $v \in \mathcal{M}_i$ , which are negative by (H.6)(b),  $\alpha_k$  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 (35) that for any  $i > j$ ,

$$\lim_{\tau \rightarrow 0^+} \tau^{-\frac{m-m_j}{2}} \int_{\mathcal{C} \cap \mathcal{U}_i} f(z) e^{\frac{\mathcal{L}(\theta, z)}{\tau}} dz' = \lim_{\tau \rightarrow 0^+} \tau^{\frac{m_j-m_i}{2}} \left( \tau^{-\frac{m-m_i}{2}} \int_{\mathcal{C} \cap \mathcal{U}_i} f(z) e^{\frac{\mathcal{L}(\theta, z)}{\tau}} dz' \right) = 0. \quad (36)$$

Combining (36) (for  $j=1$ ) and (35) (for  $i=1$ ) with (32), (33) and (34), we get

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

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

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

By the continuity assumption (H.4) on  $\nabla_\theta \mathcal{L}(\theta, z')$ , we deduce that

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

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

## A.7 Proof of Theorem 5.1

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

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

where we used that the functions  $g_i$  are path differentiable for the Clarke subdifferential by [39, Theorem 5.8]. Therefore,  $G$  is a Lyapunov function for the set  $\text{crit-}G$ . Moreover, by [41, Theorem 6],  $G(\text{crit-}G)$  has empty interior.

By the almost sure boundedness assumption,  $\max_i \|\partial^C g_i(\theta_k)\|$  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  $\zeta_k$  is a zero-mean uniformly bounded martingale difference noise. These uniform boundedness properties and the choice of the sequence  $\gamma_k$  allows to apply [43, Remark 1.5(ii) and Proposition 1.4] to get by [43, 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 (24). Combining this with [43, Theorem 3.6 and Proposition 3.27] gives the claimed results.

## A.8 Proof of Theorem 5.2

Clearly, if the sequence  $(\tau_k, N_k)_{k \in \mathbb{N}}$  is such that the right hand side in (28) vanishes as  $k \rightarrow +\infty$ , then  $\lim_{k \rightarrow +\infty} \|e_k\| = 0$  almost surely. We are then in position to invoke [52, Remark 4.4] to get that the conclusions of [43, Remark 1.5(ii) and Proposition 1.4] still hold provided that  $\gamma_k$  decays as devised. The rest of the proof is then same as that of Theorem 5.1.

## B Additional information on experiences

### B.1 Parametrization Avila

Following the recommendations of the paper that introduced the dataset [46] 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
<b>General parameters for all trainings</b>		
Epochs	1500	Number of epochs
Optimizer	SGD	Only optimization algorithm that stays within our guaranties
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 [53]	Default initialization for Pytorch modules
<b>Parameters for adversarial training</b>		
Adversarial Loss	Cross Entropy	
Iteration number	80	Iterations for adversarial attack
Attack norm	$\ell_\infty$	Norm of the attack, taken accordingly to the Sampling ball
<b>Parameters for robust training</b>		
Monte-Carlo sampling	500 000	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