Minimax Excess Risk of First-Order Methods for Statistical Learning with Data-Dependent Oracles

Kevin Scaman, Mathieu Even, Batiste Le Bars, Laurent Massoulié

Minimax Excess Risk of First-Order Methods for Statistical Learning with Data-Dependent Oracles

Kevin Scaman **Mathieu Even** **Batiste Le Bars** **Laurent Massoulié**
Inria Paris - Département d'informatique de l'ENS, PSL Research University

Abstract

In this paper, our aim is to analyse the generalization capabilities of first-order methods for statistical learning in multiple, different yet related, scenarios including supervised learning, transfer learning, robust learning and federated learning. To do so, we provide sharp upper and lower bounds for the minimax excess risk of strongly convex and smooth statistical learning when the gradient is accessed through partial observations given by a *data-dependent* oracle. This novel class of oracles can query the gradient with any given data distribution, and is thus well suited to scenarios in which the training data distribution does not match the target (or test) distribution. In particular, our upper and lower bounds are proportional to the smallest mean square error achievable by gradient estimators, thus allowing us to easily derive multiple sharp bounds in the aforementioned scenarios using the extensive literature on parameter estimation.

1 INTRODUCTION

In statistical learning, one is often interested in minimizing a population risk, also known as test loss, of the form $\mathcal{L}(x) = \mathbb{E}_{\xi\sim\mathcal{D}}[\ell(x,\xi)]$ for some loss function ℓ and (unknown) test data distribution $\mathcal{D}$. The question that arises then is how small can the *excess risk* $\mathcal{L}(\hat{x}) - \inf_x \mathcal{L}(x)$ be, for $\hat{x}$ computed using some given restricted information?

In classical supervised learning settings, $\hat{x}$ is typically computed with n i.i.d. samples drawn from $\mathcal{D}$ and usually corresponds to the minimizer of the empirical counterpart of $\mathcal{L}(x)$ computed with those samples. In that setting, the excess risk can be controlled through the concept of *generalization error*, which quantifies the degree to which minimizing the empirical risk, also known as the training loss, is similar to minimizing the test loss. Among the several approaches that have been proposed to bound generalization errors, the most prominent ones are based on the complexity of the hypothesis class like the Vapnik-Chervonenkis dimension (Vapnik, 2000; Vapnik and Chervonenkis, 2015; Blumer et al., 1989) or Rademacher complexity (Bartlett and Mendelson, 2003; Bousquet et al., 2004), algorithmic stability (Mukherjee et al., 2006; Bousquet and Elisseeff, 2002), PAC-Bayesian bounds (McAllester, 1998; Alquier, 2024), or more recently information-theoretic generalization bounds (Xu and Raginsky, 2017). Nowadays, it is commonly accepted that, in this context, the generalization error alone is not sufficient to control the excess risk. As the empirical risk minimizer cannot always be computed in an exact manner, the *optimization error* must also be taken into account, measuring the algorithm's ability to properly minimize the empirical risk, and shedding light on a generalization-optimization trade-off (Bottou and Bousquet, 2007). Over the last few years, a substantial amount of work have therefore been dedicated in controlling these errors, notably through the study of the generalization properties of *optimization algorithms* (Lin et al., 2016; London, 2017; Zhou et al., 2018; Amir et al., 2021; Neu et al., 2021), where approaches based on algorithmic stability have encountered a large success (Hardt et al., 2016; Kuzborskij and Lampert, 2018; Bassily et al., 2020; Lei and Ying, 2020a,b; Schliserman and Koren, 2022).

Above approaches are however mostly tailored for standard supervised learning and empirical risk minimization. Hence, an additional analysis is required for all the different variations and flavors of this problem, such as transfer learning / domain adaptation (Ben-David et al., 2006) in which the training distribution differs from that of the testing distribution, or robust learning in which a small portion of the training data may be corrupted by an arbitrary noise. Note further

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

that, while there exist a large panel of upper bounds on the generalization error, the optimization error (Arjevani et al., 2023; Bubeck, 2015; Drori and Taylor, 2022) or, more generally, the excess risk, the question of their optimality with respect to some lower bounds is most of the time lacking or specific to a particular algorithm or class of data-distribution (Zhang et al., 2022; Schliserman and Koren, 2023). For instance, Arjevani et al. (2023) proved lower bounds for the non-convex stochastic case, under an oracle framework that inspired our formalism; Devolder et al. (2013) considers *inexact oracles*. However, Devolder et al. (2013)'s work is quite different from ours, since their goal is to analyze different algorithms under a unified framework.

Contributions. In this work, we propose a unified framework to analyse, among others, the aforementioned statistical learning problems in a more systematic manner. Our general framework goes beyond the decomposition between generalization and optimization error and analyzes instead directly the ability of an optimization algorithm to minimize the population risk given partial, and possibly biased, information. Our contributions can be summarized as follows:

- We tackle the problem of controlling the test loss $\mathcal{L}$ trough the lens of first-order optimization methods with gradient oracles. Contrary to the stochastic optimization setting (Agarwal et al., 2012; Arjevani et al., 2023), we introduce the novel notion of **data-dependent oracle**, more adapted to the case where the gradients are computed over a fixed data set, used possibly several times during optimization.
- The data used by the oracle being arbitrary, we show that **our setting is rather generic** and contains supervised learning, transfer learning, robust learning, and federated learning.
- We provide **upper and lower bounds** for the minimax excess risk of statistical learning problems and show that they are sharp for deterministic and for more classical i.i.d. oracles. Our bounds shed light on a novel quantity called **best approximation error**, generalizing conditional expectations and conditional standard deviations.
- We show that our general bounds can be applied to **several learning settings**, allowing to obtain problem-specific excess risk bounds and recover some known results of the literature. In particular, we show that in the case of standard supervised learning, mini-batch gradient descent with increasing batch sizes and a warm start can reach an excess risk that is optimal up to a multiplicative factor, thus motivating the use of this optimization scheme in practical applications.

Outline of the paper. In Section 2 we introduce our statistical learning setting, where we define data-dependent oracles, the algorithms considered, as well as our set of assumptions. We also introduce the aforementioned quantity called *best approximation error*. In Section 3, we derive upper and lower bounds for the minimax excess risk of statistical learning with any given data-dependant oracle and discuss their optimality. Finally, in Section 4 we apply our general bounds to supervised learning, transfer learning, federated learning, robust learning and learning from a fixed predetermined dataset.

Notations. In what follows, we denote as $\mathbb{F}(\mathcal{X}, \mathcal{Y})$ (resp. $\mathbb{M}(\mathcal{X}, \mathcal{Y})$) the space of functions (resp. measurable functions) from $\mathcal{X}$ to $\mathcal{Y}$ (both measurable spaces). Let $\|x\| = \sqrt{\sum_i x_i^2}$ be the canonical norm in $\mathbb{R}^d$, and $\rho(A)$ the nuclear norm of the matrix $A \in \mathbb{R}^{d\times D}$. A function f is B-Lipschitz if $\|f(x)-f(x')\| \leq B\|x-x'\|$ for all $x, x' \in \mathcal{X}$. A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex (where $\mu \geq 0$) if $\forall x, y \in \mathbb{R}^d$, we have $f(x) - f(y) \geq \langle \nabla f(y), x - y\rangle + \frac{\mu}{2}\|x - y\|^2$, and convex if this holds for $\mu = 0$. f is L-smooth if it is differentiable and its gradient is L-Lipschitz. Finally, for two functions $a, b : \mathcal{Z} \to \mathbb{R}^+$, we write $a = \Theta(b)$ (resp. $a = O(b)$) if there exists $c, C > 0$ such that for all $z \in \mathcal{Z}$, $cb(z) \leq a(z) \leq Cb(z)$ (resp. $a(z) \leq Cb(z)$).

2 PROBLEM SETUP

We now provide precise definitions for statistical learning under data-dependent oracles, as well as the minimax estimation error used in our analysis.

2.1 Statistical learning

Consider the population risk minimization problem:

$$\inf_{x\in\mathbb{R}^d} \mathcal{L}(x) \triangleq \mathbb{E}_{\xi\sim\mathcal{D}}\left[\ell(x, \xi)\right], \tag{1}$$

where $\mathcal{D}$ is a probability distribution over the measurable space Ξ and $\ell : \mathbb{R}^d \times \Xi \to \mathbb{R}$ is a loss function that takes as input a model parameter $x \in \mathbb{R}^d$ and a data point $\xi \in \Xi$. For simplicity, for any $y \in \mathbb{R}^d$, we denote as $\nabla\ell_y : \xi \mapsto \nabla_x\ell(y, \xi)$ the gradient of the loss w.r.t. its first coordinate. Moreover, our analysis focuses on strongly-convex and smooth objective functions whose gradients belong to a given function class.

Definition 1 (function class)**.** Let $\mathcal{G} \subset \mathbb{F}(\Xi, \mathbb{R}^d)$ be a class of functions taking data points as input. We denote as $\mathcal{F}_{\text{sc}}(\mathcal{G}, \mathcal{D}, \mu, L)$ the set of μ-strongly convex and L-smooth objective functions $\mathcal{L}(x) = \mathbb{E}_{\xi\sim\mathcal{D}}\left[\ell(x, \xi)\right]$ such that $\forall x \in \mathbb{R}^d$, $\nabla\ell_x \in \mathcal{G}$.

The function class $\mathcal{G}$ is used to encode the regularity of the gradient of the loss with respect to input data, for example (assuming $\Xi = \mathbb{R}^D$ for the first two): 1) affine functions: $\mathcal{G}_{\rm Aff} = \{\xi \mapsto A\xi + b : A \in \mathbb{R}^{d\times D}, b \in \mathbb{R}^d, \rho(A) \leq B\}$, 2) Lipschitz functions: $\mathcal{G}_{\rm Lip} = \{g : \Xi \to \mathbb{R}^d : \forall \xi, \xi' \in \Xi, \|g(\xi) - g(\xi')\| \leq B\|\xi - \xi'\|\}$, and 3) bounded variations: $\mathcal{G}_{\rm Bnd} = \{g : \Xi \to \mathbb{R}^d : \exists c_g \in \mathbb{R}^d, \forall \xi \in \Xi, \|g(\xi) - c_g\| \leq B\}$. Note that all these function spaces are invariant by translation by a constant, a key property for our analysis (see Assumption 1).

Remark 1. All our results, lower and upper bounds, also apply to the minimax excess risk of non-convex smooth and μ-PL functions (see Appendix A).

Example 1 (Least squares regression)**.** Let $\ell(x, \xi = (M, v)) = \frac{1}{2}x^\top Mx - v^\top x$ for $M \in \mathbb{R}^{d\times d}$ and $v \in \mathbb{R}^d$, leading to $\mathcal{G} = \mathcal{G}_{\rm Aff}$ where B is the diameter of the space over which we optimize.

Example 2 (Regularized Lipschitz losses)**.** Let $\mathcal{L}(x) = \lambda\Omega(x) + \mathbb{E}_{\mathcal{D}}[\ell(x, \xi)]$, for some Lipschitz continuous and convex loss ℓ (in its first argument) and a convex regularizer Ω, yielding $\mathcal{G} = \mathcal{G}_{\rm Bnd}$.

2.2 Data-dependent oracles and first-order optimization algorithms

Our objective is to minimize Eq. (1) using optimization algorithms that access $\nabla\mathcal{L}$ via a *data-dependent oracle*, in a setup similar to that of Arjevani et al. (2023).

Definition 2 (Data-dependent oracle)**.** Let $\mathcal{O}, \mathcal{Z}$ be two measurable spaces and $\mathbb{F}(\Xi, \mathbb{R}^d)$ a measurable space of functions. A *data-dependent oracle* is a tuple (O, P_z) where $\mathsf{O} : \mathbb{F}(\Xi, \mathbb{R}^d) \times \mathcal{Z} \to \mathcal{O}$ is a measurable function and P_z is a probability distribution over $\mathcal{Z}$.

At each iteration, optimization algorithms will only be able to access the gradient of the objective function through the *observation* $\mathsf{O}(\nabla\ell_x, z)$, where $x \in \mathbb{R}^d$ is the current model parameter and $z \sim P_z$ is a random seed drawn *prior* to the optimization. In other words, an oracle provides a partial (and possibly random) view of the gradient, for example by accessing the gradient at i.i.d. sampled data points $\xi_i' \sim \mathcal{D}'$ drawn according to a source data distribution $\mathcal{D}' \neq \mathcal{D}$. In such a case, we have $\mathcal{O} = \mathbb{R}^{d\times n}$, $\mathcal{Z} = \Xi^n$, and $\mathsf{O}(g, (\xi_1', \dots, \xi_n')) = (g(\xi_1'), \dots, g(\xi_n'))$. Note that, contrary to the online setting of Arjevani et al. (2023), the randomness is fixed prior to the optimization, and thus each iteration of the optimization will have access to the same data points $\xi_1, \dots, \xi_n$. We now define more precisely the class of algorithms that we will consider in this analysis.

Definition 3 (Optimization algorithm)**.** Let $\mathcal{O}, \mathcal{R}$ be two measurable spaces. An *optimization algorithm* is a tuple $\mathsf{A} = (\{q^{(t)}, s^{(t)}\}_{t\geq 0}, P_r)$ where $q^{(t)} \in \mathbb{M}(\mathcal{O}^t \times \mathcal{R}, \mathbb{R}^d)$ is a query function, $s^{(t)} \in \mathbb{M}(\mathcal{O}^t \times \mathcal{R}, \{0, 1\})$ is a stopping criterion, and P_r is a distribution over $\mathcal{R}$.

For a given data-dependent oracle (O, P_z) and optimization algorithm $\mathsf{A} = (\{q^{(t)}, s^{(t)}\}_{t\geq 0}, P_r)$, we consider the following optimization protocol:

1. We first draw two random seeds: $r \sim P_r$ for the algorithm, and $z \sim P_z$ for the oracle.
2. At each iteration $t \geq 0$, we update the iterates:
$$\begin{aligned} x_{\mathsf{A}[\mathsf{O}]}^{(t)} &= q^{(t)}\left(m_{\mathsf{A}[\mathsf{O}]}^{(t)}, r\right) \\ s_{\mathsf{A}[\mathsf{O}]}^{(t)} &= s^{(t)}\left(m_{\mathsf{A}[\mathsf{O}]}^{(t)}, r\right) \end{aligned} \tag{2}$$
where $m_{\mathsf{A}[\mathsf{O}]}^{(t)} = (\mathsf{O}(\nabla\ell_{x_{\mathsf{A}[\mathsf{O}]}^{(0)}}, z), \dots, \mathsf{O}(\nabla\ell_{x_{\mathsf{A}[\mathsf{O}]}^{(t-1)}}, z))$.
3. The algorithm stops and returns the current iterate $x_{\mathsf{A}[\mathsf{O}]} = x_{\mathsf{A}[\mathsf{O}]}^{(t)}$ as soon as $s_{\mathsf{A}[\mathsf{O}]}^{(t)} = 1$.

In other words, at each iteration, the algorithm updates the model parameter based on all past observations, and then decides to stop (and return the current model parameter) or continue the optimization. If so, the algorithm receives a new observation of the gradient for the current model parameter and proceeds to the next iteration. Note that the algorithm may not terminate, in which case we consider the loss as infinite. Moreover, as discussed in Arjevani et al. (2023), fixing the randomness to a single seed r instead of drawing random seeds $r^{(t)}$ for each iteration does not lose any generality. Finally, we denote as $\mathcal{A}_{\rm rand}$ the class of all optimization algorithms as defined above and, in order to prove lower bounds on the error of optimization algorithms, we assume that the information extracted by the oracle is *invariant* with respect to translations in the following sense.

Assumption 1 (Translation invariance)**.** There exists a measurable function $\varphi : \mathcal{G} \times \mathbb{R}^d \to \mathcal{O}$ such that, for any function $g \in \mathcal{G}$ and constant $c \in \mathbb{R}^d$, $g + c \in \mathcal{G}$ and $\forall z \in \mathcal{Z}$, $\mathsf{O}(g + c, z) = \varphi(\mathsf{O}(g, z), c)$.

Intuitively, Assumption 1 means that translations do not add any information to the oracle, as the translated oracles $\mathsf{O}(g + c, z)$ can be retrieved as a function of the untranslated oracle $\mathsf{O}(g, z)$. This assumption is verified in most settings of interest (see Section 4).

2.3 Minimax excess risk

We evaluate the difficulty of optimizing functions in $\mathcal{F}_{\rm sc}(\mathcal{G}, \mathcal{D}, \mu, L)$ (abbreviated to $\mathcal{F}_{\rm sc}$ below) with a given oracle O via the *minimax excess risk* defined by

$$\varepsilon_{\rm sc}(\mathcal{G}, \mathsf{O}, \mathcal{D}, \mu, L) = \inf_{\mathsf{A}\in\mathcal{A}_{\rm rand}} \sup_{\mathcal{L}\in\mathcal{F}_{\rm sc}} \mathbb{E}\left[\mathcal{L}\left(x_{\mathsf{A}[\mathsf{O}]}\right) - \mathcal{L}^*\right],$$

where $\mathcal{L}^* = \inf_{x\in\mathbb{R}^d} \mathcal{L}(x)$ is the minimum value of the objective function. In other words, the minimax excess risk measures the best worst-case error that a first-order optimization algorithm can achieve on the objective function $\mathcal{L}$, despite only accessing to the gradients via the oracle O. For simplicity, as the terms $\mathcal{D}$, μ, L will be fixed throughout the paper, we will from now on omit them and only write $\varepsilon_{\text{sc}}(\mathcal{G}, \mathsf{O})$.

2.4 Minimax estimation error

Our upper and lower bounds on the minimax excess risk will depend on the ability to create estimators of the expectation over $\mathcal{D}$ of any function in $\mathcal{G}$. This notion, denoted as *minimax estimation error*, is defined via *best approximation errors*, a novel notion that extends conditional standard deviation to measurable functions equipped with arbitrary semi-norms.

Definition 4 (Best approximation error)**.** Let $\mathcal{X}$ and $\mathcal{Y}$ be two measurable spaces, $\mathcal{Z}$ a measurable vector space, and $\|\cdot\|_\nu$ a (possibly infinite) semi-norm over $\mathbb{M}(\mathcal{X}, \mathcal{Z})$. For $f \in \mathbb{M}(\mathcal{X}, \mathcal{Z})$ and $h \in \mathbb{M}(\mathcal{X}, \mathcal{Y})$ two measurable functions, we denote as *best approximation error* of f knowing h the quantity

$$\sigma_\nu(f|h) = \inf_{\varphi\in\mathbb{M}(\mathcal{Y},\mathcal{Z})} \|f - \varphi\circ h\|_\nu \,. \tag{3}$$

In other words, $\sigma_\nu(f|h)$ measures how well can f be approximated using g, and is thus tightly connected to estimation theory (see e.g. Polyanskiy and Wu, 2022).

Example 3 (Conditional standard deviation)**.** When $\mathcal{X}$ is a probability space and $\mathcal{Y} = \mathcal{Z} = \mathbb{R}^d$, the measurable functions $f, h \in \mathbb{M}(\mathcal{X}, \mathbb{R}^d)$ are random variables and we recover that $\sigma_2(f|h) = \sqrt{\mathbb{E}\left[\|f - \mathbb{E}\left[f|h\right]\|^2\right]}$.

Example 4 (Deviations and barycenters)**.** When h is constant (e.g. $\mathcal{Y} = \mathbb{R}$ and $h(x) = 1$), then $\sigma_\nu(f|h) = \sigma_\nu(f|1) = \inf_{c\in\mathcal{Z}} \|f - c\|_\nu$ can encode multiple notions of distance to the *barycenter* of the values $\{f(x)\}_{x\in\mathcal{X}}$, including the the median ($\nu = 1$), mean ($\nu = 2$) and Chebyshev center ($\nu = +\infty$) (Amir, 1984).

In what follows, we will mainly use this quantity for the semi-norms[1] $\|f\|_{\mathcal{G},2} = \sup_{g\in\mathcal{G}} \sqrt{\mathbb{E}\left[\|f(g,z)\|^2\right]}$ and $\|f\|_{2,\mathcal{G}} = \sqrt{\mathbb{E}\left[\sup_{g\in\mathcal{G}} \|f(g,z)\|^2\right]}$. Let $\mathsf{E}_\mathcal{D} : g \mapsto \mathbb{E}_{\xi\sim\mathcal{D}}\left[g(\xi)\right]$ be the expectation over the distribution $\mathcal{D}$. We denote as *minimax estimation error* the quantity

$$\sigma_{\mathcal{G},2}\left(\mathsf{E}_\mathcal{D}|\mathsf{O}\right) = \inf_{\varphi\in\mathbb{M}(\mathcal{Y},\mathcal{Z})} \|\mathsf{E}_\mathcal{D} - \varphi\circ\mathsf{O}\|_{\mathcal{G},2} \,. \tag{4}$$

[1]The supremum in $\|f\|_{2,\mathcal{G}}$ is a lattice supremum, i.e. the smallest measurable function that is almost everywhere larger than all the considered functions, thus ensuring measurability of $\sup_{g\in\mathcal{G}} \|f(g,z)\|^2$.

This quantity measures how well one can approximate the expectation of *any* function in $\mathcal{G}$ over the target distribution $\mathcal{D}$ using the oracle O as input. As we will see below, this quantity is tightly connected to the minimax excess risk.

3 EXCESS RISK BOUNDS OF DATA-DEPENDENT ORACLES

We now detail our upper and lower bounds on the minimax excess risk in various settings of interest.

3.1 General data-dependent oracles

We first provide a lower bound on the minimax excess risk that provides a link between this quantity and the minimax estimation error. The proofs of all propositions are available in the supplementary material.

Proposition 1. *For any distribution $\mathcal{D}$, function class $\mathcal{G}$ and data-dependent oracle O verifying Assumption 1, we have*

$$\varepsilon_{\text{sc}}(\mathcal{G}, \mathsf{O}) \geq \frac{\sigma_{\mathcal{G},2}\left(\mathsf{E}_\mathcal{D}|\mathsf{O}\right)^2}{2\mu} \,. \tag{5}$$

The proof of Proposition 1 relies on simple well-chosen quadratic functions for which the observations of the gradient through the iterations of the optimization algorithm do not significantly change, and whose minimization requires to find a good estimator of the expectation over $\mathcal{D}$ (i.e. a solution to Eq. (4)). Intuitively, Proposition 1 shows that optimizing functions in $\mathcal{F}_{\text{sc}}(\mathcal{G}, \mathcal{D}, \mu, L)$ is at least as difficult as estimating their gradient. Moreover, the quantity $\sigma_{\mathcal{G},2}\left(\mathsf{E}_\mathcal{D}|\mathsf{O}\right)$ can be lower bounded using any information theoretic lower bound on the variance of estimators. In particular, we will use a slight variation of Le Cam's two point method (see, e.g., Section 31.1 in Polyanskiy and Wu, 2022) adapted to our setting.

Proposition 2. *For any distribution $\mathcal{D}$, function class $\mathcal{G}$ and data-dependent oracle O, we have*

$$\sigma_{\mathcal{G},2}\left(\mathsf{E}_\mathcal{D}|\mathsf{O}\right)^2 \geq \sup_{g,g'\in\mathcal{G}} \frac{c_{g,g'}}{4} \|\mathsf{E}_\mathcal{D}(g) - \mathsf{E}_\mathcal{D}(g')\|^2 \,, \tag{6}$$

where $c_{g,g'} = 1 - d_{\text{LC}}(\mathsf{O}(g,z), \mathsf{O}(g',z))$ and $d_{\text{LC}}(p,q)$ is Le Cam's distance (see Appendix B).

In other words, if two functions $g, g' \in \mathcal{G}$ are almost indistinguishable using observations (i.e. Le Cam's distance between their respective distributions is small), then any estimator will necessarily return a similar value on both. As a consequence, if their expectations $\mathsf{E}_\mathcal{D}(g)$ and $\mathsf{E}_\mathcal{D}(g')$ are distant, then the estimator will have a large variance for at least one of the two functions. We will use this result in Section 4 to derive lower bounds in several learning setups.

We now show that, if we replace $\|\cdot\|_{\mathcal{G},2}$ by the (always greater) norm $\|\cdot\|_{2,\mathcal{G}}$, the lower bound in Eq. (5) can be achieved by a simple optimization algorithm.

Proposition 3. *For any distribution $\mathcal{D}$, function class $\mathcal{G}$ and data-dependent oracle O, we have*

$$\varepsilon_{\rm sc}(\mathcal{G},\mathsf{O}) \le \frac{\sigma_{2,\mathcal{G}}\left(\mathsf{E}_{\mathcal{D}}|\mathsf{O}\right)^2}{2\mu}. \tag{7}$$

The proof of Proposition 3 relies on using the simple iterative algorithm $x_{t+1} = x_t - \frac{1}{L}\varphi(o_t)$ where $o_t = \mathsf{O}(\nabla\ell_{x_t}, z)$ and φ is a minimizer of Eq. (4). Note that, if the oracle is $\mathsf{O}(g,z) = \mathbb{E}_{\mathcal{D}}[g(\xi)]$, this amounts to performing gradient descent. The variance of the gradient noise is then bounded by $\sigma_{2,\mathcal{G}}\left(\mathsf{E}_{\mathcal{D}}|\mathsf{O}\right)^2$ by taking the supremum over all functions $g \in \mathcal{G}$ *before* the expectation over z, thus avoiding issues related to the correlation between x_t and z. Of course, such a crude upper bound is often suboptimal, as $\sigma_{2,\mathcal{G}}\left(\mathsf{E}_{\mathcal{D}}|\mathsf{O}\right)$ allows for the function g to be chosen adversarially for each random observation z, and thus does not take advantage of the independence between these two quantities. However, we now show that two additional assumptions lead to sharper upper bounds: 1) deterministic oracles and 2) i.i.d. oracles (see Section 3.3).

3.2 Exact risk with deterministic oracles

Quite remarkably, the upper and lower bounds match in the case of deterministic oracles, thus providing an *exact* relationship between minimax excess risk and minimax estimation error.

Corollary 1. *If the observations are deterministic, i.e. $\mathsf{O}(g,z) = \tilde{\mathsf{O}}(g)$ is independent of z, then*

$$\varepsilon_{\rm sc}(\mathcal{G},\mathsf{O}) = \frac{\sigma_{\mathcal{G}}(\mathsf{E}_{\mathcal{D}}|\tilde{\mathsf{O}})^2}{2\mu}, \tag{8}$$

where $\|\tilde{\mathsf{O}}\|_{\mathcal{G}} = \sup_{g\in\mathcal{G}}\|\tilde{\mathsf{O}}(g)\|$.

In Section 4, we will use this result to compute the minimax excess risk in one scenario: learning from fixed predetermined data-points (e.g. a grid).

3.3 Refined upper bounds with i.i.d. oracles

We now focus on the case where observations are of the form:

$$\mathsf{O}_n(g,z) = \left(\mathsf{O}(g,z^{(1)}),\dots,\mathsf{O}(g,z^{(n)})\right), \tag{9}$$

where $z = (z^{(1)},\dots,z^{(n)})$ and the $z^{(i)}$ are i.i.d. random variables. For example, if the random variables $z^{(i)}$ are sampled from $\mathcal{D}$ and $\mathsf{O}(g,z) = g(z)$, this amounts to classical supervised learning with n samples. We first provide an upper bound on the minimax excess risk using a simple mini-batch algorithm with warmup.

Algorithm 1 Minibatch GD with warmup

Input: iterations T, sizes $(n_t)_{t<T}$, functions $(\varphi_k)_{k\in\mathbb{N}^*}$
Output: current iterate x

```
m ← 0, x ← 0, o ← O(∇ℓ_x, z^(1))
T_wu ← κ ln( (‖φ_1(o)‖² + ‖E_D − φ_1 ∘ O‖²_{2,G}) / (εμ) )
for t ∈ ⟦0, T_wu − 1⟧ do
    o ← O(∇ℓ_x, z^(1))
    x ← x − (1/L) φ_1(o)
end for
for t ∈ ⟦0, T − 1⟧ do
    o_i ← O(∇ℓ_x, z^(m+i)) for i ∈ ⟦1, n_t⟧
    x ← x − (1/L) φ_{n_t}(o_1, ..., o_{n_t})
    m ← m + n_t
end for
```

Proposition 4. *Let $a > 0$ and $\kappa = L/\mu$. For any distribution $\mathcal{D}$, function class $\mathcal{G}$ and i.i.d. oracle O_n,*

$$\varepsilon_{\rm sc}(\mathcal{G},\mathsf{O}_n) \le \frac{\sigma_{\mathcal{G},2}\left(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_{\tilde{n}}\right)^2}{2\mu} + \frac{\widetilde{\Delta}}{n^a},$$

where $\tilde{n} = \left\lfloor \frac{n-1}{1+a\kappa\log n}\right\rfloor$, and $\widetilde{\Delta} = \frac{\sigma_{2,\mathcal{G}}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_1)^2}{2\mu}$.

Similarly to Proposition 3, the proof of Proposition 4 relies on the use of an iterative algorithm akin to a gradient descent variant, here mini-batch gradient descent with a warm-up phase. The algorithm, described in Alg. (1), requires functions φ_k for $k\in\mathbb{N}^*$ minimizing Eq. (4) for the oracle O_k, and mini-batch sizes $n_t = \tilde{n}$. Note that, apart from the first sample used during the warmup phase, samples are only used once. This may seem suboptimal, as stability theory shows that one can often reuse samples without a significant cost. However, note that the stability of Alg. (1) depends on the regularity of the functions φ_k, which is not controlled in general. Moreover, Proposition 4 shows that, even with a simple mini-batch scheme without replacement, one can already obtain matching upper and lower bounds up to logarithmic factors, as $\tilde{n} = \Omega(n/\kappa\log n)$. In particular, for oracles of the form of Eq. (9) for which $\sigma_{2,\mathcal{G}}\left(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_1\right) < +\infty$ and $\sigma_{\mathcal{G},2}\left(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n\right) = \Theta(n^{-b})$, Proposition 4 implies that $\varepsilon_{\rm sc}(\mathcal{G},\mathsf{O}_n) = \widetilde{\Theta}(n^{-2b})$, where $\widetilde{\Theta}$ hides logarithmic factors (using Proposition 4 with $a \ge 2b$). Finally, the logarithmic factor can be removed when the minimax estimation error is bounded by a quantity of the form $a + b/n$ (see Section 4 for examples of such a bound).

Proposition 5. *Let $a, b > 0$, $n \ge 3$, and assume that $\sigma_{\mathcal{G},2}\left(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n\right)^2 \le a + b/n$. Then, for any function class $\mathcal{G}$, distribution $\mathcal{D}$ and i.i.d. oracle O_n, we have*

$$\varepsilon_{\rm sc}(\mathcal{G},\mathsf{O}_n) \le \frac{a}{2\mu} + \frac{6\kappa b}{\mu n} + \widetilde{\Delta}e^{-\frac{n}{6\kappa}},$$

where $\widetilde{\Delta} = \frac{\sigma_{2,\mathcal{G}}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_1)^2}{2\mu}$ and $\kappa = L/\mu$.

This result is obtained with exponentially increasing mini-batch sizes $n_t = \lceil n(1-c)c^{T-t-1}/2 \rceil$ where $c = \sqrt{1-\kappa^{-1}}$, and $T = \lfloor n/2 \rfloor$. This allows to have more precision at the end of the optimization, when the error is low and greater precision is required.

4 APPLICATIONS

In this section, we specify the general upper and lower bounds obtained in the previous sections to several statistical learning scenarios. The list of results obtained for the Lipschitz and bounded variation function classes $\mathcal{G}_{\mathrm{Lip}}$ and $\mathcal{G}_{\mathrm{Bnd}}$ are reported in Table 1.

4.1 Supervised learning

We now consider the typical supervised learning setup, in which the training samples are drawn i.i.d. according to the target distribution, i.e.

$$\mathsf{O}_n^{\mathrm{SL}}(g) = (g(\xi_1), \dots, g(\xi_n)),$$

where $\xi_i \sim \mathcal{D}$ are i.i.d. random variables. A classical approach when using this oracle consists in minimizing the empirical loss, by computing $\hat{x}_n \in \operatorname{argmin}_{x \in \mathbb{R}^d} \frac{1}{n} \sum_i \ell(x, \xi_i)$ using for instance a gradient descent algorithm.

Applying Proposition 4 to the oracle $\mathsf{O}_n^{\mathrm{SL}}$ gives an upper bound involving $\sigma_{\mathcal{G},2}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_{\tilde{n}}^{\mathrm{SL}})^2/2\mu$ on the minimax excess risk (for $\tilde{n}$ specified in Proposition 4, of order $\widetilde{\mathcal{O}}(n/\kappa)$). This quantity is however hard to compute in most cases, as $\sigma_{\mathcal{G},2}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_{\tilde{n}}^{\mathrm{SL}})$ is a minimum over all measurable functions $\varphi \in \mathbb{M}(\mathcal{O}, \mathbb{R}^d)$ (see Eq. (4)). However, a simple upper bound can be obtained by using the average over the samples $\varphi(g(\xi_1), \dots, g(\xi_n)) = \frac{1}{n}\sum_i g(\xi_i)$. The use of such a function leads to the usual (mini-batch) gradient descent algorithm on the empirical risk, and to the following proposition, which shows that $\sigma_{2,\mathcal{G}}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n^{\mathrm{SL}})$ can be upper bounded by the form specified in Proposition 5.

Proposition 6. *For $n \geq 1$, we have*

$$\sigma_{\mathcal{G},2}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n^{\mathrm{SL}})^2 \quad \leq \quad \frac{\|\mathsf{V}_{\mathcal{D}}\|_{\mathcal{G}}}{n}, \tag{10}$$

where $\|\mathsf{V}_{\mathcal{D}}\|_{\mathcal{G}} = \sup_{g \in \mathcal{G}} \operatorname{var}(g(\xi))$ and $\xi \sim \mathcal{D}$.

The quantity $\|\mathsf{V}_{\mathcal{D}}\|_{\mathcal{G}} = \sup_{g \in \mathcal{G}} \operatorname{var}(g(\xi_1))$ controls the variation of the gradient of the loss over the data distribution, and is easy to compute for simple function classes defined in Section 2.1: 1) Affine functions: $\|\mathsf{V}_{\mathcal{D}}\|_{\mathcal{G}_{\mathrm{Aff}}} \leq B^2 \operatorname{var}(\xi)$ (with equality if $\Xi = \mathbb{R}^D$ and $D \leq d$), 2) Lipschitz functions: $\|\mathsf{V}_{\mathcal{D}}\|_{\mathcal{G}_{\mathrm{Lip}}} \leq B^2 \operatorname{var}(\xi)$ (with equality if $\Xi = \mathbb{R}^D$ and $D \leq d$), and 3) Bounded variation: $\|\mathsf{V}_{\mathcal{D}}\|_{\mathcal{G}_{\mathrm{Bnd}}} \leq B^2$ (with equality if $\exists A \subset \Xi$ mesurable s.t. $\mathbb{P}_{\mathcal{D}}(A) = \frac{1}{2}$).

In particular, our results provide new excess risk bounds for the set of smooth and strongly convex (or PL, see Appendix A) functions whose gradient is Lipschitz w.r.t. x and w.r.t. input data (i.e. $\mathcal{G} = \mathcal{G}_{\mathrm{Lip}}$), by applying the bound $\|\mathsf{V}_{\mathcal{D}}\|_{\mathcal{G}_{\mathrm{Lip}}} \leq B^2 \operatorname{var}(\xi)$ to Proposition 6 and Proposition 5.

We now explicit our bounds on a classical setting for which we can compare our results: the bounded variation function class $\mathcal{G}_{\mathrm{Bnd}}$, in which the gradients are contained in a ball, and includes regularized Lipschitz losses (see Example 2).

Proposition 7. *Assume that $\forall c \in [0,1]$, $\exists A \subset \Xi$ measurable s.t. $\mathbb{P}_{\mathcal{D}}(A) = c$. Then, for $n \geq 3$, we have*

$$\frac{B^2}{8\mu n} \leq \varepsilon_{\mathrm{sc}}(\mathcal{G}_{\mathrm{Bnd}}, \mathsf{O}_n^{\mathrm{SL}}) \leq \frac{11\kappa B^2}{\mu n}. \tag{11}$$

The upper and lower bounds in Proposition 7 match up to a multiplicative factor proportional to κ, thus providing a relatively tight approximation of the minimax excess risk in this setting. Also, note that the assumption w.r.t. the measure $\mathcal{D}$ in the previous proposition is only necessary for the lower bound to hold, and is automatically verified for continuous distributions.

Comparison with the literature. As mentioned above, the bounded variation setup includes the widely studied case of B-Lipschitz continuous loss functions. This setting is handled by Sridharan et al. (2008); Bartlett et al. (2005) and Bach (2021, Chapter 4.5.5) for regularized risk minimization, and under additional structural assumptions on ℓ. It is also treated by the stability community, such as in Bousquet and Elisseeff (2002) for regularized objectives, or more generally for μ-strongly convex functions in Hardt et al. (2016). In all these works, the authors study the excess risk $\mathcal{L}(\hat{x}_n) - \inf_{x \in \mathbb{R}^d} \mathcal{L}(x)$, for $\hat{x}_n$ function of $\mathsf{O}_n^{\mathrm{SL}}$ and corresponding to the empirical minimizer. They provide bounds of the form $O\left(\frac{B^2}{\mu n}\right)$, which, up to a multiplicative factor proportional to κ, is the same as ours in Equation (11). Such multiplicative factor could be avoided by considering an optimization algorithm that uses the training samples several times, therefore minimizing effectively the empirical risk (see the discussion below Proposition 4). Such analysis could however necessitate the introduction of a notion of stability w.r.t. the optimization algorithm and is currently kept for future work. Finally, note that our lower bound demonstrates the optimality of the state-of-the-art upper bound in $O\left(\frac{B^2}{\mu n}\right)$, a result which was not provided by the aforementioned works.

Table 1: Our upper and lower bounds on the minimax excess risk $\varepsilon_{\rm sc}(\mathcal{G}, \mathsf{O})$ in several learning scenarios, and up to multiplicative universal constants whose values are available in the appendix.

SCENARIO	LOWER BOUND	UPPER BOUND
Supervised ($\mathcal{G}_{\rm Bnd}$)	$\frac{B^2}{\mu n}$	$\frac{\kappa B^2}{\mu n}$
Transfer ($\mathcal{G}_{\rm Bnd}$)	$\frac{B^2}{\mu}\left(d_{\rm TV}(\mathcal{D},\mathcal{D}')^2 + \frac{1}{n}\right)$	$\frac{B^2}{\mu}\left(d_{\rm TV}(\mathcal{D},\mathcal{D}')^2 + \frac{\kappa}{n}\right)$
Federated ($\mathcal{G}_{\rm Bnd}$)	unknown	$\inf_{q\in\mathbb{R}^m} \frac{B^2}{\mu}\left(d_{\rm TV}(\mathcal{D},\mathcal{D}_q)^2 + \sum_{i=1}^m \frac{\kappa q_i^2}{n_i}\right)$
Robust ($\mathcal{G}_{\rm Bnd}$)	$\frac{B^2}{\mu}\left(\eta^2 + \frac{1}{n}\right)$	$\frac{B^2}{\mu}\left(\eta^2 + \frac{\kappa}{n}\right)$
Robust ($\mathcal{G}_{\rm Lip}$)	unknown	$\frac{B^2}{\mu}\operatorname{var}(\xi)\left(\eta + \frac{\kappa}{n}\right)$
Fixed data ($\mathcal{G}_{\rm Bnd}$)	$\frac{B^2}{\mu}\left(1 - \mathbb{P}_{\mathcal{D}}\left(\{\xi_i'\}_{i\llbracket 1,n\rrbracket}\right)\right)^2$	$\frac{B^2}{\mu}\left(1 - \mathbb{P}_{\mathcal{D}}\left(\{\xi_i'\}_{i\llbracket 1,n\rrbracket}\right)\right)^2$
Fixed data ($\mathcal{G}_{\rm Lip}$)	$\frac{B^2}{\mu}\mathbb{E}\left[\min_i \|\xi - \xi_i\|\right]^2$	$\frac{B^2}{\mu}\mathbb{E}\left[\min_i \|\xi - \xi_i\|\right]^2$

4.2 Transfer learning

We now turn to the Transfer Learning (TL) oracle, defined as

$$\mathsf{O}_n^{\rm TL}(g) = (g(\xi_1'), \dots, g(\xi_n'))\,,$$

where $\xi_i' \sim \mathcal{D}'$ are i.i.d. random variables and $\mathcal{D}' \neq \mathcal{D}$. This oracle typically encompasses applications in transfer learning such as domain adaptation (Ben-David et al., 2006), where a task is learnt on a training dataset $\mathcal{D}'$ that differs from the test distribution $\mathcal{D}$.

Proposition 8. *We have*

$$\sigma_{\mathcal{G},2}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n^{\rm TL})^2 \leq d_{\mathcal{G}}(\mathcal{D},\mathcal{D}')^2 + \frac{\|\mathsf{V}_{\mathcal{D}'}\|_{\mathcal{G}}}{n}\,, \qquad (12)$$

where $d_{\mathcal{G}}(\mathcal{D},\mathcal{D}') = \sup_{g\in\mathcal{G}} \|\mathsf{E}_{\mathcal{D}}(g) - \mathsf{E}_{\mathcal{D}'}(g)\|$ is an integral probability metric (IPM).

The Integral Probability Metrics $d_{\mathcal{G}}(\mathcal{D},\mathcal{D}')$ for the function classes $\mathcal{G}_{\rm Bnd}$, $\mathcal{G}_{\rm Lip}$, and $\mathcal{G}_{\rm Aff}$ defined in Section 2.1 give respectively, the total variation distance, the Wasserstein distance, and the distance between expectations $\|\mathbb{E}_{\mathcal{D}}[\xi] - \mathbb{E}_{\mathcal{D}'}[\xi']\|$ in $\mathbb{R}^D$. From Eq. (12), note also that $\sigma_{\mathcal{G},2}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n^{\rm TL})^2$ is of the form specified in Proposition 5. Finally, the case $\mathcal{G}_{\rm Bnd}$ also provides a lower bound on the minimax estimation error.

Proposition 9. *Assume that $\mathcal{D} \ll \mathcal{D}'$ and $\forall c \in [0,1], \exists q \in \mathbb{R}$ s.t. $\mathbb{P}_{\mathcal{D}'}\left(\frac{d\mathcal{D}}{d\mathcal{D}'}(\xi') \geq q\right) = c$. Then, the minimax estimation error $\sigma_{\mathcal{G}_{\rm Bnd},2}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n^{\rm TL})^2$ is*

$$\Theta\left(B^2\left(d_{\rm TV}(\mathcal{D},\mathcal{D}')^2 + \frac{1}{n}\right)\right)\,, \qquad (13)$$

where $d_{\rm TV}$ is the Total Variation distance (Appendix B). Using Proposition 5, we obtain upper and lower bounds on the minimax excess risk that are within a multiplicative factor proportional to κ (see Table 1).

Comparison with the literature. Our bound can be put into perspective with the generalization bounds derived by the Domain Adaptation (DA) community. For instance, Ben-David et al. (2006); Blitzer et al. (2007) provide excess risk bounds for DA of the form $O\left(\sqrt{d_{\rm VC}/n} + d_{\mathcal{H}}(\mathcal{D},\mathcal{D}')\right)$ for algorithms learning from n samples drawn from a training data distribution $\mathcal{D}'$, on a test distribution $\mathcal{D}$, for a hypothesis class $\mathcal{H}$ of VC dimension $d_{\rm VC}$. From Proposition 8, we can prove (see Table 1), that in the bounded variation setting, the minimax excess risk can be upper bounded by $\varepsilon_{\rm sc}(\mathcal{G}_{\rm Bnd}, \mathsf{O}_n^{\rm TL}) = \mathcal{O}(\frac{B^2}{\mu}\left(d_{\rm TV}(\mathcal{D},\mathcal{D}')^2 + \frac{\kappa}{n}\right))$. At first sight we could conclude that our bound is always better than the first one since it exhibits a fast rate with respect to n and since $d_{\rm VC}$ is significantly larger (it can be infinite) than the constants of our bound. However, these two bounds cannot directly be compared as the setups are not perfectly matching. In particular, we are able to obtain a fast rate component in $\mathcal{O}(1/n)$ thanks to the strong convexity assumption, a setup which, to the best of our knowledge, was not explicitly considered in previous analyses (see Redko et al. (2020) for a survey on the theoretical guarantees of DA). A more detailed discussion on the difference between our distance on the gradient $d_{\mathcal{G}}(\mathcal{D},\mathcal{D}')$ and that of prior works on the function value $d_{\mathcal{H}}(\mathcal{D},\mathcal{D}')$ is available in Appendix C.

4.3 Federated learning

We now consider a setting in which m local agents are willing to collaborate in order to minimize their shared (or sometimes personal) excess risk, a setup known as (Personalized) Federated Learning (FL) (Kairouz et al., 2021). Let $(\mathcal{D}_i)_{i\in\llbracket 1,m\rrbracket}$ be a set of *local* distributions, and let $\mathsf{O}_n^{\rm FL}(g,z) = (g(\xi_j^i))_{i\in\llbracket 1,m\rrbracket, j\in\llbracket 1,n_i\rrbracket}$ for $\xi_j^i \sim \mathcal{D}_i$ i.i.d. random variables (n_i samples from agent i). We have the following minimax excess risk upper bounds that extend previous results (Even et al.,

2022; Ding and Wang, 2022), for the (P)FL oracle and objective (where the objective distribution $\mathcal{D}$ is $\mathcal{D} = \sum_i p_i \mathcal{D}_i$ for FL and $\mathcal{D} = \mathcal{D}_1$ for PFL).

Proposition 10. *We have*

$$\sigma_{\mathcal{G},2}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n^{\mathrm{FL}})^2 \leq \inf_{q \in \mathbb{R}^m} d_{\mathcal{G}}\big(\mathcal{D}, \mathcal{D}_q\big)^2 + \sum_{i=1}^m \frac{q_i^2 \|\mathsf{V}_{\mathcal{D}_i}\|_{\mathcal{G}}}{n_i} \,,$$

where $\mathcal{D}_q = \sum_{i=1}^m q_i \mathcal{D}_i$.

Intuitively, this bound allows to trade bias on the target distribution (first term) with variance of the local gradient (second term). Note that $\mathcal{D}_q$ may not be a probability measure, as the weights q_i are not necessarily positive and summing to 1.

4.4 Robust learning

We now consider a setting in which a fraction η of the data points may be arbitrarily corrupted (Klivans et al., 2018). To simplify the analysis, we will assume that these outliers are drawn according to an unknown (potentially very bad) distribution $\mathcal{D}_o$. The oracle is thus defined as

$$\mathsf{O}_n^{\mathrm{RL}}(g, z) = \big(g(\xi_1'), \dots, g(\xi_n')\big) \,, \tag{14}$$

where $\xi_i' \sim (1-\eta)\mathcal{D} + \eta \mathcal{D}_o$ are i.i.d. random variables. Note that this setting can be considered as a particular case of transfer learning. However, our objective is to obtain bounds that do not depend on the outlier distribution $\mathcal{D}_o$, and we thus focus on distant outliers such that $d_{\mathrm{TV}}\big(\mathcal{D}, \mathcal{D}_o\big) = 1$ (highest possible value for the total variation). First, the bounded variation setting is here very simple, as the ouliers cannot perturb the estimation to a large degree. In such a case, applying Proposition 8 where $\mathcal{D}'$ is the corrupted training dataset gives the upper bound:

$$\sigma_{\mathcal{G}_{\mathrm{Bnd}},2}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n^{\mathrm{RL}})^2 = \Theta\left(B^2\left(\eta^2 + \frac{1}{n}\right)\right) . \tag{15}$$

In the more challenging (and realistic) case of Lipschitz gradients w.r.t. the data points, the outliers can reach very large gradient values and thus completely break the average. To avoid this issue, we can use the robust mean estimation algorithm in Steinhardt et al. (2018, Algorithm 1) as estimator of the gradient of the population risk. This gives the following upper bound:

$$\sigma_{\mathcal{G}_{\mathrm{Lip}},2}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n^{\mathrm{RL}})^2 \leq cB^2 \operatorname{var}(\xi)\left(\eta + \frac{1}{n}\right) , \tag{16}$$

where c is a universal constant, $\eta \leq 1/4$ is the fraction of outliers, n is the total number of samples and $\operatorname{var}(\xi)$ is the variance of the true data distribution (without the outliers).

4.5 Learning from fixed data-points

In supervised learning, the i.i.d. assumption on the training dataset is key to obtain fast convergence w.r.t. the number of samples. However, training data-points are sometimes imposed and predetermined, for example following a pattern such as once every day or year for temporal data, or on a 2d grid for geophysical data (e.g. weather forecasts). In such a case, the minimax excess risk will depend on the distance between this training data and the target distribution. We thus consider the oracle defined as

$$\mathsf{O}_n^{\mathrm{FD}}(g, z) = \big(g(\xi_1'), \dots, g(\xi_n')\big) \,, \tag{17}$$

where $(\xi_1', \dots, \xi_n')$ are fixed prior to the optimization. As the oracle is deterministic, Corollary 1 allows to obtain the exact value of the minimax excess risk.

Proposition 11. *We have*

$$\varepsilon_{\mathrm{sc}}(\mathcal{G}_{\mathrm{Bnd}}, \mathsf{O}_n^{\mathrm{FD}}) = \frac{2B^2}{\mu}\left(1 - \mathbb{P}_{\mathcal{D}}\left(\{\xi_i'\}_{i[\![1,n]\!]}\right)\right)^2 \tag{18}$$

and

$$\varepsilon_{\mathrm{sc}}(\mathcal{G}_{\mathrm{Lip}}, \mathsf{O}_n^{\mathrm{FD}}) = \frac{B^2}{2\mu}\mathbb{E}\left[\min_i \|\xi - \xi_i'\|\right]^2 . \tag{19}$$

As $\mathcal{G}_{\mathrm{Bnd}}$ does not assume any local regularity w.r.t. data, knowing the value of the gradient on the data points does not provide any information on the gradient on the rest of the distribution. However, the Lipschitz assumption allows for smaller minimax excess risk that tends to 0 as the number of samples n tends to $+\infty$.

5 CONCLUSION

In this paper, we introduced a novel unified framework for the minimax excess risk control of a large panel of statistical learning problems. We focused on first-order optimization methods with data-dependent gradient oracles and showed, thanks to the new notion of *best approximation error*, that what matters is the ability of the given gradient oracle to approximate the true gradient of the population risk. Thanks to our general framework that encompasses numerous applications, we showed that this notion leads to sharp minimax excess risk bounds in most considered cases.

Our work focuses on specific regularity assumptions and applications due to lack of space and for clarity of exposition; we believe that our promising results and framework extend to other classical regularity assumption sets, and to other applications mentioned in our paper, which we leave for future work.

Acknowledgements

This work was supported by the French government managed by the Agence Nationale de la Recherche (ANR) through France 2030 program with the reference ANR-23-PEIA-005 (REDEEM project). It was also funded in part by the Groupe La Poste, sponsor of the Inria Foundation, in the framework of the FedMalin Inria Challenge. Laurent Massoulié was supported by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR19-P3IA-0001 (PRAIRIE 3IA Institute).

References

Agarwal, A., Bartlett, P. L., Ravikumar, P., and Wainwright, M. J. (2012). Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249.

Alquier, P. (2024). User-friendly introduction to pac-bayes bounds. *Foundations and Trends® in Machine Learning*, 17(2):174–303.

Amir, D. (1984). Best simultaneous approximation (chebyshev centers). In *International Series of Numerical Mathematics / Internationale Schriftenreihe zur Numerischen Mathematik / Série internationale d'Analyse numérique*, pages 19–35. Birkhäuser Basel.

Amir, I., Koren, T., and Livni, R. (2021). Sgd generalizes better than gd (and regularization doesn't help). In *Conference on Learning Theory*, pages 63–92. PMLR.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. (2023). Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199:165–214.

Bach, F. (2021). *Learning Theory from First Principles*. Draft book.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537.

Bartlett, P. L. and Mendelson, S. (2003). Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3(null):463–482.

Bassily, R., Feldman, V., Guzmán, C., and Talwar, K. (2020). Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33:4381–4391.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. (2006). Analysis of representations for domain adaptation. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Wortman, J. (2007). Learning bounds for domain adaptation. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. (1989). Learnability and the vapnik-chervonenkis dimension. *J. ACM*, 36(4):929–965.

Bottou, L. and Bousquet, O. (2007). The tradeoffs of large scale learning. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Bousquet, O., Boucheron, S., and Lugosi, G. (2004). Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, pages 169–207. Springer Berlin Heidelberg.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526.

Bubeck, S. (2015). Convex optimization: Algorithms and complexity.

Devolder, O., Glineur, F., and Nesterov, Y. (2013). First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1–2):37–75.

Ding, S. and Wang, W. (2022). Collaborative learning by detecting collaboration partners. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Drori, Y. and Taylor, A. (2022). On the oracle complexity of smooth strongly convex minimization. *J. Complex.*, 68(C).

Even, M., Massoulié, L., and Scaman, K. (2022). On sample optimality in personalized collaborative and federated learning. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Hardt, M., Recht, B., and Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konecný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210.

Klivans, A., Kothari, P. K., and Meka, R. (2018). Efficient algorithms for outlier-robust regression.

Kuzborskij, I. and Lampert, C. (2018). Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 2815–2824. PMLR.

Lei, Y. and Ying, Y. (2020a). Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819. PMLR.

Lei, Y. and Ying, Y. (2020b). Sharper generalization bounds for learning with gradient-dominated objective functions. In *International Conference on Learning Representations*.

Lin, J., Camoriano, R., and Rosasco, L. (2016). Generalization properties and implicit regularization for multiple passes sgm. In *International Conference on Machine Learning*, pages 2340–2348. PMLR.

London, B. (2017). A pac-bayesian analysis of randomized learning with application to stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30.

McAllester, D. A. (1998). Some pac-bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, page 230–234, New York, NY, USA. Association for Computing Machinery.

Mukherjee, S., Niyogi, P., Poggio, T., and Rifkin, R. (2006). Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1-3):161–193.

Neu, G., Dziugaite, G. K., Haghifam, M., and Roy, D. M. (2021). Information-theoretic generalization bounds for stochastic gradient descent. In *Conference on Learning Theory*, pages 3526–3545. PMLR.

Polyanskiy, Y. and Wu, Y. (2022). Information theory: From coding to learning. *Book draft*.

Redko, I., Morvant, E., Habrard, A., Sebban, M., and Bennani, Y. (2020). A survey on domain adaptation theory: learning bounds and theoretical guarantees. *arXiv preprint arXiv:2004.11829*.

Schliserman, M. and Koren, T. (2022). Stability vs implicit bias of gradient methods on separable data and beyond. In *Conference on Learning Theory*, pages 3380–3394. PMLR.

Schliserman, M. and Koren, T. (2023). Tight risk bounds for gradient descent on separable data. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 68749–68759. Curran Associates, Inc.

Sridharan, K., Shalev-shwartz, S., and Srebro, N. (2008). Fast rates for regularized objectives. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Steinhardt, J., Charikar, M., and Valiant, G. (2018). Resilience: A criterion for learning in the presence of arbitrary outliers. In Karlin, A. R., editor, *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, volume 94 of *LIPIcs*, pages 45:1–45:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*. Springer New York.

Vapnik, V. N. and Chervonenkis, A. Y. (2015). On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of Complexity*, pages 11–30. Springer International Publishing.

Xu, A. and Raginsky, M. (2017). Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 30.

Zhang, Y., Zhang, W., Bald, S., Pingali, V., Chen, C., and Goswami, M. (2022). Stability of sgd: Tightness analysis and improved bounds. In *Uncertainty in artificial intelligence*, pages 2364–2373. PMLR.

Zhou, Y., Liang, Y., and Zhang, H. (2018). Generalization error bounds with probabilistic guarantee for sgd in nonconvex optimization. *arXiv preprint arXiv:1802.06903*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Link with the Polyak-Łojasiewicz condition

Recall the definition of the Polyak-Łojasiewicz condition, which is weaker than strong convexity and can be satisfied by non-convex functions.

Definition 5. (Polyak-Łojasiewicz) Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, and $\mu > 0$. We say that f is μ-Polyak-Łojasiewicz (μ-PL for short) if it is bounded from below, and if for all $x \in \mathbb{R}^d$

$$f(x) - \inf f \leq \frac{1}{2\mu}\|\nabla f(x)\|^2 \,. \tag{20}$$

Importantly, μ-strongly convex functions are also μ-PL and we therefore have $\mathcal{F}_{\rm sc}(\mathcal{G}, \mathcal{D}, \mu, L) \subset \mathcal{F}_{\rm pl}(\mathcal{G}, \mathcal{D}, \mu, L)$ (abbreviated to $\mathcal{F}_{\rm pl}$ below), the set of μ-PL and L-smooth objective functions $\mathcal{L}(x) = \mathbb{E}_{\xi\sim\mathcal{D}}\left[\ell(x, \xi)\right]$ such that $\forall x \in \mathbb{R}^d$, $\nabla \ell_x \in \mathcal{G}$. As an immediat consequence, we also have the relation $\varepsilon_{\rm sc}(\mathcal{G}, \mathsf{O}) \leq \varepsilon_{\rm pl}(\mathcal{G}, \mathsf{O})$ and all the lower bounds presented for $\varepsilon_{\rm sc}(\mathcal{G}, \mathsf{O})$ in the main paper are also immediately valid for $\varepsilon_{\rm pl}(\mathcal{G}, \mathsf{O})$.

To make all our results valid for both set of functions $\mathcal{F}_{\rm sc}$ and $\mathcal{F}_{\rm pl}$, all the upper bounds derived in the paper are actually proved, in the following sections, for μ-PL objective functions.

B Le Cam's distance between probability distributions

First, we recall the definition of two standard divergences between probability distributions. Let P, Q be two probability distributions such that $dP(dx) = p(x)d\mu(x)$ and $dQ(x) = q(x)d\mu(x)$ for some common dominating measure μ.

- **f-divergences:** Let $f : \mathbb{R}_+ \mapsto \mathbb{R} \cup \{+\infty\}$ be a convex function such that $f(1) = 0$ and $\lim_{t\to 0^+} f(t) = f(0)$. The f-divergence between P and Q is defined as

$$D_f(P, Q) = \int f\left(\frac{p(x)}{q(x)}\right) q(x)d\mu(x) \,. \tag{21}$$

- **Total variation:** The total variation distance is defined as

$$d_{\rm TV}(P, Q) = \frac{1}{2}\int |p(x) - q(x)|d\mu(x) \,. \tag{22}$$

- **Kullback-Leibler:** The Kullback-Leibler divergence is defined as

$$d_{\rm KL}(P, Q) = \int \ln\left(\frac{p(x)}{q(x)}\right) p(x)d\mu(x) \,. \tag{23}$$

Note that both $d_{\rm TV}$ and $d_{\rm KL}$ are f-divergences with, respectively, $f(t) = |t - 1|/2$ and $f(t) = t\ln(t)$. Below, we recall a useful property of f-divergences that is going to be used later.

Property 1. $D_f = D_h$ if and only if $f(t) = h(t) + c(t - 1)$ for some constant $c \in \mathbb{R}$.

We now provide a definition for Le Cam's distance.

Definition 6. Let P, Q be two probability distributions such that $dP(x) = p(x)\mu(x)$ and $dQ(x) = q(x)d\mu(x)$ for some common dominating measure μ. We denote as *Le Cam's distance* between P and Q the quantity

$$d_{\rm LC}(P, Q) = \frac{1}{2}\int \frac{(p(x) - q(x))^2}{p(x) + q(x)}d\mu(x) \,. \tag{24}$$

Another definition for $d_{\rm LC}$ is the f-divergence obtained with the (convex) function $f(t) = \frac{(1-t)^2}{2(1+t)}$, or equivalently thanks to Property 1, with the function $h(t) = f(t) - \frac{1}{2}(t - 1) = \frac{1-t}{1+t}$.

By definition, $d_{\rm LC}$ is symmetric, and we have the following relationship between Le Cam's distance and other standard f-divergences, thanks again to Property 1.

Lemma 1. *For any P, Q, we have*

$$d_{\mathrm{LC}}(P,Q) \le d_{\mathrm{TV}}(P,Q) \quad \text{and} \quad d_{\mathrm{LC}}(P,Q) \le d_{\mathrm{KL}}(P,Q)\,. \tag{25}$$

Proof. A simple functional analysis gives $\frac{1-t}{1+t} + \frac{t-1}{2} \le \frac{1}{2}|t-1|$ for $t \ge 0$, thus directly implying the first inequality. For the second, we have $\frac{1-t}{1+t} \le -\ln(t) + \frac{t-1}{2}$ for $t \ge 0$ and, as the f-divergence with $f(t) = -\ln(t)$ corresponds to the reverse KL and d_{LC} is symmetric, we have $d_{\mathrm{LC}}(P,Q) = d_{\mathrm{LC}}(Q,P) \le d_{\mathrm{KL}}(P,Q)$. □

The link with the Kullback-Leibler divergence will be useful to derive proofs in the i.i.d. oracle regime. We now provide a proof of Le Cam's two point method adapted to our setting.

Proof of Proposition 2. Let $g_1, g_2 \in \mathcal{G}$ be two functions in the functions class. Then,

$$\begin{aligned}
\sigma_{\mathcal{G},2}\left(\mathsf{E}_{\mathcal{D}}|\mathsf{O}\right)^2 &= \inf_{\varphi\in\mathbb{M}(\mathcal{Y},\mathcal{Z})} \sup_{g\in\mathcal{G}} \mathbb{E}_z\left[\|\mathsf{E}_{\mathcal{D}}(g) - \varphi\circ\mathsf{O}(g,z)\|^2\right] \\
&\ge \inf_{\varphi\in\mathbb{M}(\mathcal{Y},\mathcal{Z})} \sup_{g\in\{g_1,g_2\}} \mathbb{E}_z\left[\|\mathsf{E}_{\mathcal{D}}(g) - \varphi\circ\mathsf{O}(g,z)\|^2\right] \\
&\ge \inf_{\varphi\in\mathbb{M}(\mathcal{Y},\mathcal{Z})} \mathbb{E}_{G,z}\left[\|\mathsf{E}_{\mathcal{D}}(G) - \varphi\circ\mathsf{O}(G,z)\|^2\right]\,,
\end{aligned} \tag{26}$$

where $G = Bg_1 + (1-B)g_2$ and $B \sim \mathcal{B}(1/2)$ is a Bernoulli random variable of parameter $1/2$. The infimum over measurable functions φ is now attained for the conditional expectation $\mathbb{E}\left[\mathsf{E}_{\mathcal{D}}(G)|\mathsf{O}(G,z)\right]$, and a simple calculation gives

$$\mathbb{E}\left[\mathsf{E}_{\mathcal{D}}(G)|\mathsf{O}(G,z)\right] = \mathsf{E}_{\mathcal{D}}(g_1)\frac{p_1(O(G,z))}{p_1(O(G,z)) + p_2(O(G,z))} + \mathsf{E}_{\mathcal{D}}(g_2)\frac{p_2(O(G,z))}{p_1(O(G,z)) + p_2(O(G,z))}\,, \tag{27}$$

where p_1, p_2 are the Radon-Nykodym densities of, respectively, $\mathsf{O}(g_1,z)$ and $\mathsf{O}(g_2,z)$ w.r.t. to a common dominating measure μ. Combining the two previous equaitons, we get

$$\begin{aligned}
\sigma_{\mathcal{G},2}\left(\mathsf{E}_{\mathcal{D}}|\mathsf{O}\right)^2 &\ge \mathbb{E}_{G,z}\left[\|\mathsf{E}_{\mathcal{D}}(G) - \tfrac{\mathsf{E}_{\mathcal{D}}(g_1)\,p_1(O(G,z)) + \mathsf{E}_{\mathcal{D}}(g_2)\,p_2(O(G,z))}{p_1(O(G,z)) + p_2(O(G,z))}\|^2\right] \\
&= \tfrac{1}{2}\mathbb{E}_z\left[\|(\mathsf{E}_{\mathcal{D}}(g_1) - \mathsf{E}_{\mathcal{D}}(g_2))\tfrac{p_2(O(g_1,z))}{p_1(O(g_1,z)) + p_2(O(g_1,z))}\|^2 + \|(\mathsf{E}_{\mathcal{D}}(g_1) - \mathsf{E}_{\mathcal{D}}(g_2))\tfrac{p_1(O(g_2,z))}{p_1(O(g_2,z)) + p_2(O(g_2,z))}\|^2\right] \\
&= \tfrac{\|\mathsf{E}_{\mathcal{D}}(g_1) - \mathsf{E}_{\mathcal{D}}(g_2)\|^2}{2}\left(\int \tfrac{p_2(o)^2}{(p_1(o)+p_2(o))^2} p_1(o)d\mu(o) + \int \tfrac{p_1(o)^2}{(p_1(o)+p_2(o))^2} p_2(o)d\mu(o)\right) \\
&= \tfrac{\|\mathsf{E}_{\mathcal{D}}(g_1) - \mathsf{E}_{\mathcal{D}}(g_2)\|^2}{2}\int \tfrac{p_1(o)p_2(o)^2 + p_2(o)p_1(o)^2}{(p_1(o)+p_2(o))^2} d\mu(o) \\
&= \tfrac{\|\mathsf{E}_{\mathcal{D}}(g_1) - \mathsf{E}_{\mathcal{D}}(g_2)\|^2}{2}\int \tfrac{p_1(o)p_2(o)}{p_1(o)+p_2(o)} d\mu(o) \\
&= \tfrac{\|\mathsf{E}_{\mathcal{D}}(g_1) - \mathsf{E}_{\mathcal{D}}(g_2)\|^2}{4}\left(1 - \tfrac{1}{2}\int \tfrac{(p_1(o)-p_2(o))^2}{p_1(o)+p_2(o)} d\mu(o)\right) \\
&= \tfrac{\|\mathsf{E}_{\mathcal{D}}(g_1) - \mathsf{E}_{\mathcal{D}}(g_2)\|^2}{4}\Big(1 - d_{\mathrm{LC}}(O(g_1,z), O(g_2,z))\Big)
\end{aligned} \tag{28}$$

□

C Comparison between distances on gradients and function values

We now discuss the differences between $d_{\mathcal{H}}(\mathcal{D},\mathcal{D}')$ where $\mathcal{H} \supset \{\xi \mapsto \ell(x,\xi) \ : \ x \in \mathbb{R}^d\}$ contains the values of the loss, and $d_{\mathcal{G}}(\mathcal{D},\mathcal{D}')$, where $\mathcal{G} \supset \{\xi \mapsto \nabla_x \ell(x,\xi) \ : \ x \in \mathbb{R}^d\}$ contains gradients of the loss. First, note that discrepancies in loss value are usually used to control the generalisation error on the iterates, as $|\mathbb{E}[\ell(x^{(t)}_{\mathsf{A}[\mathsf{O}]},\xi)] - \mathbb{E}[\ell(x^{(t)}_{\mathsf{A}[\mathsf{O}]},\xi')]| \le d_{\mathcal{H}}(\mathcal{D},\mathcal{D}')$ as long as $\xi \mapsto \ell(x^{(t)}_{\mathsf{A}[\mathsf{O}]},\xi) \in \mathcal{H}$. Unfortunately, these discrepancies are infinite in our setting, as we now show: If $\mathcal{H}_1 = \{\xi \mapsto \ell(x,\xi) \ : \ x \in \mathbb{R}^d \text{ and } \mathbb{E}[\ell(\cdot,\xi)] \in \mathcal{F}_{sc}(\mathcal{G},\mathcal{D},\mu,L)\}$ is the set of strongly-convex and smooth loss functions considered in this paper, then choosing $\ell^g(x,\xi) = \frac{\mu}{2}\|x\|^2 + \langle g(\xi), x\rangle$ gives

$$d_{\mathcal{H}_1}(\mathcal{D},\mathcal{D}') \ge \sup_{x\in\mathbb{R}^d} |\mathbb{E}[\ell^g(x,\xi) - \ell^g(x,\xi')]| = \sup_{x\in\mathbb{R}^d} |\langle \mathbb{E}[g(\xi)] - \mathbb{E}[g(\xi')], x\rangle| = +\infty\,, \tag{29}$$

as soon as $\mathbb{E}[g(\xi)] \ne \mathbb{E}[g(\xi')]$ (i.e. $d_{\mathcal{G}}(\mathcal{D},\mathcal{D}') > 0$). The discrepancy in function value is thus unsuited to the strongly convex and smooth setting without additional assumptions on the domain of x or boundedness of the loss. However, one could argue that the difference in loss is only necessary on the algorithm's output $x_{\mathsf{A}[\mathsf{O}]}$ (or equivalently the algorithm's iterates $x^{(t)}_{\mathsf{A}[\mathsf{O}]}$) instead of the whole space. Unfortunately, this quantity is also

infinite, as, $\forall g \in \mathcal{G}$ and $\forall c \in \mathbb{R}^d$, the function $\ell^{g,c}(x,\xi) = \frac{\mu}{2}\|x\|^2 + \langle g(\xi), x - c\rangle$ is μ-strongly convex and μ-smooth, and its gradient belongs to $\mathcal{G}$ (by translation invariance of $\mathcal{G}$, see Assumption 1). Note that we have only added a constant term w.r.t. x which is thus invisible to algorithms that rely on the gradient. Thus, the constant c has no impact on the algorithm's output $x_{\mathsf{A}[\mathsf{O}]}$, and, for $\mathcal{H}_2 = \{\xi \mapsto \ell(x_{\mathsf{A}[\mathsf{O}]}, \xi) \ : \ \mathbb{E}[\ell(\cdot,\xi)] \in \mathcal{F}_{sc}(\mathcal{G},\mathcal{D},\mu,L)\}$, we have, as soon as $\mathbb{E}[g(\xi)] \neq \mathbb{E}[g(\xi')]$,

$$d_{\mathcal{H}_2}(\mathcal{D},\mathcal{D}') \geq \sup_{c\in\mathbb{R}^d} |\mathbb{E}[\ell^{g,c}(xx_{\mathsf{A}[\mathsf{O}]},\xi) - \ell^{g,c}(x_{\mathsf{A}[\mathsf{O}]},\xi')]| = \sup_{c\in\mathbb{R}^d} |\langle \mathbb{E}[g(\xi)] - \mathbb{E}[g(\xi')], x_{\mathsf{A}[\mathsf{O}]} - c\rangle| = +\infty\,. \tag{30}$$

The same result holds if one replaces $x_{\mathsf{A}[\mathsf{O}]}$ by $\arg\min_x \mathbb{E}[\ell(x,\xi')]$, $\arg\min_x \mathbb{E}[\ell(x,\xi)]$, or any set independent of c.

D Proofs of Section 3

D.1 Proof of our lower bound

We start by proving our first result (Proposition 1), the lower bound on the minimax excess risk that exhibits $\sigma_{\mathcal{G},2}\left(\mathsf{E}_{\mathcal{D}}|\mathsf{O}\right)$ as limiting factor. This result is a direct consequence of the following lemma:

Lemma 2. *Let $\varepsilon > 0$, $\mathcal{G} \subset \mathbb{F}(\Xi,\mathbb{R}^d)$ be a function space and O a data-dependent oracle verifying Assumption 1. Then, for any optimization algorithm $\mathsf{A} \in \mathcal{A}_{\text{rand}}$, there exists an objective function $\mathcal{L} \in \mathcal{F}_{sc}(\mathcal{G},\mathcal{D},\mu,L)$ such that*

$$\mathbb{E}\left[\mathcal{L}(x_{\mathsf{A}[\mathsf{O}]}) - \inf_{x\in\mathbb{R}^d} \mathcal{L}(x)\right] \geq \frac{\sigma_{\mathcal{G},2}\left(\mathsf{E}_{\mathcal{D}}|\mathsf{O}\right)^2}{2\mu} - \varepsilon\,.$$

Proof. For any $g \in \mathcal{G}$, let

$$\ell^g(x,\xi) = \frac{\mu}{2}\|x\|^2 + \langle x, g(\xi)\rangle\,. \tag{31}$$

First, note that ℓ^g is μ-strongly convex and μ-smooth w.r.t. x (and $\mu \leq L$), and $\nabla_x \ell^g(x,\xi) = \mu x + g(\xi) \in \mathcal{G}$ by stability of $\mathcal{G}$ by translation. By Assumption 1, there exists a measurable function $\varphi : \mathcal{G}\times\mathbb{R}^d \to \mathcal{O}$ such that $\mathsf{O}(\nabla\ell^g_{x^{(t)}_{\mathsf{A}[\mathsf{O}]}}, z) = \varphi(\mathsf{O}(g,z), \mu x^{(t)}_{\mathsf{A}[\mathsf{O}]})$, and we now show that the output of the algorithm $x_{\mathsf{A}[\mathsf{O}]}$ is a measurable function of r and $\mathsf{O}(g,z)$.

Lemma 3. *For any optimization algorithm A of Definition 3, there exists a function ψ_{A} such that the ouptut $x_{\mathsf{A}[\mathsf{O}]}$ of A applied to any objective function ℓ^g defined in Eq. (31) for $g \in \mathcal{G}$ is*

$$\mu\, x_{\mathsf{A}[\mathsf{O}]} \quad = \quad \psi_{\mathsf{A}}(r, \mathsf{O}(g,z))\,. \tag{32}$$

Proof. First, note that $\mu x^{(0)}_{\mathsf{A}[\mathsf{O}]} = \mu q^{(0)}(r)$ is a measurable function of r. By induction over $t \geq 0$, there exists measurable functions $\psi^{(t)}_{x,\mathsf{A}} : \mathcal{R}\times\mathcal{O} \to \mathbb{R}^d$ such that $\mu x^{(t)}_{\mathsf{A}[\mathsf{O}]} = \psi^{(t)}_{x,\mathsf{A}}(r,\mathsf{O}(g,z))$ and $\psi^{(t)}_{s,\mathsf{A}} : \mathcal{R}\times\mathcal{O} \to \{0,1\}$ such that $s^{(t)}_{\mathsf{A}[\mathsf{O}]} = \psi^{(t)}_{s,\mathsf{A}}(r,\mathsf{O}(g,z))$. Without loss of generality, we assume that $s^{(t)}_{\mathsf{A}[\mathsf{O}]} = 1$ only once, as we can replace $s^{(t)}_{\mathsf{A}[\mathsf{O}]}$ on all iterations after the first 1 by 0. Thus, we have

$$\mu\, x_{\mathsf{A}[\mathsf{O}]} \quad = \quad \mu\sum_{t=0}^{+\infty} s^{(t)}_{\mathsf{A}[\mathsf{O}]} x^{(t)}_{\mathsf{A}[\mathsf{O}]} \quad = \quad \sum_{t=0}^{+\infty} \psi^{(t)}_{s,\mathsf{A}}(r,\mathsf{O}(g,z))\psi^{(t)}_{x,\mathsf{A}}(r,\mathsf{O}(g,z))\,,$$

that is a measurable function of r and $\mathsf{O}(g,z)$ as a limit of measurable functions. □

Moreover, we have $\nabla\mathcal{L}^g(x) = \mu x + \mathbb{E}\left[g(\xi)\right]$, and $\inf_{x\in\mathbb{R}^d}\mathcal{L}^g(x) = -\|\mathbb{E}\left[g(\xi)\right]\|^2/2\mu$. This gives $\mathcal{L}^g(x_{\mathsf{A}[\mathsf{O}]}) - \inf_{x\in\mathbb{R}^d}\mathcal{L}^g(x) = \|\mu x_{\mathsf{A}[\mathsf{O}]} + \mathbb{E}\left[g(\xi)\right]\|^2/2\mu$. Now,

$$\begin{aligned}\sup_{g\in\mathcal{G}} \mathbb{E}\left[\mathcal{L}^g(x_{\mathsf{A}[\mathsf{O}]}) - \inf_{x\in\mathbb{R}^d}\mathcal{L}^g(x)\right] &= \sup_{g\in\mathcal{G}} \frac{\mathbb{E}\left[\|-\psi_{\mathsf{A}}(r,\mathsf{O}(g,z)) - \mathsf{E}_{\mathcal{D}}(g)\|^2\right]}{2\mu} \\ &\geq \frac{\sigma_{\mathcal{G},2}\left(\mathsf{E}_{\mathcal{D}}|\mathsf{O}\right)^2}{2\mu}\,,\end{aligned}$$

where the last inequality follows from Jensen's inequality on r and the definition of $\sigma_{\mathcal{G},2}\left(\mathsf{E}_{\mathcal{D}}|\mathsf{O}\right)$. □

D.2 General upper-bound and subsequent corollaries

Lemma 4. *Let $\varphi \in \mathbb{M}(\mathcal{O}, \mathbb{R}^d)$ and $\mathcal{L} \in \mathcal{F}_{pl}(\mathcal{G}, \mathcal{D}, \mu, L)$. Then, the iterates $x_0 = 0$ and $x_{t+1} = x_t - \frac{1}{L}\varphi(o_t)$ where $o_t = \mathsf{O}(\nabla \ell_{x_t}, z)$ achieve an approximation error*

$$\mathbb{E}\left[\mathcal{L}(x_t) - \inf_{x\in\mathbb{R}^d} \mathcal{L}(x)\right] \leq \Delta\rho^t + \frac{\|\mathsf{E}_{\mathcal{D}} - \varphi\circ\mathsf{O}\|_{2,\mathcal{G}}^2}{2\mu}\,,$$

where $\rho = 1 - \mu/L$ and $\Delta = \mathcal{L}(x_0) - \inf_{x\in\mathbb{R}^d}\mathcal{L}(x)$.

Proof. First, recall that, for any $x \in \mathbb{R}^d$, $\nabla\ell_x \in \mathcal{G}$ and $\mathbb{E}_{\xi\sim\mathcal{D}}\left[\nabla_x \ell(x,\xi)\right] = \mathsf{E}_{\mathcal{D}}(\nabla\ell_x)$. Thus,

$$\begin{aligned}\mathbb{E}\left[\|\nabla\mathcal{L}(x_t) - \varphi(o_t)\|^2\right] &= \mathbb{E}\left[\|\mathsf{E}_{\mathcal{D}}(\nabla\ell_{x_t}) - \varphi\circ\mathsf{O}(\nabla\ell_{x_t}, z)\|^2\right]\\ &\leq \mathbb{E}\left[\sup_{g\in\mathcal{G}}\|\mathsf{E}_{\mathcal{D}}(g) - \varphi\circ\mathsf{O}(g,z)\|^2\right]\\ &= \|\mathsf{E}_{\mathcal{D}} - \varphi\circ\mathsf{O}\|_{2,\mathcal{G}}^2\end{aligned}$$

Then, by smoothness, we have

$$\begin{aligned}\mathcal{L}(x_{t+1}) - \mathcal{L}(x_t) &\leq -\tfrac{1}{L}\langle\nabla\mathcal{L}(x_t), \varphi(o_t)\rangle + \tfrac{1}{2L}\|\varphi(o_t)\|^2\\ &= -\tfrac{1}{2L}\|\nabla\mathcal{L}(x_t)\|^2 + \tfrac{1}{2L}\|\nabla\mathcal{L}(x_t) - \varphi(o_t)\|^2\end{aligned}$$

Moreover, as $\mathcal{L}$ is μ-PL, we have

$$\|\nabla\mathcal{L}(x_t)\|^2 \geq 2\mu\left(\mathcal{L}(x_t) - \inf_{x\in\mathbb{R}^d}\mathcal{L}(x)\right).$$

Combining the two previous equations and taking the expectation gives

$$\mathbb{E}\left[\mathcal{L}(x_{t+1}) - \mathcal{L}(x_t)\right] \leq -\frac{1}{\kappa}\mathbb{E}\left[\mathcal{L}(x_t) - \inf_{x\in\mathbb{R}^d}\mathcal{L}(x)\right] + \frac{\|\mathsf{E}_{\mathcal{D}} - \varphi\circ\mathsf{O}\|_{2,\mathcal{G}}^2}{2L}\,.$$

A simple recurrence gives

$$\mathbb{E}\left[\mathcal{L}(x_t) - \inf_{x\in\mathbb{R}^d}\mathcal{L}(x)\right] \leq \left(1 - \frac{1}{\kappa}\right)^t\mathbb{E}\left[\mathcal{L}(x_0) - \inf_{x\in\mathbb{R}^d}\mathcal{L}(x)\right] + \frac{\|\mathsf{E}_{\mathcal{D}} - \varphi\circ\mathsf{O}\|_{2,\mathcal{G}}^2}{2\mu}\,.$$

□

Proof of Proposition 3. We need to select a number of steps in Lemma 4 sufficient to reduce the first term in $\Delta(1-1/\kappa)^t$ to any given precision $\varepsilon > 0$. After the first iteration, we fix the number of iterations as $T_z = \kappa\ln\left(\frac{\|\varphi(o_0)\|^2 + \|\mathsf{E}_{\mathcal{D}} - \varphi\circ\mathsf{O}\|_{2,\mathcal{G}}^2}{\varepsilon\mu}\right)$. Note that this stopping time depends only on the observation o_0 at the first iteration, and can thus be computed after this iteration. Then, we have

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}(x_{T_z}) - \inf_{x\in\mathbb{R}^d}\mathcal{L}(x)\right] &\leq \mathbb{E}\left[\left(1-\frac{1}{\kappa}\right)^{T_z}(\mathcal{L}(x_0) - \inf_{x\in\mathbb{R}^d}\mathcal{L}(x))\right]\\ &\quad + \mathbb{E}\left[\frac{1}{2L}\sum_{t=0}^{T_z-1}\left(1-\frac{1}{\kappa}\right)^{T_z-t-1}\|\nabla\mathcal{L}(x_t) - \varphi(o_t)\|^2\right]\\ &\leq \varepsilon\mu\mathbb{E}\left[\frac{\mathcal{L}(x_0) - \inf_{x\in\mathbb{R}^d}\mathcal{L}(x)}{\|\varphi(o_0)\|^2 + \|\mathsf{E}_{\mathcal{D}} - \varphi\circ\mathsf{O}\|_{2,\mathcal{G}}^2}\right]\\ &\quad + \mathbb{E}\left[\frac{1}{2L}\sum_{t=0}^{T_z-1}\left(1-\frac{1}{\kappa}\right)^{T_z-t-1}\sup_{g\in\mathcal{G}}\|\mathsf{E}_{\mathcal{D}}(g) - \varphi\circ\mathsf{O}(g,z)\|^2\right]\\ &\leq 2\varepsilon + \frac{\mathbb{E}\left[\sup_{g\in\mathcal{G}}\|\mathsf{E}_{\mathcal{D}}(g) - \varphi\circ\mathsf{O}(g,z)\|^2\right]}{2\mu}\\ &= 2\varepsilon + \frac{\|\mathsf{E}_{\mathcal{D}} - \varphi\circ\mathsf{O}\|_{2,\mathcal{G}}^2}{2\mu}\,,\end{aligned}$$

where the last inequality follows from the μ-PL condition and $\mathcal{L}(x_0) - \inf_{x\in\mathbb{R}^d}\mathcal{L}(x) \le \|\nabla\mathcal{L}(x_0)\|^2/2\mu \le (2\|\varphi(o_0)\|^2 + 2\|\mathsf{E}_{\mathcal{D}}(\nabla\ell_{x_0}) - \varphi\circ\mathsf{O}(\nabla\ell_{x_0}, z)\|^2)/2\mu$ (see the beginning of the proof in Lemma 4), leading to

$$\begin{aligned}\mathbb{E}\left[\frac{\mathcal{L}(x_0) - \inf_{x\in\mathbb{R}^d}\mathcal{L}(x)}{\|\varphi(o_0)\|^2 + \|\mathsf{E}_{\mathcal{D}} - \varphi\circ\mathsf{O}\|_{2,\mathcal{G}}^2}\right] &\le \frac{1}{2\mu}\mathbb{E}\left[\frac{2\|\varphi(o_0)\|^2 + 2\|\mathsf{E}_{\mathcal{D}}(\nabla\ell_{x_0}) - \varphi\circ\mathsf{O}(\nabla\ell_{x_0}, z)\|^2}{\|\varphi(o_0)\|^2 + \|\mathsf{E}_{\mathcal{D}} - \varphi\circ\mathsf{O}\|_{2,\mathcal{G}}^2}\right] \\ &\le \frac{1}{\mu}\mathbb{E}\left[1 + \frac{\sup_{g\in\mathcal{G}}\|\mathsf{E}_{\mathcal{D}}(g) - \varphi\circ\mathsf{O}(g,z)\|^2}{\|\mathsf{E}_{\mathcal{D}} - \varphi\circ\mathsf{O}\|_{2,\mathcal{G}}^2}\right] \\ &= \frac{2}{\mu}\,.\end{aligned}$$

Finally, taking $\varphi \in \mathbb{M}(\mathcal{O}, \mathbb{R}^d)$ such that $\|\mathsf{E}_{\mathcal{D}} - \varphi\circ\mathsf{O}\|_{2,\mathcal{G}} \le \sigma_{2,\mathcal{G}}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}) + \varepsilon$ and $\varepsilon \to 0$ gives the desired result. □

Proof of Corollary 1. If $\mathsf{O}(g,z) = \tilde{\mathsf{O}}(g)$ is independent of z, then $\sigma_{\mathcal{G},2}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}) = \sigma_{2,\mathcal{G}}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}) = \sigma_{\mathcal{G}}(\mathsf{E}_{\mathcal{D}}|\tilde{\mathsf{O}})$ as all norms are equal, and Proposition 3 immediately gives the desired result. □

E Proofs of Section 3.3

Recall that for the following proofs, the oracle is assumed to be of the form O_n (n i.i.d. observations).

Lemma 5. *For any $n \ge 1$, let φ_n be such that $\|\mathsf{E}_{\mathcal{D}} - \varphi_n\circ\mathsf{O}_n\|_{\mathcal{G},2}^2 \le \sigma_{\mathcal{G},2}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n)^2 + \varepsilon$, and $(n_1,\dots,n_T)$ be non-negative integers such that $\sum_{t<T} n_t \le n$. Then, the iterates $x_0 = 0$ and*

$$x_{t+1} = x_t - \frac{1}{L}\varphi_{n_t}(o_{t,1},\dots,o_{t,n_t})\,, \tag{33}$$

where $o_{t,k} = \mathsf{O}(\nabla\ell_{x_t}, z^{(N_t+k)})$ is a (fresh) i.i.d. observation and $N_t = \sum_{i<t} n_i$, achieve after T iterations an approximation error

$$\mathbb{E}\left[\mathcal{L}(x_T) - \mathcal{L}^*\right] \le \Delta\rho^T + \sum_{t=0}^{T-1}\frac{\sigma_t^2\rho^{T-t-1}}{2L} + \frac{\varepsilon}{2\mu}\,, \tag{34}$$

where $\sigma_t = \sigma_{\mathcal{G},2}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_{n_t})$, $\mathcal{L}^* = \inf_{x\in\mathbb{R}^d}\mathcal{L}(x)$, $\rho = 1-\mu/L$ and $\Delta = \mathcal{L}(x_0) - \mathcal{L}^*$.

Proof. As in Lemma 4, we have, using smoothness and the μ-PL condition:

$$\mathcal{L}(x_{t+1}) - \mathcal{L}(x_t) \le -\frac{1}{\kappa}\left(\mathcal{L}(x_t) - \inf_{x\in\mathbb{R}^d}\mathcal{L}(x)\right) + \frac{1}{2L}\|\nabla\mathcal{L}(x_t) - \varphi_{n_t}(o_t)\|^2\,.$$

where $o_t = \mathsf{O}(\nabla\ell_{x_t}, z^{(N_t+1)}), \dots, \mathsf{O}(\nabla\ell_{x_t}, z^{(N_t+n_t)})$. Thus, by a simple recursion:

$$\mathcal{L}(x_T) - \inf_{x\in\mathbb{R}^d}\mathcal{L}(x) \le \Delta\left(1-\frac{1}{\kappa}\right)^T + \frac{1}{2L}\sum_{t=0}^{T-1}\left(1-\frac{1}{\kappa}\right)^{T-t-1}\|\nabla\mathcal{L}(x_t) - \varphi_{n_t}(o_t)\|^2\,.$$

To conclude, we take the expectation on both sides of the inequality: since $\varphi_{n_t}(o_t)$ is independent from $(z^{(i)})_{1\le i\le N_t}$ and $\nabla\mathcal{L}(x_t)$ only depends on $(z^{(i)})_{1\le i\le N_t}$, $\nabla\mathcal{L}(x_t)$ and $\varphi_{n_t}(o_t)$ are independent and thus $\mathbb{E}\|\nabla\mathcal{L}(x_t) - \varphi_{n_t}(o_t)\|^2 \le \sigma_{\mathcal{G},2}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_{n_t})^2 + \varepsilon$, by definition of φ_{n_t}. □

Proof of Proposition 4. We first start with a warm-up phase in which we only use the first observation $\mathsf{O}(\nabla\ell_{x_t}, z^{(1)})$ for a number of steps sufficient to reduce the first term in Lemma 4 in $\Delta(1-1/\kappa)^t$ to any given precision $\varepsilon > 0$. We then fix the number of iterations of this warm-up phase (after the first iteration) as $T_z = \kappa\ln\left(\frac{\|\varphi(o_0)\|^2 + \|\mathsf{E}_{\mathcal{D}} - \varphi\circ\mathsf{O}_1\|_{2,\mathcal{G}}^2}{\varepsilon\mu}\right)$. This gives

$$\mathbb{E}\left[\mathcal{L}(x_{T_z}) - \inf_{x\in\mathbb{R}^d}\mathcal{L}(x)\right] \quad\le\quad 2\varepsilon + \frac{\|\mathsf{E}_{\mathcal{D}} - \varphi\circ\mathsf{O}_1\|_{2,\mathcal{G}}^2}{2\mu} \quad\le\quad 2\varepsilon + \frac{\sigma_{2,\mathcal{G}}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_1)^2 + \varepsilon}{2\mu}\,,$$

for a function $\varphi \in \mathbb{M}(\mathcal{O}, \mathbb{R}^d)$ such that $\|\mathsf{E}_{\mathcal{D}} - \varphi \circ \mathsf{O}_1\|_{2,\mathcal{G}}^2 \leq \sigma_{2,\mathcal{G}}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_1)^2 + \varepsilon$. Then, we apply Lemma 5 starting at $x_0' = x_{T_z}$ with a number of steps $T = \lceil a\kappa \log n \rceil$ and a fixed mini-batch size of $n_t = N = \left\lfloor \frac{n-1}{1+a\kappa\log n} \right\rfloor$. This ensure that $\sum_{t<T} n_t \leq n-1$ and gives

$$\begin{aligned} \mathbb{E}\left[\mathcal{L}(x_{T+T_z}) - \inf_{x\in\mathbb{R}^d} \mathcal{L}(x)\right] &\leq \widetilde{\Delta}\left(1 - \tfrac{1}{\kappa}\right)^T + \tfrac{\sigma_{\mathcal{G},2}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_N)^2}{2\mu} + \tfrac{\varepsilon}{2\mu} \\ &\leq \widetilde{\Delta} e^{-T/\kappa} + \tfrac{\sigma_{\mathcal{G},2}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_N)^2}{2\mu} + \tfrac{\varepsilon}{2\mu} \\ &\leq \widetilde{\Delta} n^{-a} + \tfrac{\sigma_{\mathcal{G},2}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_N)^2}{2\mu} + \tfrac{\varepsilon}{2\mu}, \end{aligned} \tag{35}$$

where $\widetilde{\Delta} = \mathbb{E}\left[\mathcal{L}(x_{T_z}) - \inf_{x\in\mathbb{R}^d} \mathcal{L}(x)\right] \leq 2\varepsilon + \frac{\sigma_{2,\mathcal{G}}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_1)^2 + \varepsilon}{2\mu}$. Finally, letting ε tend to 0 concludes the proof. □

Proof of Proposition 5. Similarly to the proof of Proposition 4, we start with a warm-up phase of $T_z = \kappa \ln\left(\frac{\|\varphi(o_0)\|^2 + \|\mathsf{E}_{\mathcal{D}} - \varphi\circ\mathsf{O}_1\|_{2,\mathcal{G}}^2}{\varepsilon\mu}\right)$ steps using only the first observation $\mathsf{O}(\nabla\ell_{x_t}, z^{(1)})$. This gives, for a function $\varphi \in \mathbb{M}(\mathcal{O}, \mathbb{R}^d)$ such that $\|\mathsf{E}_{\mathcal{D}} - \varphi \circ \mathsf{O}_1\|_{2,\mathcal{G}}^2 \leq \sigma_{2,\mathcal{G}}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_1)^2 + \varepsilon$, an approximation error

$$\mathbb{E}\left[\mathcal{L}(x_{T_z}) - \inf_{x\in\mathbb{R}^d} \mathcal{L}(x)\right] \qu\leq \quad 2\varepsilon + \frac{\|\mathsf{E}_{\mathcal{D}} - \varphi\circ\mathsf{O}_1\|_{2,\mathcal{G}}^2}{2\mu} \quad \leq \quad 2\varepsilon + \frac{\sigma_{2,\mathcal{G}}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_1)^2 + \varepsilon}{2\mu}.$$

We then apply Lemma 5 starting at $x_0' = x_{T_z}$ with a number of steps $T = \lfloor (n-1)/2 \rfloor$ and an increasing mini-batch of size $n_t = \left\lceil (n-1)(1-c)c^{T-t-1}/2 \right\rceil$ where $c = \sqrt{1-\kappa^{-1}}$. This ensures that $\sum_{t<T} n_t \leq \sum_{t<T}(1 + (n-1)(1-c)c^{T-t-1}/2) \leq T + \frac{(n-1)(1-c)}{2(1-c)} \leq \frac{n-1}{2} + \frac{n-1}{2} = n-1$. Thus, applying Lemma 5 gives

$$\begin{aligned} \mathbb{E}\left[\mathcal{L}(x_T) - \inf_{x\in\mathbb{R}^d} \mathcal{L}(x)\right] &\leq \widetilde{\Delta} c^{2T} + \tfrac{1}{2L}\textstyle\sum_{t=0}^{T-1} c^{2(T-t-1)}\left(a + \frac{b}{n_t}\right) + \frac{\varepsilon}{2\mu} \\ &\leq \widetilde{\Delta} e^{-T/\kappa} + \tfrac{1}{2L}\textstyle\sum_{t=0}^{T-1} \frac{2bc^{T-t-1}}{(n-1)(1-c)} + \frac{a+\varepsilon}{2\mu} \\ &\leq \widetilde{\Delta} e^{-\frac{n-2}{2\kappa}} + \tfrac{b}{(n-1)L(1-c)^2} + \tfrac{a+\varepsilon}{2\mu}, \end{aligned} \tag{36}$$

where $\widetilde{\Delta} = \mathbb{E}\left[\mathcal{L}(x_{T_z}) - \inf_{x\in\mathbb{R}^d} \mathcal{L}(x)\right] \leq 2\varepsilon + \frac{\sigma_{2,\mathcal{G}}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_1)^2 + \varepsilon}{2\mu}$. We conclude by noting that $L(1-c)^2 = \mu\kappa(1 - \sqrt{1-\kappa^{-1}})^2 \geq \mu\kappa(2\kappa)^{-2} = \mu/4\kappa$ and, if $n \geq 3$, we have $n - 2 \geq n/3$ and $n - 1 \geq 2n/3$. □

F Proofs of Section 4

F.1 Empirical risk minimization

Proof of Proposition 6. Let $\varphi(g_1, \ldots, g_n) = \frac{1}{n}\sum_i g_i$ be the average over the n data points, then we directly have $\|\mathsf{E}_{\mathcal{D}} - \varphi \circ \mathsf{O}_n^{\mathrm{SL}}\|_{\mathcal{G},2}^2 = \sup_{g\in\mathcal{G}} \mathbb{E}\left[\|\frac{1}{n}\sum_i (g(\xi_i) - \mathbb{E}_{\xi\sim\mathcal{D}}[g(\xi)])\|^2\right] = \sup_{g\in\mathcal{G}} \operatorname{var}(g(\xi_1))/n$ as the data samples ξ_i are i.i.d. random variables. □

We next prove the upper bounds on $\|\mathsf{V}_{\mathcal{D}}\|_{\mathcal{G}} = \sup_{g\in\mathcal{G}} \operatorname{var}(g(\xi_1))$ provided after Proposition 6 in the main text.

Proof of the upper bounds on $\|\mathsf{V}_{\mathcal{D}}\|_{\mathcal{G}}$. We begin with affine functions. Let $(g : \xi \mapsto A\xi + b) \in \mathcal{G}_{\mathrm{Aff}}$. We have $\operatorname{var}(g(\xi)) = \operatorname{var}(A\xi) = \mathbb{E}\left[\|A(\xi - \mathbb{E}\xi)\|^2\right] \leq B^2\mathbb{E}\left[\|\xi - \mathbb{E}\xi\|\right]^2 = B^2\operatorname{var}(\xi)$, since $\rho(A) \leq B$. By taking a supremum over $g \in \mathcal{G}_{\mathrm{Aff}}$, we have the desired result. The equality is obtained by taking any rank D projection for A, in the case $D \leq d$.

For Lipschitz functions: let $g \in \mathcal{G}_{\mathrm{Lip}}$. We have $\operatorname{var}(g(\xi)) = \mathbb{E}\left[\|g(\xi) - \mathbb{E}g(\xi)\|^2\right] = \frac{1}{2}\mathbb{E}\left[\|g(\xi) - g(\xi')\|^2\right]$, where $\xi' \sim \mathcal{D}$ is independent from ξ. Thus, using the Lipschitzness of g, $\operatorname{var}(g(\xi)) \leq B^2\frac{1}{2}\mathbb{E}\left[\|\xi - \xi'\|^2\right] = B^2\operatorname{var}(\xi)$, and we take the supremum over g.

For functions with bounded variations: let $g \in \mathcal{G}_{\mathrm{Bnd}}$. There exists $c \in \mathbb{R}^d$ such that for all ξ, $\|g(\xi) - c\| \leq B$. Using $\operatorname{var}(g(\xi)) = \mathbb{E}\left[\|g(\xi) - \mathbb{E}[g(\xi)]\|^2\right] \leq \mathbb{E}\left[\|g(\xi) - c\|^2\right] \leq B^2$, we have $\operatorname{var}(g(\xi)) \leq B^2$, and we take the supremum over g. For the equality, we take g such that $g(\xi) = (B, 0, \ldots, 0)$ for $\xi \in A$ and $g(\xi) = (-B, 0, \ldots, 0)$ for $\xi \notin A$. □

We finally specify our SL results for $\mathcal{G}_{\mathrm{Bnd}}$ with the example or regularized SL.

Lemma 6. *Assume that* $\forall p \in [0,1]$, $\exists A \subset \Xi$ *measurable s.t.* $\mathbb{P}_{\mathcal{D}}(A) = p$. *Then, for* $n \geq 1$, *we have*

$$\frac{B}{1+\sqrt{n}} \leq \sigma_{\mathcal{G}_{\mathrm{Bnd}},2}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n^{\mathrm{SL}}) \leq \frac{B}{\sqrt{n}}\,. \tag{37}$$

Moreover, the average $\varphi(x) = \frac{1}{n}\sum_i x_i$ *is asymptotically optimal, as* $\|\mathsf{E}_{\mathcal{D}} - \varphi \circ \mathsf{O}_n^{\mathrm{SL}}\|^2_{\mathcal{G}_{\mathrm{Bnd}},2} = B^2/n$.

Proof. First, the upper bound is a direct application of Proposition 6 for $\mathcal{G}_{\mathrm{Bnd}}$, and is achieved for the average $\varphi(x) = \frac{1}{n}\sum_i x_i$. To prove the lower bound, we replace the supremum over all functions in $\mathcal{G}_{\mathrm{Bnd}}$ by functions of the form $g : \Xi \to \{-Bv, Bv\}$ where $\|v\| = 1$.

$$\begin{aligned}
\sigma_{\mathcal{G}_{\mathrm{Bnd}},2}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n^{\mathrm{SL}})^2 &\geq \inf_{\varphi \in \mathbb{M}(\mathcal{O},\mathbb{R}^d)} \sup_{g \in \mathbb{F}(\Xi,\{-Bv,Bv\})} \mathbb{E}\left[\|\mathsf{E}_{\mathcal{D}}(g) - \varphi(g(\xi_1),\dots,g(\xi_n))\|^2\right] \\
&\geq B^2 \inf_{\varphi \in \mathbb{M}(\mathcal{O},\mathbb{R}^d)} \sup_{p \in [0,1]} \mathbb{E}\left[\left(2p - 1 - \varphi(G_1,\dots,G_n)\right)^2\right],
\end{aligned}$$

where $G_i = \mathbb{1}\{g(\xi_i) = Bv\}$ are i.i.d. Bernoulli random variables of parameter $p = \mathbb{P}(g(\xi_1) = Bv)$. We now replace the probability p by a random variable $P \sim \mathrm{Beta}(a,b)$ for $a, b > 0$, which gives

$$\begin{aligned}
\sigma_{\mathcal{G}_{\mathrm{Bnd}},2}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n^{\mathrm{SL}})^2 &\geq B^2 \inf_{\varphi \in \mathbb{M}(\mathcal{O},\mathbb{R}^d)} \mathbb{E}\left[\left(2P - 1 - \varphi(G_1,\dots,G_n)\right)^2\right] \\
&= 4B^2 \inf_{\varphi \in \mathbb{M}(\mathcal{O},\mathbb{R}^d)} \mathbb{E}\left[\left(P - \varphi(G_1,\dots,G_n)\right)^2\right] \\
&= 4B^2\, \mathbb{E}\left[\left(P - \mathbb{E}\left[P \mid G_1,\dots,G_n\right]\right)^2\right] \\
&= 4B^2\, \mathbb{E}\left[\left(P - \mathbb{E}\left[P \mid K\right]\right)^2\right],
\end{aligned}$$

where $K = \sum_i G_i$ is a Binomial distribution of parameters n and p, as the Bernoulli r.v. are identically distributed. A simple calculation gives that $P|K \sim \mathrm{Beta}(K+a, n-K+b)$, which allows us to compute the quantities $\mathbb{E}\left[P^2\right] = \frac{ab}{(a+b)^2(1+a+b)} + \frac{a^2}{(a+b)^2}$ and $\mathbb{E}\left[P \mid K\right] = \frac{K+a}{n+a+b}$. We thus obtain

$$\begin{aligned}
\sigma_{\mathcal{G}_{\mathrm{Bnd}},2}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n^{\mathrm{SL}})^2 &\geq 4B^2 \left(\mathbb{E}\left[P^2\right] - \mathbb{E}\left[\mathbb{E}\left[P \mid K\right]^2\right]\right) \\
&= 4B^2 \left(\frac{ab}{(a+b)^2(1+a+b)} + \frac{a^2}{(a+b)^2} - \mathbb{E}\left[\left(\frac{K+a}{n+a+b}\right)^2\right]\right) \\
&= \frac{4B^2 ab}{(a+b)(1+a+b)(n+a+b)},
\end{aligned}$$

where the last equality is obtained using $\mathbb{E}\left[(K+a)^2\right] = \mathrm{var}(K) + (\mathbb{E}\left[K\right] + a)^2 = \frac{nab(n+a+b)}{(a+b)^2(1+a+b)} + \left(\frac{a(n+a+b)}{a+b}\right)^2$. Finally, choosing $a = b = \sqrt{n}/2$ gives the desired result. □

Proof of Proposition 7. Noticing that $\sigma_{\mathcal{G},2}\left(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n\right)^2 \leq B^2/n$ and $\sigma_{2,\mathcal{G}}\left(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_1\right)^2 \leq 4B^2$, we can use Proposition 5 with $a = 0$ in order to obtain

$$\varepsilon_{\mathrm{sc}}(\mathcal{G}_{\mathrm{Bnd}}, \mathsf{O}_n^{\mathrm{SL}}) \quad \leq \quad \frac{6\kappa B^2}{\mu n} + \frac{2B^2}{\mu} e^{-\frac{n}{6\kappa}}\,.$$

The right handside of Proposition 7 is obtained by using $e^{-\frac{n}{6\kappa}} \leq \frac{6\kappa}{ne}$ and $6 + 12/e = 10.41... \leq 11$. The left handside of the desired inequality is then proved using the lower bound of Lemma 6 together with the general lower bound of Lemma 2. □

F.2 Transfer learning

Proof of Proposition 8. Let $\varphi(g_1, \dots, g_n) = \frac{1}{n}\sum_i g_i$ be the average over the n data points, then we directly have, for $\xi \sim \mathcal{D}$,

$$\begin{aligned} \|\mathsf{E}_{\mathcal{D}} - \varphi \circ \mathsf{O}_n^{\mathrm{TL}}\|_{\mathcal{G},2}^2 &= \sup_{g\in\mathcal{G}} \mathbb{E}\left[\|\mathbb{E}\left[g(\xi)\right] - \tfrac{1}{n}\textstyle\sum_i g(\xi_i')\|^2\right] \\ &= \sup_{g\in\mathcal{G}} \|\mathbb{E}\left[g(\xi)\right] - \mathbb{E}\left[g(\xi_1')\right]\|^2 + \tfrac{\mathrm{var}(g(\xi_1'))}{n} \\ &\le \sup_{g\in\mathcal{G}} \|\mathbb{E}\left[g(\xi)\right] - \mathbb{E}\left[g(\xi_1')\right]\|^2 + \sup_{g\in\mathcal{G}} \tfrac{\mathrm{var}(g(\xi_1'))}{n} \\ &= d_{\mathcal{G}}(\mathcal{D},\mathcal{D}')^2 + \tfrac{\|\mathsf{V}_{\mathcal{D}'}\|_{\mathcal{G}}}{n}\,. \end{aligned} \tag{38}$$

□

Proof of Proposition 9. By assumption, $\forall \varepsilon \in (0,1], \exists q_\varepsilon \in \mathbb{R}$ s.t. $\mathbb{P}_{\mathcal{D}'}\left(\frac{d\mathcal{D}}{d\mathcal{D}'}(\xi') \ge q_\varepsilon\right) = (1+\varepsilon)/2$. We apply Proposition 2 to $\mathsf{O}_n^{\mathrm{TN}}$ with $g(\xi) = -g'(\xi) = B\left(2\mathbb{1}\left\{\frac{d\mathcal{D}}{d\mathcal{D}'}(\xi) \ge q_\varepsilon\right\} - 1\right)e_1$ where $e_1 = (1,0,\dots)^\top$ is the first basis vector. First, note that $g, g' \in \mathcal{G}_{\mathrm{Bnd}}$, which gives

$$\sigma_{\mathcal{G},2}\left(\mathsf{E}_{\mathcal{D}}|\mathsf{O}\right)^2 \ge \left(1 - d_{\mathrm{LC}}(\mathsf{O}_n^{\mathrm{TL}}(g,z), \mathsf{O}_n^{\mathrm{TL}}(-g,z))\right)\|\mathsf{E}_{\mathcal{D}}(g)\|^2\,, \tag{39}$$

where $d_{\mathrm{LC}}(p,q)$ is Le Cam's distance (see Appendix B). By definition,

$$\begin{aligned} \|\mathsf{E}_{\mathcal{D}}(g)\|^2 &= B^2\mathbb{E}_{\mathcal{D}}\left[2\mathbb{1}\left\{\tfrac{d\mathcal{D}}{d\mathcal{D}'}(\xi') \ge q_\varepsilon\right\} - 1\right]^2 \\ &= B^2\mathbb{E}_{\mathcal{D}'}\left[\left(2\mathbb{1}\left\{\tfrac{d\mathcal{D}}{d\mathcal{D}'}(\xi') \ge q_\varepsilon\right\} - 1\right)\tfrac{d\mathcal{D}}{d\mathcal{D}'}(\xi')\right]^2 \\ &= B^2\left(\mathbb{E}_{\mathcal{D}'}\left[\left(2\mathbb{1}\left\{\tfrac{d\mathcal{D}}{d\mathcal{D}'}(\xi') \ge q_\varepsilon\right\} - 1\right)\left(\tfrac{d\mathcal{D}}{d\mathcal{D}'}(\xi') - q_\varepsilon\right)\right] + q_\varepsilon\varepsilon\right)^2 \\ &= B^2\left(\mathbb{E}_{\mathcal{D}'}\left[\left|\tfrac{d\mathcal{D}}{d\mathcal{D}'}(\xi') - q_\varepsilon\right|\right] + q_\varepsilon\varepsilon\right)^2\,. \end{aligned} \tag{40}$$

Moreover, we have $\mathbb{E}_{\mathcal{D}'}\left[\left|\frac{d\mathcal{D}}{d\mathcal{D}'}(\xi') - q_\varepsilon\right|\right] \ge 2d_{\mathrm{TV}}(\mathcal{D},\mathcal{D}') - |q_\varepsilon - 1|$ and $\mathbb{E}_{\mathcal{D}'}\left[\left|\frac{d\mathcal{D}}{d\mathcal{D}'}(\xi') - q_\varepsilon\right|\right] \ge |q_\varepsilon - 1|$. As a consequence, we have $\mathbb{E}_{\mathcal{D}'}\left[\left|\frac{d\mathcal{D}}{d\mathcal{D}'}(\xi') - q_\varepsilon\right|\right] \ge d_{\mathrm{TV}}(\mathcal{D},\mathcal{D}')$ and $\mathbb{E}_{\mathcal{D}'}\left[\left|\frac{d\mathcal{D}}{d\mathcal{D}'}(\xi') - q_\varepsilon\right|\right] + q_\varepsilon\varepsilon \ge |q_\varepsilon - 1| + q_\varepsilon\varepsilon \ge \varepsilon$, which gives

$$\|\mathsf{E}_{\mathcal{D}}(g)\|^2 \ge B^2\max\left\{d_{\mathrm{TV}}(\mathcal{D},\mathcal{D}'), \varepsilon\right\}^2 \ge \frac{B^2}{2}\left(d_{\mathrm{TV}}(\mathcal{D},\mathcal{D}')^2 + \varepsilon^2\right)\,. \tag{41}$$

Finally, we conclude by noting that, as g only takes two values ($-Be_1$ and Be_1), we have that, if $N = |\{i \in [\![1,n]\!] \mid g(\xi_i') = Be_1\}|$ and $\varepsilon \le 1/2$, then

$$\begin{aligned} d_{\mathrm{LC}}(\mathsf{O}_n^{\mathrm{TL}}(g,z), \mathsf{O}_n^{\mathrm{TL}}(-g,z)) &\le d_{\mathrm{KL}}(\mathsf{O}_n^{\mathrm{TL}}(g,z), \mathsf{O}_n^{\mathrm{TL}}(-g,z)) \\ &= (2\mathbb{E}\left[N\right] - n)\ln\left(\tfrac{1+\varepsilon}{1-\varepsilon}\right) \\ &= n\varepsilon\ln\left(\tfrac{1+\varepsilon}{1-\varepsilon}\right) \\ &\le 2\ln(3)n\varepsilon^2\,, \end{aligned} \tag{42}$$

and taking $\varepsilon = 1/(2\sqrt{n})$ gives the desired result

$$\sigma_{\mathcal{G},2}\left(\mathsf{E}_{\mathcal{D}}|\mathsf{O}\right)^2 \ge \frac{(2-\ln(3))\,B^2}{16}\left(d_{\mathrm{TV}}(\mathcal{D},\mathcal{D}')^2 + \frac{1}{n}\right)\,. \tag{43}$$

□

F.3 (Personalized) Federated Learning

We now prove the following upper bound:

$$\sigma_{\mathcal{G},2}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n^{\mathrm{FL}})^2 \le \inf_{q\in\mathbb{R}^m} d_{\mathcal{G}}\left(\mathcal{D}, \sum_i q_i\mathcal{D}_i\right)^2 + \sum_{i=1}^m q_i^2\frac{\|\mathsf{V}_{\mathcal{D}_i}\|_{\mathcal{G}}}{n_i}\,.$$

Proof of Proposition 10. We start by upper bounding using the function $\varphi((g_j^i)_{i\in[\![1,m]\!],j\in[\![1,n_i]\!]}) = \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{q_i}{n_i} g_j^i$, leading to

$$\begin{aligned}
\sigma_{\mathcal{G},2}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n^{\mathrm{FL}})^2 &\le \sup_{g\in\mathcal{G}} \mathbb{E}\left[\left\|\mathsf{E}_{\mathcal{D}}(g) - \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{q_i}{n_i} g(\xi_j^i)\right\|^2\right] \\
&= \sup_{g\in\mathcal{G}} \mathbb{E}\left[\left\|\mathsf{E}_{\mathcal{D}}(g) - \mathbb{E}\left[\sum_{i=1}^m \sum_{j=1}^{n_i} \frac{q_i}{n_i} g(\xi_j^i)\right]\right\|^2\right] + \operatorname{var}\left(\sum_{i=1}^m \sum_{j=1}^{n_i} \frac{q_i}{n_i} g(\xi_j^i)\right) \\
&= \sup_{g\in\mathcal{G}} \mathbb{E}\left[\left\|\mathsf{E}_{\mathcal{D}}(g) - \sum_{i=1}^m q_i \mathsf{E}_{\mathcal{D}_i}(g)\right\|^2\right] + \sum_{i=1}^m \frac{q_i^2}{n_i} \operatorname{var}\left(g(\xi_1^i)\right) \\
&\le d_{\mathcal{G}}\left(\mathcal{D}, \sum_i q_i \mathcal{D}_i\right)^2 + \sum_i \frac{q_i^2}{n_i} \|\mathsf{V}_{\mathcal{D}_i}\|_{\mathcal{G}} \,.
\end{aligned}$$

We conclude by taking the infimum over (q_i). □

F.4 Robust Learning

We now prove the two results of the robust learning section. First, the upper and lower bound for $\mathcal{G}_{\mathrm{Bnd}}$,

$$\sigma_{\mathcal{G}_{\mathrm{Bnd}},2}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n^{\mathrm{RL}})^2 = \Theta\left(B^2\left(\eta^2 + \frac{1}{n}\right)\right), \tag{44}$$

is obtained by applying Proposition 9 to $\mathcal{D}' = (1-\eta)\mathcal{D} + \eta\mathcal{D}_o$, and noting that,

$$d_{\mathrm{TV}}(\mathcal{D}, \mathcal{D}') = \frac{1}{2} \sup_{f\in\mathbb{F}(\Xi,[-1,1])} \mathsf{E}_{\mathcal{D}}(f) - \mathsf{E}_{\mathcal{D}'}(f) = \frac{\eta}{2} \sup_{f\in\mathbb{F}(\Xi,[-1,1])} \mathsf{E}_{\mathcal{D}}(f) - \mathsf{E}_{\mathcal{D}_o}(f) = \eta\,, \tag{45}$$

as $\mathsf{E}_{\mathcal{D}'}(f) = (1-\eta)\mathsf{E}_{\mathcal{D}}(f) + \eta\mathsf{E}_{\mathcal{D}_o}(f)$ and, by assumption, $d_{\mathrm{TV}}(\mathcal{D}, \mathcal{D}_o) = 1$. The second result, for $\mathcal{G}_{\mathrm{Lip}}$,

$$\sigma_{\mathcal{G}_{\mathrm{Lip}},2}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n^{\mathrm{RL}})^2 \le cB^2 \operatorname{var}(\xi)\left(\eta + \frac{1}{n}\right), \tag{46}$$

is proved as follows: let φ be the robust mean estimator of Steinhardt et al. (2018, Algorithm 1). First, note that, for a set of data points $(\xi_1, \ldots, \xi_m)$, we have, with $\lambda_{\max}(M)$ denoting the largest singular value of the symmetric matrix M,

$$\begin{aligned}
\lambda_{\max}\left(\tfrac{1}{m}\textstyle\sum_{i=1}^m (g(\xi_i) - \bar{g}_m)(g(\xi_i) - \bar{g}_m)^\top\right) &= \max_{x\,:\,\|x\|\le 1} \tfrac{1}{m}\textstyle\sum_{i=1}^m (x^\top(g(\xi_i) - \bar{g}_m))^2 \\
&\le \tfrac{1}{m}\textstyle\sum_{i=1}^m \|g(\xi_i) - \bar{g}_m\|^2 \\
&\le \tfrac{B^2}{m}\textstyle\sum_{i=1}^m \|\xi_i - \bar{\xi}_m\|^2 \,,
\end{aligned} \tag{47}$$

where $\bar{g}_m$ and $\bar{\xi}_m$ are, respectively, the averages of $g(\xi_i)$ and ξ_i over all data points. Thus, assuming without loss of generality that we place all the outliers at the end of the sequence $(\xi_1, \ldots, \xi_n)$, we can use $\sigma_0^2 = \frac{B^2}{(1-\eta)n}\sum_{i=1}^{(1-\eta)n} \|\xi_i - \bar{\xi}_{(1-\eta)n}\|^2$ in Proposition 16 of Steinhardt et al. (2018), and get that, if $\eta \le 1/4$, the robust mean estimator always returns a value $\varphi \circ \mathsf{O}_n^{\mathrm{RL}}(g, z)$ such that

$$\|\bar{g}_{(1-\eta)n} - \varphi \circ \mathsf{O}_n^{\mathrm{RL}}(g, z)\|^2 \le \frac{c^2B^2}{(1-\eta)n} \sum_{i=1}^{(1-\eta)n} \|\xi_i - \bar{\xi}_{(1-\eta)n}\|^2 \eta \tag{48}$$

where $c = 40$ is a universal constant. Finally, we take the expectation over the samples (i.e. z) and have

$$\begin{aligned}
\|\mathsf{E}_{\mathcal{D}} - \varphi \circ \mathsf{O}_n^{\mathrm{RL}}\|_{\mathcal{G}_{\mathrm{Lip}}}^2 &\le 2\sup_{g\in\mathcal{G}_{\mathrm{Lip}}} \mathbb{E}\left[\|\mathsf{E}_{\mathcal{D}}(g) - \bar{g}_{(1-\eta)n}\|^2 + \|\bar{g}_{(1-\eta)n} - \varphi \circ \mathsf{O}_n^{\mathrm{RL}}(g, z)\|^2\right] \\
&\le \tfrac{2B^2 \operatorname{var}(\xi)}{(1-\eta)n} + \tfrac{2c^2B^2}{(1-\eta)n} \textstyle\sum_{i=1}^{(1-\eta)n} \mathbb{E}\left[\|\xi_i - \bar{\xi}_{(1-\eta)n}\|^2\right] \eta \,.
\end{aligned} \tag{49}$$

We conclude by showing that $\mathbb{E}\left[\|\xi_i - \bar{\xi}_{(1-\eta)n}\|^2\right] = (1 - \frac{1}{n})\operatorname{var}(\xi) \le \operatorname{var}(\xi)$ and $\frac{1}{1-\eta} \le 4/3$, giving Eq. (46) with the constant $c = 3200$.

F.5 Learning with fixed data points

We now provide a proof of the upper and lower bounds for the minimax excess risk under the fixed data learning scenario.

Proof of Proposition 11. First, as $\mathsf{O}_n^{\mathrm{FD}}$ is deterministic, Corollary 1 immediately gives that

$$\varepsilon_{\mathrm{sc}}(\mathcal{G}_{\mathrm{Bnd}}, \mathsf{O}_n^{\mathrm{FD}}) = \frac{\sigma_{2,\mathcal{G}_{\mathrm{Bnd}}}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n^{\mathrm{FD}})^2}{2\mu} \quad \text{and} \quad \varepsilon_{\mathrm{sc}}(\mathcal{G}_{\mathrm{Lip}}, \mathsf{O}_n^{\mathrm{FD}}) = \frac{\sigma_{2,\mathcal{G}_{\mathrm{Lip}}}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n^{\mathrm{FD}})^2}{2\mu}\,. \tag{50}$$

We thus only need to show that $\sigma_{2,\mathcal{G}_{\mathrm{Bnd}}}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n^{\mathrm{FD}}) = 2B\left(1 - \mathbb{P}_{\mathcal{D}}\left(\{\xi_i'\}_{i[\![1,n]\!]}\right)\right)$ and $\sigma_{2,\mathcal{G}_{\mathrm{Lip}}}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n^{\mathrm{FD}}) = B\mathbb{E}\left[\min_i \|\xi - \xi_i'\|\right]$ to conclude.

Case $\mathcal{G}_{\mathrm{Bnd}}$: First, let us assume that all ξ_i are distinct, as we can otherwise remove the duplicates without any loss of generality. Then, consider the estimator $\varphi(g_1, \dots, g_n) = \sum_{i=1}^n \mathbb{P}_{\mathcal{D}}\left(\{\xi_i'\}\right) g_i + \left(1 - \mathbb{P}_{\mathcal{D}}\left(\{\xi_i'\}_{i[\![1,n]\!]}\right)\right)\frac{1}{n}\sum_i g_i$, which can be used here as the ξ_i' are fixed data points and $\mathcal{D}$ is known. Thus, using φ, we have

$$\begin{aligned}
\sigma_{2,\mathcal{G}_{\mathrm{Bnd}}}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n^{\mathrm{FD}}) &\leq \sup_{g\in\mathcal{G}_{\mathrm{Bnd}}} \|\mathsf{E}_{\mathcal{D}}(g) - \varphi(g(\xi_1'), \dots, g(\xi_n'))\| \\
&= \sup_{g\in\mathcal{G}_{\mathrm{Bnd}}} \|\mathbb{E}\left[\mathbb{1}\{\xi \notin \{\xi_i'\}_{i\in[\![1,n]\!]}\}(g(\xi) - \tfrac{1}{n}\textstyle\sum_i g(\xi_i'))\right]\| \\
&\leq \sup_{g\in\mathcal{G}_{\mathrm{Bnd}}} \mathbb{E}\left[\mathbb{1}\{\xi \notin \{\xi_i'\}_{i\in[\![1,n]\!]}\}\|g(\xi) - \tfrac{1}{n}\textstyle\sum_i g(\xi_i')\|\right] \\
&\leq 2B\left(1 - \mathbb{P}_{\mathcal{D}}\left(\{\xi_i'\}_{i[\![1,n]\!]}\right)\right)\,.
\end{aligned} \tag{51}$$

The lower bound is obtained using Proposition 2 with $g(\xi) = -g'(\xi) = \mathbb{1}\{\xi \notin \{\xi_i'\}_{i\in[\![1,n]\!]}\}2Be_1$, where e_1 is the first basis vector. This gives $\mathsf{O}_n^{\mathrm{FD}}(g, z) = \mathsf{O}_n^{\mathrm{FD}}(g', z) = (0, \dots, 0)$, and thus $d_{\mathrm{TV}}(\mathsf{O}_n^{\mathrm{FD}}(g, z), \mathsf{O}_n^{\mathrm{FD}}(g', z)) = 0$ and

$$\sigma_{2,\mathcal{G}_{\mathrm{Bnd}}}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n^{\mathrm{FD}}) \geq \|\mathsf{E}_{\mathcal{D}}(g)\| = 2B\left(1 - \mathbb{P}_{\mathcal{D}}\left(\{\xi_i'\}_{i[\![1,n]\!]}\right)\right)\,, \tag{52}$$

which concludes the proof.

Case $\mathcal{G}_{\mathrm{Lip}}$: Again, we assume, without loss of generality, that all ξ_i are distinct. For the upper bound, we use the estimator $\varphi(g_1, \dots, g_n) = \sum_{i=1}^n \mathbb{P}_{\mathcal{D}}\left(\mathrm{argmin}_j \|\xi - \xi_j'\| = i\right) g_i$, which gives

$$\begin{aligned}
\sigma_{2,\mathcal{G}_{\mathrm{Lip}}}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n^{\mathrm{FD}}) &\leq \sup_{g\in\mathcal{G}_{\mathrm{Lip}}} \|\mathsf{E}_{\mathcal{D}}(g) - \varphi(g(\xi_1'), \dots, g(\xi_n'))\| \\
&= \sup_{g\in\mathcal{G}_{\mathrm{Lip}}} \|\mathbb{E}\left[\textstyle\sum_{i=1}^n \mathbb{1}\{\mathrm{argmin}_j \|\xi - \xi_j'\| = i\}(g(\xi) - g(\xi_i'))\right]\| \\
&\leq \sup_{g\in\mathcal{G}_{\mathrm{Lip}}} \mathbb{E}\left[\textstyle\sum_{i=1}^n \mathbb{1}\{\mathrm{argmin}_j \|\xi - \xi_j'\| = i\}\|g(\xi) - g(\xi_i')\|\right] \\
&\leq B\,\mathbb{E}\left[\textstyle\sum_{i=1}^n \mathbb{1}\{\mathrm{argmin}_j \|\xi - \xi_j'\| = i\}\|\xi - \xi_i'\|\right] \\
&= B\,\mathbb{E}\left[\min_i \|\xi - \xi_i'\|\right]\,.
\end{aligned} \tag{53}$$

For the lower bound, we apply Proposition 2 with $g(\xi) = -g(\xi) = B\min_i \|\xi - \xi_i'\|e_1$, which is B-Lipschitz by construction. As $\mathsf{O}_n^{\mathrm{FD}}(g, z) = \mathsf{O}_n^{\mathrm{FD}}(g', z) = (0, \dots, 0)$, we have $d_{\mathrm{TV}}(\mathsf{O}_n^{\mathrm{FD}}(g, z), \mathsf{O}_n^{\mathrm{FD}}(g', z)) = 0$ and

$$\sigma_{2,\mathcal{G}_{\mathrm{Lip}}}(\mathsf{E}_{\mathcal{D}}|\mathsf{O}_n^{\mathrm{FD}}) \geq \|\mathsf{E}_{\mathcal{D}}(g)\| = B\,\mathbb{E}\left[\min_i \|\xi - \xi_i'\|\right]\,, \tag{54}$$

which concludes the proof. □