Beyond Least-Squares: Fast Rates for Regularized Empirical Risk Minimization through Self-Concordance

Ulysse Marteau-Ferey, Dmitrii Ostrovskii, Francis Bach, Alessandro Rudi

Beyond Least-Squares: Fast Rates for Regularized Empirical Risk Minimization through Self-Concordance

Ulysse Marteau-Ferey Dmitrii Ostrovskii Francis Bach Alessandro Rudi

INRIA - Département d'Informatique de l'École Normale Supérieure
PSL Research University
Paris, France

Abstract

We consider learning methods based on the regularization of a convex empirical risk by a squared Hilbertian norm, a setting that includes linear predictors and non-linear predictors through positive-definite kernels. In order to go beyond the generic analysis leading to convergence rates of the excess risk as $O(1/\sqrt{n})$ from n observations, we assume that the individual losses are self-concordant, that is, their third-order derivatives are bounded by their second-order derivatives. This setting includes least-squares, as well as all generalized linear models such as logistic and softmax regression. For this class of losses, we provide a bias-variance decomposition and show that the assumptions commonly made in least-squares regression, such as the source and capacity conditions, can be adapted to obtain fast non-asymptotic rates of convergence by improving the bias terms, the variance terms or both.
Keywords: Self-concordance, regularization, logistic regression, non-parametric estimation.

1 Introduction

Regularized empirical risk minimization remains a cornerstone of statistics and supervised learning, from the early days of linear regression [17] and neural networks [13], then to spline smoothing [41] and more generally kernel-based methods [31]. While the regularization by the squared Euclidean norm is applied very widely, the statistical analysis of the resulting learning methods is still not complete.

The main goal of this paper is to provide a sharp non-asymptotic analysis of regularized empirical risk minimization (ERM), or more generally regularized M-estimation, that is estimators obtained as the unique solution of

$$\min_{\theta \in \mathcal{H}} \ \frac{1}{n} \sum_{i=1}^{n} \ell_{z_i}(\theta) + \frac{\lambda}{2} \|\theta\|^2, \tag{1}$$

where $\mathcal{H}$ is a Hilbert space (possibily infinite-dimensional) and $\ell_z(\theta)$ is the convex loss associated with an observation z and the estimator $\theta \in \mathcal{H}$. We assume that the observations z_i, $i = 1, \dots, n$ are independent and identically distributed, and that the minimum of the associated unregularized expected risk $L(\theta)$ is attained at a certain $\theta^\star \in \mathcal{H}$.

In this paper, we focus on dimension-independent results [thus ultimately extending the analysis in the finite-dimensional setting from 25]. For this class of problems, two main classes of problems have been studied, depending on the regularity assumptions on the loss.

Convex Lipschitz-continuous losses (with respect to the parameter θ), such as for logistic regression or the support vector machine, lead to general *non-asymptotic* bounds for the excess risk of the form [33]:

$$\frac{B^2}{\lambda n} + \lambda\|\theta^\star\|^2, \tag{2}$$

where B is a uniform upper bound on the Lipschitz constant for all losses $\theta \mapsto \ell_z(\theta)$. The bound above already has a form that takes into account two separate terms: a *variance term* $B^2/(\lambda n)$ which depends on the sample size n but not on the optimal predictor $\theta^\star$, and a *bias term* $\lambda\|\theta^\star\|^2$ which depends on the optimal predictor but not on the sample size n. All our bounds will have this form but with smaller quantities (but asking fore more assumptions). Without further assumptions, in Eq. (2), λ is taken proportional to $1/\sqrt{n}$, and we get the usual optimal slow rate in excess risk of $O(1/\sqrt{n})$ associated with such a general set-up [see, e.g., 9].

For the specific case of quadratic losses of the form $\ell_z(\theta) = \frac{1}{2}(y - \theta \cdot \Phi(x))^2$, where $z = (x, y)$, and $y \in \mathbb{R}$ and $\Phi(x) \in \mathcal{H}$, the situation is much richer. Without further assumptions, the same rate $O(1/\sqrt{n})$ is achieved, but stronger assumptions lead to faster rates [8]. In particular, the decay of the eigenvalues of the Hessian $\mathbb{E}\big[\Phi(x) \otimes \Phi(x)\big]$ (often called the *capacity condition*) leads to an improved variance term, while the finiteness of some bounds on $\theta^\star$ for norms other than the plain Hilbertian norms $\|\theta^\star\|$ (often called the *source condition*) leads to an improved bias term. Both of these assumptions lead to faster rates than $O(1/\sqrt{n})$ for the excess risk, with the proper choice of the regularization parameter λ. For least-squares, these rates are then optimal and provide a better understanding of properties of the problem that influence the generalization capabilities of regularized ERM [see, e.g. 32, 8, 35, 11, 5].

Our main goal in this paper is to bridge the gap between Lipschitz-continuous and quadratic losses by improving on slow rates for general classes of losses beyond least-squares. We first note that: (a) there has to be an extra regularity assumption because of lower bounds [9], and (b) asymptotically, we should obtain bounds that approach the local quadratic approximation of $\ell_z(\theta)$ around $\theta^\star$ with the same optimal behavior as for plain least-squares.

Several frameworks are available for such an extension with extra assumptions on the losses, such as "exp-concavity" [19, 23], strong convexity [38] or a generalized notion of self-concordance [2, 25]. In this paper, we focus on self-concordance, which links the second and third order derivatives of the loss. This notion is quite general and corresponds to widely used losses in machine learning, and does not suffer from constants which can be exponential in problem parameters (e.g., $\|\theta^\star\|$) when applied to generalized linear models like logistic regression. See Sec. 1.1 for a comparison to related work.

With this self-concordance assumption, we will show that our problem behaves like a quadratic problem corresponding to the local approximation around $\theta^\star$, in a totally non-asymptotic way, which is the core technical contribution of this paper. As we have already mentioned, this phenomenon is naturally expected in the asymptotic regime, but is hard to capture in the non-asymptotic setting without constants which explode exponentially with the problem parameters.

The paper is organized as follows: in Sec. 2, we present our main assumptions and informal results, as well as our bias-variance decomposition. In order to introduce precise results gradually, we start in Sec. 3 with a result similar to Eq. (2) for our set-up to show that we recover with a simple argument the result from Sridharan et al. [33], which itself applies more generally. Then, in Sec. 4 we introduce the source condition allowing for a better control of the bias. Finally, in Sec. 5, we detail the capacity condition leading to an improved variance term, which, together with the improved bias leads to fast rates (which are optimal for least-squares).

1.1 Related work

Fast rates for empirical risk minimization. Rates faster than $O(1/\sqrt{n})$ can be obtained with a variety of added assumptions, such as some form of strong convexity [33, 7], noise conditions for classification [34], or extra conditions on the loss, such as self-concordance [2] or exp-concavity [19, 23], whose partial goal is to avoid exponential constants. Note that Bach [2] already considers logistic regression with Hilbert spaces, but only for well-specified models and a fixed design, and without the sharp and simpler results that we obtain in this paper.

Avoiding exponential constants for logistic regression. The problem of exponential constants (i.e., leading factors in the rates scaling as e^{RD} where D is the radius of the optimal predictor, and R the radius of the design) is long known. In fact, Hazan et al. [16] showed a lower bound, explicitly constructing an adversarial distribution (i.e., an ill-specified model) for which the problem manifests in the finite-sample regime with $n = O(e^{RD})$. Various attempts to address this problem are found in the literature. For example, Ostrovskii and Bach [25, App. C] prove the optimal d/n rate in the non-regularized d-dimensional setting but, multiplied with the curvature parameter ρ which is at worst exponential but is shown to grow at most as $(RD)^{3/2}$ in the case of Gaussian design. Another approach is due to Foster et al. [12]: they establish "1-mixability" of the logistic loss, then apply Vovk's aggregating algorithm in the online setting, and then proceed via online-to-batch conversion. While this result allows to obtain the fast $O(d/n)$ rate (and its counterparts in the nonparametric setting) without exponential constants, the resulting algorithm is *improper* (i.e., the canonical parameter $\eta = \Phi(x) \cdot \theta^\star$, see below, is estimated by a *non-linear* functional of $\Phi(x)$).

A closely related approach is to use the notion of exp-concavity instead of mixability [27, 19, 23]. The two close notions are summarized in the so-called central condition (due to Van Erven et al. [40]) which fully characterizes when the fast $O(d/n)$ rates (up to log factors and in high probability) are available for improper algorithms. However, when proper learning algorithms are concerned, this analysis requires η-mixability (or η-exp-concavity) of the *overall* loss $\ell_z(\theta)$ for which the η parameter scales with the radius of the set of predictors. This scaling is exponential for the logistic loss, leading to exponential constants.

2 Main Assumptions and Results

Let $\mathcal{Z}$ be a Polish space and Z be a random variable on $\mathcal{Z}$ with distribution ρ. Let $\mathcal{H}$ be a separable (non-necessarily finite-dimensional) Hilbert space, with norm $\|\cdot\|$, and let $\ell : \mathcal{Z} \times \mathcal{H} \to \mathbb{R}$ be a loss function, we denote by $\ell_z(\cdot)$ the function $\ell(z, \cdot)$. Our goal is to minimize the expected risk with respect to $\theta \in \mathcal{H}$:

$$\inf_{\theta \in \mathcal{H}} L(\theta) = \mathbb{E}\left[\ell_Z(\theta)\right].$$

Given $(z_i)_{i=1}^n \in \mathcal{Z}^n$, we will consider the following estimator based on regularized empirical risk minimization given $\lambda > 0$ (note that the minimizer is unique in this case):

$$\widehat{\theta}_\lambda^\star = \arg\min_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \ell_{z_i}(\theta) + \frac{\lambda}{2}\|\theta\|^2,$$

where we assume the following.

Assumption 1 (i.i.d. data)**.** *The samples $(z_i)_{1 \leqslant i \leqslant n}$ are independently and identically distributed according to ρ.*

The goal of this work is to provide upper bounds in high probability for the so-called *excess risk*

$$L(\widehat{\theta}_\lambda^\star) - \inf_{\theta \in \mathcal{H}} L(\theta),$$

and thus to provide a general framework to measure the quality of the estimator $\widehat{\theta}_\lambda^\star$. Algorithms for obtaining such estimators have been extensively studied, in both finite-dimensional regimes, where a direct optimization over θ is performed, typically by gradient descent or stochastic versions thereof [see, e.g., 6, 30] and infinite-dimensional regimes, where kernel-based methods are traditionally used [see, e.g., 18, 14, 10, 37, 29, and references therein].

Example 1 (Supervised learning). *Although formulated as a general M-estimation problem [see, e.g., 21], our main motivation comes from supervised learning, with $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ is the data space and $\mathcal{Y}$ the target space. We will consider, as examples, losses with both real-valued outputs but also the multivariate case. For learning real-valued outputs, consider we have a bounded representation of the input space $\Phi : \mathcal{X} \to \mathcal{H}$ [potentially implicit when using kernel-based methods, 1]. We will provide bounds for the following losses.*

- *The square loss $\ell_z(\theta) = \frac{1}{2}\left(y - \theta \cdot \Phi(x)\right)^2$, which is not Lipschitz-continuous.*
- *The Huber losses $\ell_z(\theta) = \psi(y - \theta \cdot \Phi(x))$ where $\psi(t) = \sqrt{1+t^2} - 1$ or $\psi(t) = \log \frac{e^t + e^{-t}}{2}$ [15], which are Lipschitz-continuous.*
- *The logistic loss $\ell_z(\theta) = \log(1 + e^{-y\theta\cdot\Phi(x)})$ commonly used in binary classification where $y \in \{-1, 1\}$, which is Lipschitz-continuous.*

Our framework goes beyond real-valued outputs, and can be applied to all generalized linear models (GLM) [22], including softmax regression: we consider a representation function $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathcal{H}$ and an a priori measure μ on $\mathcal{Y}$. The loss we consider in this case is

$$\ell_z(\theta) = -\theta \cdot \Phi(x, y) + \log \int_{\mathcal{Y}} \exp\left(\theta \cdot \Phi(x, y')\right) d\mu(y'),$$

which corresponds to the negative conditional log-likelihood when modelling y given x by the distribution $p(y|x, \theta) \sim \frac{\exp(\theta\cdot\Phi(x,y))}{\int_{\mathcal{Y}} \exp(\theta\cdot\Phi(x,y'))d\mu(y')} d\mu(y)$. Our framework applies to all of these generalized linear models with almost surely bounded features $\Phi(x, y)$, such as conditional random fields [20].

We can now introduce the main technical assumption on the loss ℓ.

Assumption 2 (Generalized self-concordance). *For any $z \in \mathcal{Z}$, the function $\ell_z(\cdot)$ is convex and three times differentiable. Moreover, there exists a set $\varphi(z) \subset \mathcal{H}$ such that it holds :*

$$\forall \theta \in \mathcal{H},\ \forall h, k \in \mathcal{H},\ \left|\nabla^3 \ell_z(\theta)[k, h, h]\right| \leqslant \sup_{g \in \varphi(z)} |k \cdot g|\ \nabla^2 \ell_z(\theta)[h, h].$$

This is a generalization of the assumptions introduced by Bach [2], by allowing a varying term $\sup_{g\in\varphi(z)} |k \cdot g|$ instead of a uniform bound proportional to $\|k\|$. This is crucial for the fast rates we want to show.

Example 2 (Checking assumptions). *For the losses in Example 1, this condition is satisfied with the following corresponding set-function φ.*

- *For the square loss $\ell_z(\theta) = \frac{1}{2}\left(y - \theta \cdot \Phi(x)\right)^2$, $\varphi(z) = \{0\}$.*
- *For the Huber losses $\ell_z(\theta) = \psi(y - \theta \cdot \Phi(x))$, if $\psi(t) = \sqrt{1+t^2} - 1$, then $\varphi(z) = \{3\Phi(x)\}$ and if $\psi(t) = \log \frac{e^t+e^{-t}}{2}$, then $\varphi(z) = \{2\Phi(x)\}$ [25]. For the logistic loss $\ell_z(\theta) = \log(1 + e^{-y\theta\cdot\Phi(x)})$, we have $\varphi(z) = \{y\Phi(x)\}$ (here, $\varphi(z)$ is reduced to a point).*

- *For generalized linear models, $\nabla^3 \ell_z(\theta)$ is a third-order cumulant, and thus $\left|\nabla^3 \ell_z(\theta)[k,h,h]\right| \leqslant \mathbb{E}_{p(y|x,\theta)}|k\cdot\Phi(x,y) - k\cdot\mathbb{E}_{p(y'|x,\theta)}\Phi(x,y')|\cdot|h\cdot\Phi(x,y) - h\cdot\mathbb{E}_{p(y'|x,\theta)}\Phi(x,y')|^2 \leqslant 2\sup_{y\in\mathcal{Y}}|k\cdot\Phi(x,y)|\ \nabla^2\ell_z(\theta)[h,h]$. Therefore $\varphi(z) = \{2\Phi(x,y'),\ y'\in\mathcal{Y}\}$ (which is not a singleton).*

Moreover we require the following two technical assumptions to guarantee that $L(\theta)$ and its first and second derivatives are well defined for any $\theta \in \mathcal{H}$.

Assumption 3 (Boundedness)**.** *There exists $R \geqslant 0$ such that $\sup_{g\in\varphi(Z)}\|g\| \leqslant R$ almost surely.*
Assumption 4 (Definition in 0)**.** *$|\ell_Z(0)|$, $\|\nabla\ell_Z(0)\|$ and $\mathrm{Tr}(\nabla^2\ell_Z(0))$ are almost surely bounded.*

The assumptions above are usually easy to check in practice. In particular, if the support of ρ is bounded, the mappings $z \mapsto \ell_z(0), \nabla\ell_z(0), \mathrm{Tr}(\nabla^2\ell_z(0))$ are continuous, and φ is uniformly bounded on bounded sets, then they hold. The main regularity assumption we make on our statistical problems follows.

Assumption 5 (Existence of a minimizer)**.** *There exists $\theta^\star \in \mathcal{H}$ such that $L(\theta^\star) = \inf_{\theta\in\mathcal{H}} L(\theta)$.*

While Assumption 3 is standard in the analysis of such models [8, 33, 35, 3], Assumption 5 imposes that the model is "well-specified", that is, for supervised learning situations from Example 1, we have chosen a rich enough representation Φ. It is possible to study the non-realizable case in our setting by requiring additional technical assumptions (see [35] or discussion after (6)), but this is out of scope of this paper. Note that our well-specified assumption (for logistic regression for simplicity of arguments) is weaker than requiring $f^\star(x) = \mathbb{E}\,[Y|X]$ being equal to $\theta^\star\cdot\Phi(x)$. We can now introduce the main definitions allowing our bias-variance decomposition.

Definition 1 (Hessian, Bias, Degrees of freedom)**.** *Let $L_\lambda(\theta) = L(\theta)+\frac{\lambda}{2}\|\theta\|^2$; define the* expected Hessian $\mathbf{H}(\theta)$*, the regularized Hessian $\mathbf{H}_\lambda(\theta)$, the* bias Bias_λ *and the* degrees of freedom df_λ *as:*

$$\mathbf{H}(\theta) = \mathbb{E}\left[\nabla^2\ell_Z(\theta)\right], \ \ \textit{and} \ \ \mathbf{H}_\lambda(\theta) = \mathbf{H}(\theta) + \lambda I, \tag{3}$$

$$\mathsf{Bias}_\lambda = \|\mathbf{H}_\lambda(\theta^\star)^{-1/2}\nabla L_\lambda(\theta^\star)\|, \tag{4}$$

$$\mathsf{df}_\lambda = \mathbb{E}\left[\|\mathbf{H}_\lambda(\theta^\star)^{-1/2}\nabla\ell_Z(\theta^\star)\|^2\right]. \tag{5}$$

Note that the bias and degrees of freedom only depend on the optimum $\theta^\star \in \mathcal{H}$ and not on the minimizer $\theta^\star_\lambda$ of the regularized expected risk. Moreover, the degrees of freedom df_λ correspond to the usual Fisher information term commonly seen in the asymptotic analysis of M-estimation [39, 21], and correspond to the usual quantities introduced in the analysis of least-squares [8]. Indeed, in the least-squares case, we recover exactly $\mathsf{Bias}_\lambda = \lambda\|\mathbf{C}_\lambda^{-1/2}\theta^\star\|$ and $\mathsf{df}_\lambda = \mathrm{Tr}(\mathbf{C}\mathbf{C}_\lambda^{-1})$, where $\mathbf{C}$ is the covariance operator $\mathbf{C} = \mathbb{E}\,[\Phi(x)\otimes\Phi(x)]$ and $\mathbf{C}_\lambda = \mathbf{C} + \lambda I$.

Our results will rely on the quadratic approximation of the losses around $\theta^\star$. Borrowing tools from the analysis of Newton's method [24], this will only be possible in the vicinity of $\theta^\star$. The proper notion of vicinity is the so-called *radius of the Dikin ellipsoid*, which we define as follows:

$$\mathsf{r}_\lambda(\theta) \quad \text{such that} \quad 1/\mathsf{r}_\lambda(\theta) = \sup_{z\in\mathrm{supp}(\rho)}\ \sup_{g\in\varphi(z)} \|\mathbf{H}_\lambda{}^{-1/2}(\theta)g\|. \tag{6}$$

Our most refined bounds will depend whether the bias term is small enough compared to $\mathsf{r}_\lambda(\theta^\star)$. We believe that in the non realizable setting, the results we obtain would still hold when the bias term is smaller than the Dikin radius, although one would have to modify the definitions to incorporate the fact that $\theta^\star$ is not in $\mathcal{H}$. The following informal result summarizes all of our results.

Assumptions	Bias	Variance	Optimal λ	Optimal Rate	
None	λ	$\frac{1}{\lambda n}$	$n^{-1/2}$	$n^{-1/2}$	Thm. 2 and Cor. 1
Source	λ^{2r+1}	$\frac{1}{\lambda n}$	$n^{-\frac{1}{2r+2}}$	$n^{-\frac{2r+1}{2r+2}}$	Thm. 3 and Cor. 2
Source + Capacity	λ^{2r+1}	$\frac{1}{\lambda^{1/\alpha} n}$	$n^{-\frac{\alpha}{2r\alpha+\alpha+1}}$	$n^{-\frac{2r\alpha+\alpha}{2r\alpha+\alpha+1}}$	Thm. 4 and Cor. 3

Table 1: Summary of convergence rates, without constants except λ, for source condition (Asm. 6): $\theta^\star \in \mathrm{Im}(\mathbf{H}(\theta^\star)^r),\ r \in (0, 1/2]$, capacity condition (Asm. 7): $\mathsf{df}_\lambda = O(\lambda^{-1/\alpha}),\ \alpha \geqslant 1$.

Theorem 1 (General bound, informal)**.** *Let $n \in \mathbb{N}$, $\delta \in (0, 1/2]$, $\lambda > 0$. Under Assumptions 1 to 5, whenever*

$$n \geqslant C_0 \frac{R^2 \mathsf{df}_\lambda \log \frac{2}{\delta}}{\lambda},$$

then with probability at least $1 - 2\delta$, it holds

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leqslant \mathsf{C}_{\text{bias}}\ \mathsf{Bias}^2_\lambda + \mathsf{C}_{\text{var}} \frac{\mathsf{df}_\lambda \log \frac{2}{\delta}}{n},$$

where C_0, C_{bias} and C_{var} are either universal or depend only on $R\|\theta^\star\|$.

This mimics a usual bias-variance decomposition, with a bias term Bias^2_λ and a variance term proportional to df_λ/n. In particular in the rest of the paper we quantify the constants and the rates under various regularity assumptions, and specify the good choices of the regularization parameter λ. In Table 1, we summarize the different assumptions and corresponding rates.

3 Slow convergence rates

Here we bound the quantity of interest without any regularity assumption (e.g., source of capacity condition) beyond some boundedness assumptions on the learning problem. We consider the various bounds on the derivatives of the loss ℓ:

$$\mathsf{B}_1(\theta) = \sup_{z \in \mathrm{supp}(\rho)} \|\nabla \ell_z(\theta)\|,\ \mathsf{B}_2(\theta) = \sup_{z \in \mathrm{supp}(\rho)} \mathrm{Tr}(\nabla^2 \ell_z(\theta)),\ \overline{\mathsf{B}}_1 = \sup_{\|\theta\| \leqslant \|\theta^\star\|} \mathsf{B}_1(\theta),\ \overline{\mathsf{B}}_2 = \sup_{\|\theta\| \leqslant \|\theta^\star\|} \mathsf{B}_2(\theta).$$

Example 3 (Bounded derivatives)**.** *In all the losses considered above, assume the feature representation ($\Phi(x)$ for the Huber losses and the square loss, $y\Phi(x)$ for the logistic loss, and $\Phi(x, y)$ for GLMs) is bounded by $\bar{R}$. Then the losses considered above apart from the square loss are Lipschitz-continuous and B_1 is uniformly bounded by $\bar{R}$. For these losses, B_2 is also uniformly bounded by $\bar{R}^2$. Using Example 2, one can take $\bar{R}$ to be equal to a constant times R ($1/2$ and $1/3$ for the respective Huber losses, 1 for logistic regression and $1/2$ for canonical GLMs). For the square loss (where $R = 0$ because the third-order derivative is zero), $\overline{\mathsf{B}}_2 \leqslant \bar{R}^2$ and $\overline{\mathsf{B}}_1 \leqslant \bar{R}\|y\|_\infty + \bar{R}^2\|\theta^\star\|$, where $\|y\|_\infty$ is an almost sure bound on the output y.*

Theorem 2 (Basic result)**.** *Let $n \in \mathbb{N}$ and $0 < \lambda \leqslant \overline{\mathsf{B}}_2$. Let $\delta \in (0, 1/2]$. If*

$$n \geqslant 512 \left(\|\theta^\star\|^2 R^2 \vee 1\right) \log \frac{2}{\delta}, \quad n \geqslant 24 \frac{\overline{\mathsf{B}}_2}{\lambda} \log \frac{8\overline{\mathsf{B}}_2}{\lambda\delta}, \quad n \geqslant 256 \frac{R^2 \overline{\mathsf{B}}_1^2}{\lambda^2} \log \frac{2}{\delta},$$

then with probability at least $1 - 2\delta$,

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leqslant 84\, \frac{\overline{\mathsf{B}}_1^2}{\lambda n} \log \frac{2}{\delta} + 2\lambda\|\theta^\star\|^2. \tag{7}$$

This result shown in Appendix C.3 as a consequence of Thm. 6 (also see the proof sketch in Sec. 6) matches the one obtained with Lipschitz-continuous losses [33] and the one for least-squares when assuming the existence of $\theta^\star$ [8]. The following corollary (proved as Thm. 8 in Appendix E) gives the bound optimized in λ, with explicit rates.

Corollary 1 (Basic Rates)**.** *Let* $\delta \in (0, 1/2]$*. Under Assumptions 1 to 5, when* $n \geqslant N, \lambda = C_0\sqrt{\log(2/\delta)/n}$*, then with probability at least* $1 - 2\delta$*,*

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leqslant C_1\, n^{-1/2}\, \log^{1/2}\frac{2}{\delta}.$$

with $C_0 = 16\overline{\mathsf{B}}_1 \max(1, R), C_1 = 48\overline{\mathsf{B}}_1 \max(1, R)\max(1, \|\theta^\star\|^2)$ *and with* N *defined in Eq. (41) and satisfying* $N = O(\mathrm{poly}(\overline{\mathsf{B}}_1, \overline{\mathsf{B}}_2, R\|\theta^\star\|))$ *where* poly *denotes a polynomial function of the inputs.*

Both bias and variance terms are of order $O(1/\sqrt{n})$ and we recover up to constants terms the result of Sridharan et al. [33]. In the next section, we will improve both bias and variance terms to obtain faster rates.

4 Faster Rates with Source Conditions

Here we provide a more refined bound, where we introduce a *source condition* on $\theta^\star$ allowing to improve the bias term and to achieve learning rates as fast as $O(n^{-2/3})$. We first define the localized versions of $\overline{\mathsf{B}}_1, \overline{\mathsf{B}}_2$:

$$\mathsf{B}^\star_1 = \mathsf{B}_1(\theta^\star), \qquad \mathsf{B}^\star_2 = \mathsf{B}_2(\theta^\star),$$

and recall the definition of the bias

$$\mathsf{Bias}_\lambda = \|\mathbf{H}_\lambda(\theta^\star)^{-1/2}\nabla L_\lambda(\theta^\star)\|. \tag{8}$$

Note that since $\theta^\star$ is the minimizer of L, we have $\nabla L(\theta^\star) = 0$, so that $\nabla L_\lambda(\theta^\star) = \nabla L(\theta^\star) + \lambda\theta^\star = \lambda\theta^\star$, and $\mathsf{Bias}_\lambda = \lambda\|\mathbf{H}_\lambda(\theta^\star)^{-1/2}\theta^\star\|$. This characterization is always bounded by $\lambda\|\theta^\star\|^2$, but allows a finer control of the regularity of $\theta^\star$, leading to improved rates compared to Sec. 3.

Note that in the least-squares case, we recover exactly the bias of ridge regression $\mathsf{Bias}_\lambda = \lambda\|\mathbf{C}_\lambda^{-1/2}\theta^\star\|$, where $\mathbf{C}$ is the covariance operator $\mathbf{C} = \mathbb{E}\,[\Phi(x) \otimes \Phi(x)]$.

Using self-concordance, we will relate quantities at $\theta^\star$ to quantities at $\theta^\star_\lambda$ using:

$$\mathsf{t}_\lambda = \sup_{z \in \mathrm{supp}(\rho)}\ \sup_{g \in \varphi(z)} |(\theta^\star_\lambda - \theta^\star) \cdot g|.$$

The following theorem, proved in Appendix D.4, relates Bias_λ to the excess risk.

Theorem 3 (Decomposition with refined bias)**.** *Let* $n \in \mathbb{N}$, $\delta \in (0, 1/2]$, $0 < \lambda \leqslant \mathsf{B}^\star_2$*. Whenever*

$$n \geqslant \triangle_1 \frac{\mathsf{B}^\star_2}{\lambda}\log\frac{8\square_1^2\mathsf{B}^\star_2}{\lambda\delta}, \quad n \geqslant \triangle_2\,\frac{(\mathsf{B}^\star_1 R)^2}{\lambda^2}\,\log\frac{2}{\delta},$$

then with probability at least $1 - 2\delta$*, it holds*

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leqslant \mathsf{C}_{\text{bias}}\,\mathsf{Bias}_\lambda^2 + \mathsf{C}_{\text{var}}\,\frac{(\mathsf{B}^\star_1)^2}{\lambda n}\,\log\frac{2}{\delta}, \tag{9}$$

where $\square_1 \leqslant e^{\mathsf{t}_\lambda/2}, \triangle_1 \leqslant 2304e^{4\mathsf{t}_\lambda}(1/2 \vee R\|\theta^\star\|), \triangle_2 \leqslant 256e^{2\mathsf{t}_\lambda}, \mathsf{C}_{\text{bias}} \leqslant 6e^{2\mathsf{t}_\lambda}, \mathsf{C}_{\text{var}} \leqslant 256e^{3\mathsf{t}_\lambda}$*.*

It turns out that the *radius of the Dikin ellipsoid* $\mathsf{r}_\lambda(\theta^\star)$ defined in Eq. (6) provides the sufficient control over the constants above: when the bias is of the same order of the radius of the Dikin ellipsoid, the quantities $\mathsf{C}_{\text{bias}}, \mathsf{C}_{\text{var}}, \triangle_1, \triangle_2$ become universal constants instead of depending exponentially on $R\|\theta^\star\|$, as shown by the lemma below, proved in Lemma 4 in Appendix D.

Lemma 1. *When* $\mathsf{Bias}_\lambda \leqslant \frac{\mathsf{r}_\lambda(\theta^\star)}{2}$ *then* $\mathsf{t}_\lambda \leqslant \log 2$ *else* $\mathsf{t}_\lambda \leqslant 2R\|\theta^\star\|$*.*

Interestingly, regularity of $\theta^\star$, like the source condition below, can induce this effect, allowing a better dependence on λ for the bias term.

Assumption 6 (Source condition)**.** *There exists $r \in (0, 1/2]$ and $v \in \mathcal{H}$ such that $\theta^\star = \mathbf{H}(\theta^\star)^r v$.*

In particular we denote by $\mathsf{L} := \|v\|$. Assumption 6 is commonly made in least-squares regression [8, 35, 5] and is equivalent to requiring that, when expressing $\theta^\star$ with respect to the eigenbasis of $\mathbf{H}(\theta^\star)$, i.e., $\theta^\star = \sum_{j\in\mathbb{N}} \alpha_j u_j$, where λ_j, u_j is the eigendecomposition of $\mathbf{H}(\theta^\star)$, and $\alpha_j = \theta \cdot u_j$, then α_j decays as λ_j^r. In particular, with this assumption, defining $\beta_j = v \cdot u_j$,

$$\mathsf{Bias}_\lambda^2 = \lambda^2 \sum_j \frac{\alpha_j^2}{\lambda_j + \lambda} = \lambda^2 \sum_j \frac{\lambda_j^{2r}\beta_j^2}{\lambda_j + \lambda} \leqslant \lambda^2 \Big(\sup_j \frac{\lambda_j^{2r}}{\lambda_j + \lambda} \Big) \sum_j \beta_j^2 \leqslant \lambda^{1+2r} \|v\|^2.$$

Note moreover that $\mathbf{H}(\theta^\star) \preccurlyeq \mathsf{B}_2^\star \mathbf{C}$, meaning that the usual sufficient conditions leading to the source conditions for least-squares also apply here. For example, for logistic regression, if the log-odds ratio is smooth enough, then it is in $\mathcal{H}$. So, when $\mathcal{H}$ corresponds to a Sobolev space of smoothness m and the marginal of ρ on the input space is a density bounded away from 0 and infinity with bounded support, then the source condition corresponds essentially to requiring $\theta^\star$ to be $(1 + 2r)m$-times differentiable [see discussion after Thm. 9 of 35, for more details]. A precise example can be found in Sec. 4.1 of [26].

In conclusion, the effect of additional regularity for $\theta^\star$ as Assumption 6, has two beneficial effects: (a) on one side it allows to obtain faster rates as shown in the next corollary, (b) as mentioned before, somewhat surprisingly, it reduces the constants to universal, since it allows the bias to go to zero faster than the Dikin radius (indeed, the squared radius $\mathsf{r}_\lambda^2(\theta^\star)$ is always larger than λ/R^2, which is strictly larger than $\lambda^{1+2r}\|v\|^2$ if $r > 0$ and λ small enough). This is why we do not the get exponential constants imposed by Hazan et al. [16].

Corollary 2 (Rates with source condition)**.** *Let $\delta \in (0, 1/2]$. Under Assumptions 1 to 5 and Assumption 6, whenever $n \geqslant N$ and $\lambda = (C_0/n)^{1/(2+2r)}$, then with probability at least $1 - 2\delta$,*

$$L(\widehat{\theta}_\lambda^\star) - L(\theta^\star) \quad \leqslant \quad C_1 \;\; n^{-\frac{1+2r}{2+2r}} \log \frac{2}{\delta},$$

with $C_0 = 256\ (\mathsf{B}_1^\star/\mathsf{L})^2,\ C_1 = 8\ (256)^\gamma((\mathsf{B}_1^\star)^\gamma \mathsf{L}^{1-\gamma})^2, \gamma = \frac{1+2r}{2+2r}$ and with N defined in Eq. (48) and satisfying $N = O(\mathrm{poly}(\mathsf{B}_1^\star, \mathsf{B}_2^\star, \mathsf{L}, R, \log(1/\delta)))$.

The corollary above, derived in Appendix F, is obtained by minimizing in λ the r.h.s. side of Eq. (9) in Thm. 3, and considering that when $\theta^\star$ satisfies the source condition, then $\mathsf{Bias}_\lambda \leqslant \lambda^{1+2r}\mathsf{L}$, while the variance is still of the form $1/(\lambda n)$. When r is close to 0, the rate $1/\sqrt{n}$ is recovered. When instead the target function is more regular, implying $r = 1/2$, a rate of $n^{-2/3}$ is achieved. Two considerations are in order: (a) the obtained rate is the same as least-squares and minimax optimal [8, 35, 5], (b) the fact that regularized ERM is adaptive to the regularity of the function up to $r = 1/2$ is a byproduct of Tikhonov regularization as already shown for the least-squares case by Gerfo et al. [14]. Using different regularization techniques may remove the limit $r = 1/2$.

5 Fast Rates with both Source and Capacity Conditions

In this section, we consider improved results with a finer control of the effective dimension df_λ (often called degrees of freedom), which, together with the source condition allows to achieve rates as fast as $1/n$:

$$\mathsf{df}_\lambda = \mathbb{E}\left[\|\mathbf{H}_\lambda(\theta^\star)^{-1/2}\nabla \ell_Z(\theta^\star)\|^2\right],$$

As mentioned earlier this definition of df_λ corresponds to the usual asymptotic term in M-estimation. Moreover, in the case of least-squares, it corresponds to the standard notion of effective dimension $\mathsf{df}_\lambda = \mathrm{Tr}(\mathbf{C}\mathbf{C}_\lambda^{-1})$ [8, 5]. Note that by definition, we always have $\mathsf{df}_\lambda \leqslant \mathsf{B}_1^{\star 2}/\lambda$, but we can have in general a much finer control. For example, for least-squares, $\mathsf{df}_\lambda = O(\lambda^{-1/\alpha})$ if the eigenvalues of the covariance operator $\mathbf{C}$ decay as $\lambda_j(\mathbf{C}) = O(j^{-\alpha})$, for $\alpha \geqslant 1$. Moreover note that since $\mathbf{C}$ is trace-class, by Asm. 3, the eigenvalues form a summable sequence and so $\mathbf{C}$ satisfies $\lambda_j(\mathbf{C}) = O(j^{-\alpha})$ with α always larger than 1.

Example 4 (Generalized linear models)**.** *For generalized linear models, an extra assumption makes the degrees of freedom particularly simple: if the probabilistic model is well-specified, that is, there exists $\theta^\star$ such that almost surely, $p(y|x) = p(y|x,\theta^\star) = \frac{\exp(\theta^\star \cdot \Phi(x,y))}{\int_{\mathcal{Y}} \exp(\theta^\star \cdot \Phi(x,y')) d\mu(y')}$, then from the usual Bartlett identities [4] relating the expected squared derivatives and Hessians, we have $\mathbb{E}\left[\nabla \ell_z(\theta^\star) \otimes \nabla \ell_z(\theta^\star)\right] = \mathbf{H}(\theta^\star)$, leading to $\mathsf{df}_\lambda = \mathrm{Tr}(\mathbf{H}_\lambda(\theta^\star)^{-1}\mathbf{H}(\theta^\star))$.*

As we have seen in the previous example there are interesting problems for which $\mathsf{df}_\lambda = \mathrm{Tr}(\mathbf{H}(\theta^\star) + \lambda I)^{-1}\mathbf{H}(\theta^\star))$. Since we have $\mathbf{H}(\theta^\star) \preceq \mathsf{B}_2^\star \mathbf{C}$, df_λ still enjoys a polynomial decay depending on the eigenvalue decay of $\mathbf{C}$ as observed for least-squares. In the finite-dimensional setting where $\mathcal{H}$ is of dimension d, note that in this case, df_λ is always bounded by d. Now we are ready to state our result in the most general form, proved in Appendix D.4.

Theorem 4 (General bound)**.** *Let $n \in \mathbb{N}$, $\delta \in (0, 1/2]$, $0 < \lambda \leqslant \mathsf{B}_2^\star$. Whenever*

$$n \geqslant \triangle_1 \frac{\mathsf{B}_2^\star}{\lambda} \log \frac{8\square_1^2 \mathsf{B}_2^\star}{\lambda\delta}, \quad n \geqslant \triangle_2 \frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2}{\mathsf{r}_\lambda(\theta^\star)^2} \log \frac{2}{\delta},$$

with $(\mathsf{Q}^\star)^2 = \mathsf{B}_1^{\star 2}/\mathsf{B}_2^\star$, then with probability at least $1 - 2\delta$, it holds

$$L(\widehat{\theta}_\lambda^\star) - L(\theta^\star) \leqslant \mathsf{C}_{\text{bias}}\, \mathsf{Bias}_\lambda^2 + \mathsf{C}_{\text{var}} \frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2}{n} \log \frac{2}{\delta}, \tag{10}$$

where, $\mathsf{C}_{\text{bias}}, \mathsf{C}_{\text{var}}, \square_1 \leqslant 414$, $\triangle_1$, $\triangle_2 \leqslant 5184$ when $\mathsf{Bias}_\lambda \leqslant \mathsf{r}_\lambda(\theta^\star)/2$; otherwise C_{bias}, C_{var}, $\square_1 \leqslant 256 e^{6R\|\theta^\star\|}$, $\triangle_1$, $\triangle_2 \leqslant 2304(1 + R\|\theta^\star\|)^2 e^{8R\|\theta^\star\|}$.

As shown in the theorem above, the variance term depends on df_λ/n, implying that, when df_λ has a better dependence in λ than $1/\lambda$, it is possible to achieve faster rates. We quantify this with the following assumption.

Assumption 7 (Capacity condition)**.** *There exists $\alpha > 0$ and $\mathsf{Q} \geqslant 0$ such that $\mathsf{df}_\lambda \leqslant \mathsf{Q}\lambda^{-1/\alpha}$.*

Assumption 7 is standard in the context of least-squares, [8] and in many interesting settings is implied by the eigenvalue decay order of $\mathbf{H}(\theta^\star)$, or $\mathbf{C}$ as discussed above. In the following corollary we quantify the effect of df_λ in the learning rates.

Corollary 3. *Let $\delta \in (0, 1/2]$. Under Assumptions 1 to 5, Assumption 6 and Assumption 7, when $n \geqslant N$ and $\lambda = (C_0/n)^{\alpha/(1+\alpha(1+2r))}$, then with probability at least $1 - 2\delta$,*

$$L(\widehat{\theta}_\lambda^\star) - L(\theta^\star) \leqslant C_1 n^{-\frac{\alpha(1+2r)}{1+\alpha(1+2r)}} \log \frac{2}{\delta},$$

with $C_0 = 256(\mathsf{Q}/\mathsf{L})^2$, $C_1 = 8(256)^\gamma\ (\mathsf{Q}^\gamma\ \mathsf{L}^{1-\gamma})^2$, $\gamma = \frac{\alpha(1+2r)}{1+\alpha(1+2r)}$ and N defined in Eq. (48) and satisfying $N = O(\mathrm{poly}(\mathsf{B}_1^\star, \mathsf{B}_2^\star, \mathsf{L}, \mathsf{Q}, R, \log(1/\delta)))$.

The result above is derived in Cor. 4 in Appendix F and is obtained by bounding Bias_λ with $\lambda^{1+2r}\mathsf{L}$ due to the source condition, and df_λ with $\lambda^{-1/\alpha}$ due to the capacity condition and then optimizing the r.h.s. of Eq. (10) in λ. Note that (a) the learning rate under the considered assumptions is the same as least-squares and minimax optimal [8], and (b) when $\alpha = 1$ the same rate of Cor. 2 is achieved, which can be as fast as $n^{-2/3}$, otherwise, when $\alpha \gg 1$, we achieve a learning rate in the order of $1/n$, for $\lambda = n^{-1/(1+2r)}$.

6 Sketch of the proof

In this section we will use the notation $\|v\|_{\mathbf{A}} := \|\mathbf{A}^{1/2}v\|$, with $v \in \mathcal{H}$ and $\mathbf{A}$ a bounded positive semi-definite operator on $\mathcal{H}$. Here we prove that the excess risk decomposes using the bias term Bias_λ defined in Eq. (8) and a variance term V_λ, where V_λ is defined as

$$V_\lambda := \|\nabla \widehat{L}_\lambda(\theta^\star_\lambda)\|_{\mathbf{H}_\lambda^{-1}(\theta^\star_\lambda)}, \quad \text{with} \quad \widehat{L}_\lambda(\cdot) = \frac{1}{n}\sum_{i=1}^n \ell_{z_i}(\cdot) + \frac{\lambda}{2}\|\cdot\|^2,$$

which in turn is a random variable that concentrate in high probability to $\sqrt{\mathsf{df}_\lambda/n}$.

Required tools. To proceed with the proof we need two main tools. The first is a result on the equivalence of norms of the empirical Hessian $\widehat{\mathbf{H}}_\lambda(\theta) = \nabla^2\widehat{L}_\lambda(\theta)$ w.r.t. the true Hessian $\mathbf{H}_\lambda(\theta) = \nabla^2 L_\lambda(\theta)$ for $\lambda > 0$ and $\theta \in \mathcal{H}$. The result is proven in Lemma 6 of Appendix D.3, using Bernstein inequalities for Hermitian operators [36], and essentially states that for $\delta \in (0, 1]$, whenever $n \geqslant \frac{24\mathsf{B}_2(\theta)}{\lambda}\log\frac{8\mathsf{B}_2(\theta)}{\lambda\delta}$, then with probability $1-\delta$, it holds

$$\|\cdot\|_{\mathbf{H}_\lambda(\theta)} \leqslant 2\|\cdot\|_{\widehat{\mathbf{H}}_\lambda(\theta)}, \qquad \|\cdot\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta)} \leqslant 2\|\cdot\|_{\mathbf{H}_\lambda^{-1}(\theta)}. \tag{11}$$

The second result is about localization properties induced by generalized self-concordance on the risk. We express the result with respect to a generic probability μ (we will use it with $\mu = \rho$ and $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$). Let μ be a probability distribution with support contained in the support of ρ. Denote by $L_\mu(\theta)$ the risk $L_\mu(\theta) = \mathbb{E}_{z\sim\mu}[\ell_z(\theta)]$ and by $L_{\mu,\lambda}(\theta) = L_\mu(\theta) + \frac{\lambda}{2}\|\theta\|^2$ (then $L_{\mu,\lambda} = L_\lambda$ when $\mu = \rho$, or $\widehat{L}_\lambda$ when $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$).

Proposition 1. *Under Assumptions 2 to 4, the following holds: (a) $L_{\mu,\lambda}(\theta), \nabla L_{\mu,\lambda}(\theta), \mathbf{H}_{\mu,\lambda}(\theta)$ are defined for all $\theta \in \mathcal{H}, \lambda \geqslant 0$, (b) for all $\lambda > 0$, there exists a unique $\theta^\star_{\mu,\lambda} \in \mathcal{H}$ minimizing $L_{\mu,\lambda}$ over $\mathcal{H}$, and (c) for all $\lambda > 0$ and $\theta \in \mathcal{H}$,*

$$\mathbf{H}_{\mu,\lambda}(\theta) \preceq e^{\mathsf{t}_0}\mathbf{H}_{\mu,\lambda}(\theta^\star_{\mu,\lambda}), \tag{12}$$

$$L_{\mu,\lambda}(\theta) - L_{\mu,\lambda}(\theta^\star_{\mu,\lambda}) \leqslant \psi(\mathsf{t}_0)\|\theta - \theta^\star_{\mu,\lambda}\|^2_{\mathbf{H}_{\mu,\lambda}(\theta^\star_{\mu,\lambda})}, \tag{13}$$

$$\underline{\phi}(\mathsf{t}_0)\|\theta - \theta^\star_{\mu,\lambda}\|_{\mathbf{H}_{\mu,\lambda}(\theta)} \leqslant \|\nabla L_{\mu,\lambda}(\theta)\|_{\mathbf{H}^{-1}_{\mu,\lambda}(\theta)}, \tag{14}$$

(d) Eqs. (12) and (13) hold also for $\lambda = 0$, provided that $\theta^\star_{\mu,0}$ exists. Here $\mathsf{t}_0 := \mathsf{t}(\theta - \theta^\star_{\mu,\lambda})$ and $\underline{\phi}(t) = (1-e^{-t})/t, \ \ \psi(t) = (e^t - t - 1)/t^2$.

The result above is proved in Appendix B.1 and is essentially an extension of results by [2] applied to $L_{\mu,\lambda}$ under Assumptions 2 to 4.

Sketch of the proof. Now we are ready to decompose the excess risk using our bias and variance terms. In particular we will sketch the decomposition without studying the terms that lead to constants terms. For the complete proof of the decomposition see Thm. 7 in Appendix D.1. Since $\theta^\star$ exists by Assumption 5, using Eq. (13), applied with $\mu = \rho$ and $\lambda = 0$, we have $L(\theta) - L(\theta^\star) \leqslant \psi(\mathsf{t}(\theta - \theta^\star))\|\theta - \theta^\star\|^2_{\mathbf{H}(\theta^\star)}$ for any $\theta \in \mathcal{H}$. By setting $\theta = \widehat{\theta}^\star_\lambda$, we obtain

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leqslant \psi(\mathsf{t}(\widehat{\theta}^\star_\lambda - \theta^\star))\|\widehat{\theta}^\star_\lambda - \theta^\star\|^2_{\mathbf{H}(\theta^\star)}.$$

The term $\psi(\mathsf{t}(\widehat{\theta}^\star_\lambda - \theta^\star))$ will become a constant. For the sake of simplicity, in this sketch of proof we will not deal with it nor with other terms of the form $\mathsf{t}(\cdot)$ leading to constants. On the other hand, the term $\|\widehat{\theta}^\star_\lambda - \theta^\star\|^2_{\mathbf{H}(\theta^\star)}$ will yield our bias and variance terms. Using the fact that $\mathbf{H}(\theta^\star) \preceq \mathbf{H}(\theta^\star) + \lambda I =: \mathbf{H}_\lambda(\theta^\star)$, by adding and subtracting $\theta^\star_\lambda$, we have

$$\|\theta^\star_\lambda - \theta^\star\|_{\mathbf{H}(\theta^\star)} \leqslant \|\theta^\star_\lambda - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)} \leqslant \|\theta^\star_\lambda - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)} + \|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star)},$$

so

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leqslant \text{const.} \times (\|\theta^\star_\lambda - \theta^\star\|^2_{\mathbf{H}_\lambda(\theta^\star)} + \|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star)})^2.$$

By applying Eq. (12) with $\mu = \rho$ and $\theta = \theta^\star$, we have $\mathbf{H}_\lambda(\theta^\star) \preceq e^{\mathsf{t}_\lambda}\mathbf{H}_\lambda(\theta^\star_\lambda)$ and so we further bound $\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star)}$ with $e^{\mathsf{t}_\lambda/2}\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)}$ obtaining

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leqslant \text{const.} \times (\|\theta^\star_\lambda - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)} + e^{\mathsf{t}_\lambda/2}\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)})^2.$$

The term $\|\theta^\star_\lambda - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)}$ will lead to the *bias terms*, while the term $\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)}$ will lead to the *variance term*.

Bounding the bias terms. Recall the definition of bias $\mathsf{Bias}_\lambda = \|\nabla L_\lambda(\theta^\star)\|_{\mathbf{H}_\lambda^{-1}(\theta^\star)}$ and of the constant $\mathsf{t}_\lambda := \mathsf{t}(\theta^\star - \theta^\star_\lambda)$. We bound $\|\theta^\star - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star)}$ by applying Eq. (14) with $\mu = \rho$ and $\theta = \theta^\star$

$$\|\theta^\star - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star)} \leqslant 1/\underline{\phi}(\mathsf{t}_\lambda)\ \|\nabla L_\lambda(\theta^\star)\|_{\mathbf{H}_\lambda^{-1}(\theta^\star)} = 1/\underline{\phi}(\mathsf{t}_\lambda)\ \mathsf{Bias}_\lambda.$$

Bounding the variance terms. To bound the term $\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)}$, we assume n large enough to apply Eq. (11) in high probability. Thus, we obtain

$$\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)} \leqslant 2\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star_\lambda)}.$$

Applying Eq. (14) with $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$ and $\theta = \widehat{\theta}^\star_\lambda$, since $L_{\mu,\lambda} = \widehat{L}_\lambda$ for the given choice of μ,

$$\|\theta^\star_\lambda - \widehat{\theta}^\star_\lambda\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star_\lambda)} \leqslant \|\nabla \widehat{L}_\lambda(\theta^\star_\lambda)\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star_\lambda)}\ /\ \underline{\phi}(\mathsf{t}(\theta^\star_\lambda - \widehat{\theta}^\star_\lambda)),$$

and applying Eq. (11) in high probability again, we obtain

$$\|\nabla \widehat{L}_\lambda(\theta^\star_\lambda)\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star_\lambda)} \leqslant 2\|\nabla \widehat{L}_\lambda(\theta^\star_\lambda)\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star_\lambda)}.$$

Bias-variance decomposition. A technical part of the proof relates $\|\nabla \widehat{L}_\lambda(\theta^\star_\lambda)\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star_\lambda)}$ with $\|\nabla \widehat{L}_\lambda(\theta^\star)\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)} =: V_\lambda$, by many applications of Prop. 1. Here we assume it is done, obtaining

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leqslant \text{const.} \times (\mathsf{Bias}^2_\lambda + V^2_\lambda).$$

From V_λ to $\sqrt{\mathsf{df}_\lambda/n}$. By construction, $\nabla \widehat{L}_\lambda(\theta^\star_\lambda) = \frac{1}{n}\sum_{i=1}^n \zeta_i$, with $\zeta_i := \nabla \ell_{z_i}(\theta^\star_\lambda) + \lambda\theta^\star_\lambda$. Moreover since the z_i's are i.i.d. samples from ρ, $\mathbb{E}[\zeta_i] = \nabla L_\lambda(\theta^\star_\lambda)$. Finally since $\theta^\star_\lambda$ is the minimizer of L_λ, $\nabla L_\lambda(\theta^\star_\lambda) = 0$. Thus $\nabla \widehat{L}_\lambda(\theta^\star_\lambda)$ is the average of n i.i.d. zero-mean random vectors, and so the variance of V_λ is exactly

$$\mathbb{E}\left[V^2_\lambda\right] = \frac{1}{n}\mathbb{E}\left[\|\mathbf{H}_\lambda^{-1/2}(\theta^\star)\nabla \ell_Z(\theta^\star)\|^2\right] = \frac{\mathsf{df}_\lambda}{n}.$$

Finally, by using Bernstein inequality for random vectors [e.g., 42, Thm. 3.3.4], we bound V_λ roughly with $\sqrt{\mathsf{df}_\lambda \log(2/\delta)/n}$ in high probability.

7 Conclusion

In this paper we have presented non-asymptotic bounds with faster rates than $O(1/\sqrt{n})$, for regularized empirical risk minimization with self-concordant losses such as the logistic loss. It would be interesting to extend our work to algorithms used to minimize the empirical risk, in particular stochastic gradient descent or Newton's method.

Acknowledgments

The second author is supported by the ERCIM Alain Bensoussan Fellowship. We acknowledge support from the European Research Council (grant SEQUOIA 724063).

References

[1] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[2] Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

[3] Francis Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(1):595–627, 2014.

[4] M. S. Bartlett. Approximate confidence intervals. *Biometrika*, 40(1/2):12–19, 1953.

[5] Gilles Blanchard and Nicole Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18(4):971–1013, 2018.

[6] Léon Bottou and Olivier Bousquet. The trade-offs of large scale learning. In *Advances in Neural Information Processing systems*, pages 161–168, 2008.

[7] Stéphane Boucheron and Pascal Massart. A high-dimensional Wilks phenomenon. *Probability Theory and Related Fields*, 150(3-4):405–433, 2011.

[8] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, 7(3):331–368, July 2007.

[9] Nicolo Cesa-Bianchi, Yishay Mansour, and Ohad Shamir. On the complexity of learning with kernels. In *Conference on Learning Theory*, pages 297–325, 2015.

[10] Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

[11] Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithm. *arXiv preprint arXiv:1702.07254*, 2017.

[12] Dylan J. Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic regression: The importance of being improper. *Proceedings of COLT*, 2018.

[13] Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

[14] L. Lo Gerfo, Lorenzo Rosasco, Francesca Odone, E. De Vito, and Alessandro Verri. Spectral algorithms for supervised learning. *Neural Computation*, 20(7):1873–1897, 2008.

[15] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons, 2011.

[16] Elad Hazan, Tomer Koren, and Kfir Y. Levy. Logistic regression: tight bounds for stochastic and online optimization. In *Proceedings of The 27th Conference on Learning Theory*, volume 35, pages 197–209, 2014.

[17] Arthur E. Hoerl and Robert W. Kennard. Ridge regression iterative estimation of the biasing parameter. *Communications in Statistics-Theory and Methods*, 5(1):77–88, 1976.

[18] S. Sathiya Keerthi, K. B. Duan, Shirish Krishnaj Shevade, and Aun Neow Poo. A fast dual algorithm for kernel logistic regression. *Machine learning*, 61(1-3):151–165, 2005.

[19] Tomer Koren and Kfir Levy. Fast rates for exp-concave empirical risk minimization. In *Advances in Neural Information Processing Systems*, pages 1477–1485, 2015.

[20] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2001.

[21] Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer Science & Business Media, 2006.

[22] P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Chapman and Hall, 1989.

[23] Nishant A. Mehta. Fast rates with high probability in exp-concave statistical learning. Technical Report 1605.01288, ArXiv, 2016.

[24] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13. SIAM, 1994.

[25] Dmitrii Ostrovskii and Francis Bach. Finite-sample analysis of M-estimators using self-concordance. Technical Report 1810.06838, arXiv, 2018.

[26] Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 8125–8135, 2018.

[27] Alexander Rakhlin and Karthik Sridharan. Online nonparametric regression with general loss functions. *arXiv preprint arXiv:1501.06598*, 2015.

[28] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems 30*, pages 3215–3225. Curran Associates, Inc., 2017.

[29] Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. Falkon: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems*, pages 3888–3898, 2017.

[30] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for SVM. *Mathematical programming*, 127(1):3–30, 2011.

[31] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[32] Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007.

[33] Karthik Sridharan, Shai Shalev-Shwartz, and Nathan Srebro. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems 21*, pages 1545–1552. 2009.

[34] Ingo Steinwart and Clint Scovel. Fast rates for support vector machines using gaussian kernels. *The Annals of Statistics*, 35(2):575–607, 2007.

[35] Ingo Steinwart, Don R. Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *Proc. COLT*, 2009.

[36] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

[37] Stephen Tu, Rebecca Roelofs, Shivaram Venkataraman, and Benjamin Recht. Large scale kernel learning using block coordinate descent. *arXiv preprint arXiv:1602.05310*, 2016.

[38] Sara A. Van de Geer. High-dimensional generalized linear models and the Lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

[39] Aad W. Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

[40] Tim Van Erven, Peter D. Grünwald, Nishant A. Mehta, Mark D. Reid, and Robert C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16:1793–1861, 2015.

[41] Grace Wahba. *Spline Models for Observational Data*, volume 59. SIAM, 1990.

[42] Vadim Yurinsky. *Sums and Gaussian vectors, volume 1617 of Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1995.

Organization of the Appendix

A Setting, definitions, assumptions

Let $\mathcal{Z}$ be a Polish space and Z a random variable on $\mathcal{Z}$ whith law ρ. Let $\mathcal{H}$ be a separable (non-necessarily finite) Hilbert space and let $\ell : \mathcal{Z} \times \mathcal{H} \to \mathbb{R}$ be a loss function; we denote by $\ell_z(\cdot)$ the function $\ell(z, \cdot)$. Our goal is to solve

$$\inf_{\theta \in \mathcal{H}} L(\theta), \quad \text{with} \quad L(\theta) = \mathbb{E}\left[\ell_Z(\theta)\right].$$

Given $(z_i)_{i=1}^n$ we will consider the following estimator

$$\widehat{\theta}^\star_\lambda = \arg\min_{\theta \in \mathcal{H}} \widehat{L}_\lambda(\theta), \quad \text{with} \quad \widehat{L}_\lambda(\theta) := \frac{1}{n}\sum_{i=1}^n \ell_{z_i}(\theta) + \frac{\lambda}{2}\|\theta\|^2.$$

The goal of this work is to give upper bounds in high probability to the so called *excess risk*

$$L(\widehat{\theta}^\star_\lambda) - \inf_{\theta \in \mathcal{H}} L(\theta).$$

In the rest of this introduction we will introduce the basic assumptions required to make $\widehat{\theta}^\star_\lambda$ and the *excess risk* well defined, and we will introduce basic objects that are needed for the proofs.

First we introduce some notation we will use in the rest of the appendix: let $\lambda \geqslant 0$, $\theta \in \mathcal{H}$ and $\mathbf{A}$ be a bounded positive semidefinite Hermitian operator on $\mathcal{H}$, we denote by $\mathbf{I}$, the identity operator and

$$\|f\|_{\mathbf{A}} := \|\mathbf{A}^{1/2} f\|, \tag{15}$$

$$\mathbf{A}_\lambda := \mathbf{A} + \lambda \mathbf{I}, \tag{16}$$

$$\ell_z^\lambda(\theta) := \ell_z(\theta) + \frac{\lambda}{2}\|\theta\|^2, \tag{17}$$

$$L_\lambda(\theta) := L(\theta) + \frac{\lambda}{2}\|\theta\|^2. \tag{18}$$

Now we recall the assumptions we require on the loss function ℓ, ρ, $(z_i)_{1 \leqslant i \leqslant n}$.

Assumption 1 (i.i.d. data)**.** *The samples $(z_i)_{1 \leqslant i \leqslant n}$ are independently and identically distributed according to ρ.*

Assumption 8 (Generalized self-concordance)**.** *The mapping $z \mapsto \ell_z(\theta)$ is measurable for all $\theta \in \mathcal{H}$ and for any $z \in \mathcal{Z}$, the function ℓ_z is convex and three times differentiable. Moreover, there exists a set $\varphi(z) \subset \mathcal{H}$ such that it holds:*

$$\forall \theta \in \mathcal{H},\ \forall h, k \in \mathcal{H},\ \left|\nabla^3 \ell_z(\theta)[k, h, h]\right| \leqslant \sup_{g \in \varphi(z)} |k \cdot g| \ \nabla^2 \ell_z(\theta)[h, h].$$

Assumption 3 (Boundedness)**.** *There exists $R \geqslant 0$ such that $\sup_{g \in \varphi(Z)} \|g\| \leqslant R$ almost surely.*

Assumption 4 (Definition in 0)**.** $|\ell_Z(0)|$, $\|\nabla \ell_Z(0)\|$ *and* $\mathrm{Tr}(\nabla^2 \ell_Z(0))$ *are almost surely bounded.*

Introduce the following definitions.

Definition 2. *Let $\lambda > 0$, $\theta \in \mathcal{H}$. We introduce*

$$\mathsf{B}_1(\theta) = \sup_{z \in \mathrm{supp}(\rho)} \|\nabla \ell_z(\theta)\|, \qquad \mathsf{B}_2(\theta) = \sup_{z \in \mathrm{supp}(\rho)} \mathrm{Tr}\left(\nabla^2 \ell_z(\theta)\right). \tag{19}$$

$$\mathbf{H}(\theta) = \mathbb{E}\left[\nabla^2 \ell_Z(\theta)\right], \qquad \mathbf{H}_\lambda(\theta) = \mathbf{H}(\theta) + \lambda \mathbf{I}. \tag{20}$$

$$\theta_\lambda^\star = \arg\min_{\theta \in \mathcal{H}} L_\lambda(\theta). \tag{21}$$

Proposition 2. *Under Assumptions 3, 4 and 8, $\mathsf{B}_1(\theta), \mathsf{B}_2(\theta), L(\theta), \nabla L(\theta), \mathbf{H}(\theta), \theta_\lambda^\star$ exist for any $\theta \in \mathcal{H}, \lambda > 0$. Moreover $\nabla L = \mathbb{E}\left[\nabla \ell_Z(\theta)\right]$, $\mathbf{H}(\theta) = \nabla^2 L(\theta)$ and $\mathbf{H}(\theta)$ is trace class.*

Proof. We start by proving, using the assumptions, that $\mathsf{B}_2, \mathsf{B}_1$ and $\theta \mapsto \sup_{z \in \mathrm{supp}(\rho)} |\ell_z(\theta)|$ are all locally bounded (see Lemmas 11 to 13). This allows us to show that $\ell_z(\theta)$, $\nabla \ell_z(\theta)$ and $\mathrm{Tr}(\nabla^2 \ell_z(\theta))$ are uniformly integrable on any ball of finite radius. The fact that $\theta_\lambda^\star$ exists is due to the strong convexity of the function L_λ.

Proposition 3. *Under Assumptions 1, 4 and 8, when $\lambda > 0$, $\widehat{\theta}_\lambda^\star$ exists and is unique.*

Proof. By Assumption 1 we know that $z_1, \ldots, z_n$ are in the support of ρ. Thus, by Assumption 4, $\frac{1}{n}\sum_{i=1}^n \ell_{z_i}$ is finite valued in 0. Since $\frac{1}{n}\sum_{i=1}^n \ell_{z_i}$ is convex three times differentiable as a sum of such functions, it is real-valued on $\mathcal{H}$ and hence $\widehat{L}_\lambda$ is real-valued on $\mathcal{H}$; by strong convexity, $\widehat{\theta}_\lambda^\star$ exists and is unique.

Recall that we also make the following regularity assumption.

Assumption 5 (Existence of a minimizer)**.** *There exists $\theta^\star \in \mathcal{H}$ such that $L(\theta^\star) = \inf_{\theta \in \mathcal{H}} L(\theta)$.*

Finally we conclude with the following definitions that will be used later.

Definition 3. *For $\theta \in \mathcal{H}$, denote by $\mathsf{t}(\theta)$ the function*

$$\mathsf{t}(\theta) = \sup_{z \in \mathrm{supp}(\rho)} \left(\sup_{g \in \varphi(z)} |\theta \cdot g| \right),$$

and define

$$\mathsf{Bias}_\lambda = \|\nabla L_\lambda(\theta^\star)\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)}, \tag{22}$$

$$\widehat{\mathsf{Var}}_\lambda = \|\mathbf{H}_\lambda^{1/2}(\theta_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta_\lambda^\star)\|^2 \, \|\nabla \widehat{L}_\lambda(\theta_\lambda^\star)\|_{\mathbf{H}_\lambda^{-1}(\theta_\lambda^\star)}, \tag{23}$$

$$\mathsf{df}_\lambda = \mathbb{E}\left[\|\nabla \ell_Z(\theta^\star)\|^2_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)}\right], \tag{24}$$

$$\mathsf{t}_\lambda = \mathsf{t}(\theta^\star - \theta_\lambda^\star), \tag{25}$$

$$\mathsf{r}_\lambda(\theta) \quad \textit{such that} \quad 1/\mathsf{r}_\lambda(\theta) = \sup_{z \in \mathrm{supp}(\rho)} \left(\sup_{g \in \varphi(z)} \|g\|_{\mathbf{H}_\lambda{}^{-1}(\theta)} \right). \tag{26}$$

B Preliminary results on self concordant losses

In this section, we show how our definition/assumption of self concordance (see Assumption 8) enables a fine control on the excess risk. In particular, we clearly relate the difference in function values to the quadratic approximation at the optimum as well as the renormalized gradient. We start by presenting a general bounds in Appendix B.1 before applying them to the problem of localizing the optimum Appendix B.2.

B.1 Basic results on self-concordance

In this section, as in the rest of the appendix, we are under the conditions of Assumption 8. **In this section only**, we give ourselves a probability measure μ on $\mathcal{Z}$. We will apply the results of this section to $\mu = \rho, \widehat{\rho}, \delta_z$, where $\widehat{\rho} = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$ and z is sampled from ρ.

First of all, let us introduce the following notation. For any probability measure μ on $\mathcal{Z}$ and any $\theta \in \mathcal{H}$, define

- $R^\mu = \sup_{z \in \mathrm{supp}(\mu)} \left(\sup_{g \in \varphi(z)} \|g\| \right)$,
- $\mathsf{t}^\mu(\theta) = \sup_{z \in \mathrm{supp}(\mu)} \left(\sup_{g \in \varphi(z)} |\theta \cdot g| \right)$.

In order to be able to define $L_\mu(\theta) = \mathbb{E}_\mu[\ell_z(\theta)]$ and to derive under the expectation, we assume that Assumptions 3 and 4 are satisfied for μ (replace ρ by μ in the assumption).

Since μ and ℓ satisfy Assumptions 3, 4 and 8, Proposition 8 ensures that we can define $L_\mu(\theta) = \mathbb{E}_\mu[\ell_z(\theta)]$ and $L_{\mu,\lambda}(\theta) = L_\mu(\theta) + \frac{\lambda}{2}\|\theta\|^2$, as well as their respective Hessians $\mathbf{H}_\mu(\theta)$ and $\mathbf{H}_{\mu,\lambda}(\theta)$.

The following result is greatly inspired from results in [2] on generalized self concordant losses, and their refinement in [25]. However, while Eqs. (27), (29) and (30) appear more or less explicitly, Eq. (28) provides an easier way to deal with certain bounds afterwards and was not used in this form before.

Proposition 4 (using the self-concordance of ℓ)**.** *Let $\theta_0, \theta_1 \in \mathcal{H}$ and $\lambda \geqslant 0$. Assume that $(\ell_z)_z$ and μ satisfy Assumptions 3, 4 and 8. We have the following inequalities:*

- *Bounds on Hessians*

$$\mathbf{H}_{\mu,\lambda}(\theta_1) \preceq \exp\left(\mathsf{t}^\mu(\theta_1-\theta_0)\right)\mathbf{H}_{\mu,\lambda}(\theta_0). \tag{27}$$

- *Bounds on gradients (if $\lambda > 0$)*

$$\underline{\phi}\left(\mathsf{t}^\mu(\theta_1-\theta_0)\right)\ \|\theta_1-\theta_0\|_{\mathbf{H}_{\mu,\lambda}(\theta_0)} \leqslant \|\nabla L_{\mu,\lambda}(\theta_1)-\nabla L_{\mu,\lambda}(\theta_0)\|_{\mathbf{H}^{-1}_{\mu,\lambda}(\theta_0)}, \tag{28}$$

$$\|\nabla L_{\mu,\lambda}(\theta_1)-\nabla L_{\mu,\lambda}(\theta_0)\|_{\mathbf{H}^{-1}_{\mu,\lambda}(\theta_0)} \leqslant \overline{\phi}\left(\mathsf{t}^\mu(\theta_1-\theta_0)\right)\ \|\theta_1-\theta_0\|_{\mathbf{H}_{\mu,\lambda}(\theta_0)}, \tag{29}$$

where $\overline{\phi}(t) = (e^t-1)/t$ *and* $\underline{\phi}(t) = (1-e^{-t})/t$.

- *Bounds on function values*

$$L_{\mu,\lambda}(\theta_1) - L_{\mu,\lambda}(\theta_0) - \nabla L_{\mu,\lambda}(\theta_0)(\theta_1-\theta_0) \leqslant \psi\left(\mathsf{t}^\mu(\theta_1-\theta_0)\right)\|\theta_1-\theta_0\|^2_{\mathbf{H}_{\mu,\lambda}(\theta_0)}, \tag{30}$$

where $\psi(t) = (e^t-t-1)/t^2$.

Proof. First of all, note that for any μ and λ, given $\theta \in \mathcal{H}$ and $k, h \in \mathcal{H}$,

$$\begin{aligned}
\left|\nabla^3 L_{\mu,\lambda}(\theta)[h,k,k]\right| &= \left|\mathbb{E}_{z\sim\mu}\left[\nabla^3 \ell_z^\lambda(\theta)[h,k,k]\right]\right| \\
&\leqslant \mathbb{E}_{z\sim\mu}\left[\left|\nabla^3\ell_z(\theta)[h,k,k]\right|\right] \\
&\leqslant \mathbb{E}_{z\sim\mu}\left[\sup_{g\in\varphi(z)} |h\cdot g|\ \nabla^2\ell_z(\theta)[k,k]\right] \\
&\leqslant \mathsf{t}^\mu(h)\,\mathbb{E}_{z\sim\mu}\left[\nabla^2\ell_z(\theta)[k,k]\right] = \mathsf{t}^\mu(h)\ \nabla^2 L_\mu(\theta)[k,k].
\end{aligned}$$

This yields the following fundamental inequality :

$$\left|\nabla^3 L_{\mu,\lambda}(\theta)[h,k,k]\right| \leqslant\ \mathsf{t}^\mu(h)\ \nabla^2 L_{\mu,\lambda}(\theta)[k,k]. \tag{31}$$

We now define, for any $t\in\mathbb{R}$, $\theta_t := \theta_0 + t(\theta_1-\theta_0)$.

Point 1. For the first inequality, let $h\in\mathcal{H}$ be a fixed vector, and consider the function $\varphi : t \in \mathbb{R} \mapsto \nabla^2 L_{\mu,\lambda}(\theta_t)[h,h]$. Since $\varphi'(t) = \nabla^3 L_{\mu,\lambda}(\theta_t)[\theta_1-\theta_0,h,h]$, using Eq. (31), we get that $\varphi'(t) \leqslant \mathsf{t}^\mu(\theta_1-\theta_0)\ \varphi(t)$. Using Lemma 10, we directly find that $\varphi(1) \leqslant \exp(\mathsf{t}^\mu(\theta_1-\theta_0))\varphi(0)$, which, rewriting the definition of φ, yields

$$\nabla^2 L_{\mu,\lambda}(\theta_1)[h,h] \leqslant \exp(\mathsf{t}^\mu(\theta_1-\theta_0))\nabla^2 L_{\mu,\lambda}(\theta_0)[h,h].$$

This being true for any direction h, we have (27).

Point 2. To prove Eq. (28), let us look at the quantity $(\theta_1-\theta_0)\cdot(\nabla L_{\mu,\lambda}(\theta_1)-\nabla L_{\mu,\lambda}(\theta_0))$. Since $\nabla L_{\mu,\lambda}(\theta_1)-\nabla L_{\mu,\lambda}(\theta_0) = \int_0^1 \nabla^2 L_{\mu,\lambda}(\theta_t)(\theta_1-\theta_0)dt$, we have

$$(\theta_1-\theta_0)\cdot(\nabla L_{\mu,\lambda}(\theta_1)-\nabla L_{\mu,\lambda}(\theta_0)) = \int_0^1 \nabla^2 L_{\mu,\lambda}(\theta_t)[\theta_1-\theta_0,\theta_1-\theta_0]dt.$$

Applying Eq. (27) to θ_0 and θ_t and the reverse, we find that

$$\forall t\in[0,1],\ e^{-t\mathsf{t}^\mu(\theta_1-\theta_0)}\nabla^2 L_{\mu,\lambda}(\theta_0) \preceq \nabla^2 L_{\mu,\lambda}(\theta_t).$$

Hence, integrating the previous equation, we have

$$(\theta_1-\theta_0)\cdot(\nabla L_{\mu,\lambda}(\theta_1)-\nabla L_{\mu,\lambda}(\theta_0)) \geqslant \underline{\phi}\left(\mathsf{t}^\mu(\theta_1-\theta_0)\right)\|\theta_1-\theta_0\|^2_{\mathbf{H}_{\mu,\lambda}(\theta_0)}.$$

Finally, bounding $(\theta_1-\theta_0)\cdot(\nabla L_{\mu,\lambda}(\theta_1)-\nabla L_{\mu,\lambda}(\theta_0))$ by $\|\theta_1-\theta_0\|_{\mathbf{H}_{\mu,\lambda}(\theta_0)}\ \|\nabla L_{\mu,\lambda}(\theta_1)-\nabla L_{\mu,\lambda}(\theta_0)\|_{\mathbf{H}^{-1}_{\mu,\lambda}(\theta_0)}$, and simplifying by $\|\theta_1-\theta_0\|_{\mathbf{H}_{\mu,\lambda}(\theta_0)}$, we obtain Eq. (28).

Point 3. To prove Eq. (29), first write

$$\begin{aligned}
\|\nabla L_{\mu,\lambda}(\theta_1) - \nabla L_{\mu,\lambda}(\theta_0)\|_{\mathbf{H}^{-1}_{\mu,\lambda}(\theta_0)} &= \|\int_0^1 \mathbf{H}^{-1/2}_{\mu,\lambda}(\theta_0)\mathbf{H}_{\mu,\lambda}(\theta_t)(\theta_1 - \theta_0)dt\| \\
&= \|\int_0^1 \mathbf{H}^{-1/2}_{\mu,\lambda}(\theta_0)\mathbf{H}_{\mu,\lambda}(\theta_t)\mathbf{H}^{-1/2}_{\mu,\lambda}(\theta_0)\ \mathbf{H}^{1/2}_{\mu,\lambda}(\theta_0)(\theta_1 - \theta_0)dt\| \\
&\leqslant \left(\int_0^1 \|\mathbf{H}^{-1/2}_{\mu,\lambda}(\theta_0)\mathbf{H}_{\mu,\lambda}(\theta_t)\mathbf{H}^{-1/2}_{\mu,\lambda}(\theta_0)\|\ dt\right)\ \|\theta_1 - \theta_0\|_{\mathbf{H}_{\mu,\lambda}(\theta_0)}.
\end{aligned}$$

Then apply Eq. (27) to have

$$\forall t \in [0,1],\ \mathbf{H}_{\mu,\lambda}(\theta_t) \preceq e^{t\mathsf{t}^\mu(\theta_1-\theta_0)}\mathbf{H}_{\mu,\lambda}(\theta_0).$$

This implies

$$\forall t \in [0,1],\ \mathbf{H}^{-1/2}_{\mu,\lambda}(\theta_0)\mathbf{H}_{\mu,\lambda}(\theta_t)\mathbf{H}^{-1/2}_{\mu,\lambda}(\theta_0) \preceq e^{t\mathsf{t}^\mu(\theta_1-\theta_0)}\ I.$$

And hence in particular

$$\forall t \in [0,1],\ \|\mathbf{H}^{-1/2}_{\mu,\lambda}(\theta_0)\mathbf{H}_{\mu,\lambda}(\theta_t)\mathbf{H}^{-1/2}_{\mu,\lambda}(\theta_0)\| \leqslant e^{t\mathsf{t}^\mu(\theta_1-\theta_0)}.$$

Finally, integrating this, we get

$$\int_0^1 \|\mathbf{H}^{-1/2}_{\mu,\lambda}(\theta_0)\mathbf{H}_{\mu,\lambda}(\theta_t)\mathbf{H}^{-1/2}_{\mu,\lambda}(\theta_0)\|\ dt \leqslant \overline{\phi}\left(\mathsf{t}^\mu(\theta_1-\theta_0)\right).$$

Thus Eq. (29) is proved.

Point 4. To prove Eq. (30), define $\forall t \in \mathbb{R},\ \varphi(t) = L_{\mu,\lambda}(\theta_t) - L_{\mu,\theta}(\theta_0) - t\,\nabla L_{\mu,\lambda}(\theta_0)(\theta_1 - \theta_0)$ We have $\varphi''(t) = \|\theta_1 - \theta_0\|^2_{\mathbf{H}_{\mu,\lambda}(\theta_t)} \leqslant e^{t\ \mathsf{t}^\mu(\theta_1-\theta_0)}\varphi''(0)$. Then using the fact that $\varphi(0), \varphi'(0) = 0$ and integrating this inequality two times, we get the result.

Proof.of Proposition 1. First note that since the support of μ is included in the support of ρ, Assumption 3 and Assumption 4 also hold for μ. Hence, since Assumptions 2 to 4 are satisfied, by Proposition 8, $L_{\mu,\lambda}$, $\nabla L_{\mu,\lambda}$ and $\nabla^2 L_{\mu,\lambda}$ are well-defined.

Assuming the existence of a minimizer $\theta^\star_{\mu,\lambda}$ of $L_{\mu,\lambda}$, the reported equations are the same than those of Proposition 4 when taking $\theta_1 = \theta$ and $\theta_0 = \theta^\star_{\mu,\lambda}$, with the fact that $\mathsf{t}^\mu(v) \leqslant \mathsf{t}(v)$ for any $v \in \mathcal{H}$ since the support of μ is a subset of the support of ρ, and $\nabla L_{\mu,\lambda}(\theta^\star_{\mu,\lambda}) = 0$. Note that since $L_{\mu,\lambda}$ is defined on $\mathcal{H}$, if $\lambda > 0$, then $\theta^\star_{\mu,\lambda}$ always exists and is unique by strong convexity.

B.2 Localization properties for t_λ

The aim of this section is to localize the optima $\theta^\star_\lambda$ and $\widehat{\theta}^\star_\lambda$ using the re-normalized gradient. This type of result is inspired by Proposition 2 of [2] or Proposition 3.5 of [25]. However, their proof is based on a slightly different result, namely Eq. (28), and its formulation is slightly different. Indeed, while the two propositions mentioned above concentrate on performing a quadratic approximation directly, we bound the term that could have been too large in that quadratic approximation.

Proposition 5 (localisation). *Let $\theta \in \mathcal{H}$, then the following holds*

$$\|\nabla L_\lambda(\theta)\|_{\mathbf{H}^{-1}_\lambda(\theta)} \leqslant \frac{\mathsf{r}_\lambda(\theta)}{2} \quad\implies\quad \mathsf{t}(\theta - \theta^\star_\lambda) = \mathsf{t}_\lambda \leqslant \log 2, \tag{32}$$

$$\|\nabla \widehat{L}_\lambda(\theta)\|_{\mathbf{H}_\lambda{}^{-1}(\theta)}\ \|\widehat{\mathbf{H}}^{-1/2}_\lambda(\theta)\mathbf{H}_\lambda{}^{1/2}(\theta)\|^2 \leqslant \frac{\mathsf{r}_\lambda(\theta)}{2} \quad\implies \mathsf{t}(\theta - \widehat{\theta}^\star_\lambda) \leqslant \log 2. \tag{33}$$

Proof. To prove Eq. (32), we first write

$$\mathsf{t}(\theta-\theta^\star_\lambda) = \sup_{z\in\mathrm{supp}(\rho)}\sup_{g\in\varphi(z)} |(\theta-\theta^\star_\lambda)\cdot g| \leqslant \|\theta-\theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta)} \sup_{z\in\mathrm{supp}(\rho)}\sup_{g\in\varphi(z)} \|g\|_{\mathbf{H}_\lambda^{-1}(\theta)}.$$

Now we use Eq. (14) to bound $\|\theta-\theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta)}$, and putting things together, we get

$$\mathsf{t}(\theta-\theta^\star_\lambda)\underline{\phi}\left(\mathsf{t}(\theta-\theta^\star_\lambda)\right) \leqslant \frac{\|\nabla L_\lambda(\theta)\|_{\mathbf{H}_\lambda^{-1}(\theta)}}{\mathsf{r}_\lambda(\theta)}.$$

Using the fact that $t\underline{\phi}(t) = 1-e^{-t}$ is an increasing function, we see that if $t\underline{\phi}(t) \leqslant 1/2$, then $t\leqslant \log 2$ hence the result.

To prove Eq. (33), we use the same reasoning. First, we bound

$$\mathsf{t}(\theta-\widehat{\theta}^\star_\lambda) = \sup_{z\in\mathrm{supp}(\rho)}\sup_{g\in\varphi(z)} \left|(\theta-\widehat{\theta}^\star_\lambda)\cdot g\right| \leqslant \|\theta-\widehat{\theta}^\star_\lambda\|_{\widehat{\mathbf{H}}_\lambda(\theta)}\, \|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta)\mathbf{H}_\lambda{}^{1/2}(\theta)\| \sup_{z\in\mathrm{supp}(\rho)}\sup_{g\in\varphi(z)} \|g\|_{\mathbf{H}_\lambda^{-1}(\theta)}.$$

Now using Eq. (14) to the function $\widehat{L}_\lambda$, we get

$$\mathsf{t}(\theta-\widehat{\theta}^\star_\lambda)\underline{\phi}\left(\mathsf{t}^{\hat\rho}(\theta-\widehat{\theta}^\star_\lambda)\right) \leqslant \|\nabla\widehat{L}_\lambda(\theta)\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta)}\, \|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta)\mathbf{H}_\lambda{}^{1/2}(\theta)\|\, \frac{1}{\mathsf{r}_\lambda(\theta)}.$$

Now using the fact that $\mathsf{t}^{\hat\rho}(\theta-\widehat{\theta}^\star_\lambda) \leqslant \mathsf{t}(\theta-\widehat{\theta}^\star_\lambda)$ and that $\underline{\phi}$ is a decreasing function, and that $\|\nabla\widehat{L}_\lambda(\theta)\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta)} \leqslant \|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta)\mathbf{H}_\lambda{}^{1/2}(\theta)\|\, \|\nabla\widehat{L}_\lambda(\theta)\|_{\mathbf{H}_\lambda{}^{-1}(\theta)}$, this yields

$$\mathsf{t}(\theta-\widehat{\theta}^\star_\lambda)\underline{\phi}\left(\mathsf{t}(\theta-\widehat{\theta}^\star_\lambda)\right) \leqslant \|\nabla\widehat{L}_\lambda(\theta)\|_{\mathbf{H}_\lambda{}^{-1}(\theta)}\, \|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta)\mathbf{H}_\lambda{}^{1/2}(\theta)\|^2\, \frac{1}{\mathsf{r}_\lambda(\theta)}.$$

We conclude using the same argument as before.

C Main result, simplified

In this section, we perform a simplified analysis in the case where we assume nothing on Bias_λ more than just the fact that $\theta^\star$ exists. In this section we assume that ℓ_z and ρ satisfy Assumptions 3 to 5 and 8.

Definition 4 (Definition of $\overline{\mathsf{B}}_1$, $\overline{\mathsf{B}}_2$ and $\overline{\mathsf{df}}_\lambda$)**.** *Under assumptions Assumptions 3 to 5 and 8, the following quantities are well-defined and real-valued.*

$$\overline{\mathsf{B}}_1 = \sup_{\|\theta\|\leqslant\|\theta^\star\|} \mathsf{B}_1(\theta) \quad \overline{\mathsf{B}}_2 = \sup_{\|\theta\|\leqslant\|\theta^\star\|} \mathsf{B}_2(\theta), \quad \overline{\mathsf{df}}_\lambda = \mathbb{E}\left[\|\nabla\ell_z(\theta^\star_\lambda)\|^2_{\mathbf{H}_\lambda{}^{-1}(\theta^\star_\lambda)}\right].$$

Proposition 6. *The quantities in Definition 4 are finite and moreover*

$$\overline{\mathsf{df}}_\lambda \leqslant \frac{\overline{\mathsf{B}}_1^2}{\lambda}.$$

Proof. These are well defined thanks to Lemmas 11 and 12.

Definition 5 (Constants)**.** *In this section, we will use the following constants.*

$$K_{\text{var}} = \frac{1+\psi(\log 2)}{\underline{\phi}(\log 2)^2} \leqslant 4, \quad \triangle = 2\sqrt{2}\left(1+\frac{1}{2\sqrt{3}}\right) \leqslant 4,$$

$$\mathsf{C}_{\text{bias}} = 1+\frac{K_{\text{var}}}{8} \leqslant 2, \qquad \mathsf{C}_{\text{var}} = 2K_{\text{var}}\triangle^2 \leqslant 84.$$

C.1 Analytic results

Theorem 5 (Analytic decomposition)**.** *For any* $\lambda > 0$ *and* $n \in \mathbb{N}$, *if* $\frac{R}{\sqrt{\lambda}}\widehat{\mathsf{Var}}_\lambda \leqslant \frac{1}{2}$,

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leqslant K_{\text{var}}\, \widehat{\mathsf{Var}}_\lambda^2 + \lambda\|\theta^\star\|^2, \tag{34}$$

where K_{var} *is defined in Definition 5.*

Proof.
First decompose the excess risk of $\widehat{\theta}^\star_\lambda$ in the following way:

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) = \underbrace{L_\lambda(\widehat{\theta}^\star_\lambda) - L_\lambda(\theta^\star_\lambda)}_{\text{variance}} + \underbrace{L(\theta^\star_\lambda) - L(\theta^\star)}_{\text{bias}} + \underbrace{\frac{\lambda}{2}\left(\|\theta^\star_\lambda\|^2 - \|\widehat{\theta}^\star_\lambda\|^2\right)}_{\text{mixed}}.$$

1) Variance term: For the variance term, use Eq. (13)

$$L_\lambda(\widehat{\theta}^\star_\lambda) - L_\lambda(\theta^\star_\lambda) \leqslant \psi\left(\mathsf{t}(\theta^\star_\lambda - \widehat{\theta}^\star_\lambda)\right) \|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|^2_{\mathbf{H}_\lambda(\theta^\star_\lambda)}.$$

2) Bias term: For the bias term, note that since $\|\theta^\star_\lambda\| \leqslant \|\theta^\star\|$,

$$L(\theta^\star_\lambda) - L(\theta^\star) = L_\lambda(\theta^\star_\lambda) - L_\lambda(\theta^\star) + \frac{\lambda}{2}\|\theta^\star\|^2 - \frac{\lambda}{2}\|\theta^\star_\lambda\|^2 \leqslant \frac{\lambda}{2}\|\theta^\star\|^2.$$

3) Mixed term: For the mixed term, since $\|\theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)^{-1}} \leqslant \|\mathbf{H}_\lambda(\theta^\star_\lambda)^{-1/2}\|\|\theta^\star_\lambda\| \leqslant \lambda^{-1/2}\|\theta^\star_\lambda\| \leqslant \lambda^{-1/2}\|\theta^\star\|$, we have

$$\begin{aligned}
\frac{\lambda}{2}\left(\|\theta^\star_\lambda\|^2 - \|\widehat{\theta}^\star_\lambda\|^2\right) &= \frac{\lambda}{2}\left(\theta^\star_\lambda - \widehat{\theta}^\star_\lambda\right)\cdot\left(\theta^\star_\lambda + \widehat{\theta}^\star_\lambda\right) \\
&\leqslant \frac{\lambda}{2}\|\theta^\star_\lambda - \widehat{\theta}^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)}\ \left(\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)^{-1}} + 2\|\theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)^{-1}}\right) \\
&\leqslant \frac{1}{2}\|\theta^\star_\lambda - \widehat{\theta}^\star_\lambda\|^2_{\mathbf{H}_\lambda(\theta^\star_\lambda)} + \sqrt{\lambda}\|\theta^\star\|\ \|\theta^\star_\lambda - \widehat{\theta}^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)} \\
&\leqslant \|\theta^\star_\lambda - \widehat{\theta}^\star_\lambda\|^2_{\mathbf{H}_\lambda(\theta^\star_\lambda)} + \frac{\lambda}{2}\|\theta^\star\|^2.
\end{aligned}$$

where we get the last inequality by using $ab \leqslant \frac{a^2}{2} + \frac{b^2}{2}$.
4) Putting things together

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leqslant \left(1 + \psi\left(\mathsf{t}(\theta^\star_\lambda - \widehat{\theta}^\star_\lambda)\right)\right)\|\theta^\star_\lambda - \widehat{\theta}^\star_\lambda\|^2_{\mathbf{H}_\lambda(\theta^\star_\lambda)} + \lambda\|\theta^\star\|^2.$$

By using Eq. (14) we have

$$\begin{aligned}
\|\theta^\star_\lambda - \widehat{\theta}^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)} &\leqslant \|\mathbf{H}_\lambda{}^{1/2}(\theta^\star_\lambda)\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta^\star_\lambda)\|\ \|\theta^\star_\lambda - \widehat{\theta}^\star_\lambda\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star_\lambda)} \\
&\leqslant \frac{1}{\underline{\phi}\left(\mathsf{t}^{\widehat{\rho}}(\theta^\star_\lambda - \widehat{\theta}^\star_\lambda)\right)}\ \|\mathbf{H}_\lambda{}^{1/2}(\theta^\star_\lambda)\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta^\star_\lambda)\|\ \|\nabla\widehat{L}_\lambda(\theta^\star_\lambda)\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star_\lambda)}.
\end{aligned}$$

Note that by multiplying and dividing for $\mathbf{H}_\lambda{}^{1/2}(\theta^\star_\lambda)$,

$$\begin{aligned}
\|\nabla\widehat{L}_\lambda(\theta^\star_\lambda)\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star_\lambda)} &= \|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta^\star_\lambda)\nabla\widehat{L}_\lambda(\theta^\star_\lambda)\| = \|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta^\star_\lambda)\mathbf{H}_\lambda{}^{-1/2}(\theta^\star_\lambda)\mathbf{H}_\lambda{}^{1/2}(\theta^\star_\lambda)\nabla\widehat{L}_\lambda(\theta^\star_\lambda)\| \\
&\leqslant \|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta^\star_\lambda)\mathbf{H}_\lambda{}^{-1/2}(\theta^\star_\lambda)\|\|\mathbf{H}_\lambda{}^{1/2}(\theta^\star_\lambda)\nabla\widehat{L}_\lambda(\theta^\star_\lambda)\| \\
&= \|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta^\star_\lambda)\mathbf{H}_\lambda{}^{-1/2}(\theta^\star_\lambda)\|\|\nabla\widehat{L}_\lambda(\theta^\star_\lambda)\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star_\lambda)}.
\end{aligned}$$

Then,

$$\|\theta_\lambda^\star - \widehat{\theta}_\lambda^\star\|_{\mathbf{H}_\lambda(\theta_\lambda^\star)} \leqslant \frac{1}{\underline{\phi}\left(\mathsf{t}(\theta_\lambda^\star - \widehat{\theta}_\lambda^\star)\right)} \|\mathbf{H}_\lambda{}^{1/2}(\theta_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta_\lambda^\star)\|^2 \, \|\nabla \widehat{L}_\lambda(\theta_\lambda^\star)\|_{\mathbf{H}_\lambda{}^{-1}(\theta_\lambda^\star)}$$
$$= \frac{1}{\underline{\phi}\left(\mathsf{t}(\theta_\lambda^\star - \widehat{\theta}_\lambda^\star)\right)} \widehat{\mathsf{Var}}_\lambda.$$

Now we know that using Eq. (33), if $\widehat{\mathsf{Var}}_\lambda \leqslant \frac{\mathsf{r}_\lambda(\theta_\lambda^\star)}{2}$, then $\mathsf{t}(\theta_\lambda^\star - \widehat{\theta}_\lambda^\star) \leqslant \log 2$, which yields the following bound:

$$L(\widehat{\theta}_\lambda^\star) - L(\theta^\star) \leqslant \frac{(1+\psi(\log 2))}{\underline{\phi}(\log 2)^2}\widehat{\mathsf{Var}}_\lambda + \lambda\|\theta^\star\|^2.$$

Finally, we can bound $\frac{1}{\mathsf{r}_\lambda(\theta_\lambda^\star)} \leqslant \frac{R}{\lambda^{1/2}}$ to have the final form of the proposition.

C.2 Probabilistic results

Lemma 2 (bounding $\|\nabla \widehat{L}_\lambda(\theta_\lambda^\star)\|_{\mathbf{H}_\lambda(\theta_\lambda^\star)}$). *Let $n \in \mathbb{N}$, $\lambda > 0$ and $\delta \in (0,1]$. For $k \geqslant 1$, if*

$$n \geqslant 24\frac{\overline{\mathsf{B}}_2}{\lambda}\log\frac{2}{\delta}, n \geqslant k^2 2\log\frac{2}{\delta},$$

then with probability at least $1-\delta$,

$$\|\mathbf{H}_\lambda(\theta_\lambda^\star)^{-1/2}\nabla \widehat{L}_\lambda(\theta_\lambda^\star)\| \leqslant \triangle/2\sqrt{\frac{\overline{\mathsf{df}}_\lambda \vee (\overline{\mathsf{B}}_1^2/\overline{\mathsf{B}}_2)\,\log\frac{2}{\delta}}{n}} + \frac{2}{k}\sqrt{\lambda}\|\theta^\star\|$$

where $\triangle$ is defined in Definition 5.

Proof.1) First use Bernstein inequality for random vectors [e.g. Thm. 3.3.4 of 42]: for any $n \in \mathbb{N}$ and $\delta \in (0,1]$, with probability at least $1-\delta$, we have

$$\|\mathbf{H}_\lambda(\theta_\lambda^\star)^{-1/2}\nabla \widehat{L}_\lambda(\theta_\lambda^\star)\| \leqslant \frac{2M\log\frac{2}{\delta}}{n} + \sigma\sqrt{\frac{2\log\frac{2}{\delta}}{n}},$$

where $M = \sup_{z\in\mathrm{supp}(\rho)} \|\nabla \ell_z^\lambda(\theta_\lambda^\star)\|_{\mathbf{H}_\lambda{}^{-1}(\theta_\lambda^\star)}$ and $\sigma = \mathbb{E}\left[\|\nabla \ell_z^\lambda(\theta_\lambda^\star)\|^2_{\mathbf{H}_\lambda{}^{-1}(\theta_\lambda^\star)}\right]^{1/2}$.

2) Using the fact that $\nabla \ell_z^\lambda(\theta_\lambda^\star) = \nabla \ell_z(\theta_\lambda^\star) + \lambda\theta_\lambda^\star$, we bound M as follows:

$$M = \sup_{z\in\mathrm{supp}(\rho)} \|\nabla \ell_z^\lambda(\theta_\lambda^\star)\|_{\mathbf{H}_\lambda(\theta_\lambda^\star)} \leqslant \sup_{z\in\mathrm{supp}(\rho)} \|\nabla \ell_z(\theta_\lambda^\star)\|_{\mathbf{H}_\lambda(\theta_\lambda^\star)} + \lambda\|\theta_\lambda^\star\|_{\mathbf{H}_\lambda{}^{-1}(\theta_\lambda^\star)} \leqslant \frac{\overline{\mathsf{B}}_1}{\sqrt{\lambda}} + \sqrt{\lambda}\|\theta^\star\|,$$

where in the last inequality, we use the fact that $\|\theta_\lambda^\star\|_{\mathbf{H}_\lambda{}^{-1}(\theta_\lambda^\star)} \leqslant \frac{1}{\sqrt{\lambda}}\|\theta_\lambda^\star\| \leqslant \frac{1}{\sqrt{\lambda}}\|\theta^\star\|$. Similarly, we bound σ

$$\sigma \leqslant \mathbb{E}\left[\|\nabla \ell_z(\theta_\lambda^\star)\|^2_{\mathbf{H}_\lambda{}^{-1}(\theta_\lambda^\star)}\right]^{1/2} + \lambda\|\theta_\lambda^\star\|_{\mathbf{H}_\lambda{}^{-1}(\theta_\lambda^\star)} \leqslant \overline{\mathsf{df}}_\lambda^{1/2} + \sqrt{\lambda}\|\theta^\star\|.$$

3) Injecting these bounds in the concentration inequality,

$$\|\mathbf{H}_\lambda(\theta_\lambda^\star)^{-1/2}\nabla \widehat{L}_\lambda(\theta_\lambda^\star)\| \leqslant \sqrt{\frac{2\overline{\mathsf{B}}_2\,\log\frac{2}{\delta}}{\lambda n}}\sqrt{\frac{2(\overline{\mathsf{B}}_1^2/\overline{\mathsf{B}}_2)\,\log\frac{2}{\delta}}{n}} + \sqrt{\frac{2\overline{\mathsf{df}}_\lambda\,\log\frac{2}{\delta}}{n}}$$
$$+ \sqrt{\lambda}\|\theta^\star\|\left(\frac{2\log\frac{2}{\delta}}{n} + \sqrt{\frac{2\log\frac{2}{\delta}}{n}}\right),$$

where we have decomposed $\frac{2\overline{\mathsf{B}}_1^2 \log\frac{2}{\delta}}{\sqrt{\lambda}n} = \sqrt{\frac{2\overline{\mathsf{B}}_2\ \log\frac{2}{\delta}}{\lambda n}}\sqrt{\frac{2(\overline{\mathsf{B}}_1^2/\overline{\mathsf{B}}_2)\ \log\frac{2}{\delta}}{n}}$ for the first term. Reordering the terms, this yields

$$\|\mathbf{H}_\lambda(\theta_\lambda^\star)^{-1/2}\nabla\widehat{L}_\lambda(\theta_\lambda^\star)\| \leqslant \left(1+\sqrt{\frac{2\overline{\mathsf{B}}_2\ \log\frac{2}{\delta}}{\lambda n}}\right)\sqrt{\frac{2\overline{\mathsf{df}}_\lambda \vee \left(\overline{\mathsf{B}}_1^2/\overline{\mathsf{B}}_2\right)\ \log\frac{2}{\delta}}{n}}$$
$$+\sqrt{\lambda}\|\theta^\star\|\left(\frac{2\log\frac{2}{\delta}}{n}+\sqrt{\frac{2\log\frac{2}{\delta}}{n}}\right).$$

4) Now assuming that

$$n \geqslant 24\frac{\overline{\mathsf{B}}_2}{\lambda}\log\frac{2}{\delta}, n \geqslant k^2 2\log\frac{2}{\delta},$$

this yields

$$\|\mathbf{H}_\lambda(\theta_\lambda^\star)^{-1/2}\nabla\widehat{L}_\lambda(\theta_\lambda^\star)\| \leqslant \left(1+\frac{1}{2\sqrt{3}}\right)\sqrt{\frac{2\overline{\mathsf{df}}_\lambda \vee (\overline{\mathsf{B}}_1^2/\overline{\mathsf{B}}_2)\ \log\frac{2}{\delta}}{n}}+\frac{2}{k}\sqrt{\lambda}\|\theta^\star\|.$$

Combining the two previous lemmas, we get:

Lemma 3 (Bounding $\widehat{\mathsf{Var}}_\lambda$)**.** *Let* $n \in \mathbb{N}$ *and* $0 < \lambda \leqslant \overline{\mathsf{B}}_2$. *Let* $\delta \in (0,1]$. *If for* $k \geqslant 1$

$$n \geqslant 24\frac{\overline{\mathsf{B}}_2}{\lambda}\log\frac{8\overline{\mathsf{B}}_2}{\lambda\delta}, \quad n \geqslant 2k^2\log\frac{2}{\delta},$$

then with probability at least $1-2\delta$*,*

$$\widehat{\mathsf{Var}}_\lambda \leqslant \triangle\sqrt{\frac{\overline{\mathsf{df}}_\lambda \vee (\overline{\mathsf{B}}_1^2/\overline{\mathsf{B}}_2)\ \log\frac{2}{\delta}}{n}}+\frac{4}{k}\sqrt{\lambda}\|\theta^\star\|,$$

where $\triangle$ *is a constant defined in Definition 5.*

Proof. Recall that $\widehat{\mathsf{Var}}_\lambda = \|\mathbf{H}_\lambda{}^{1/2}(\theta_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta_\lambda^\star)\|^2\ \|\nabla\widehat{L}_\lambda(\theta_\lambda^\star)\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta_\lambda^\star)}$. Using Lemma 6, under the conditions of this lemma, we have $\|\mathbf{H}_\lambda{}^{1/2}(\theta_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta_\lambda^\star)\|^2 \leqslant 2$. Combining this with the bound for $\|\nabla\widehat{L}_\lambda(\theta_\lambda^\star)\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta_\lambda^\star)}$ obtained in Lemma 2, we get the result (the probability $1-2\delta$ comes from the fact that we perform a union bound).

C.3 Final result

Theorem 6 (General bound, simplified setting)**.** *Let* $n \in \mathbb{N}$ *and* $0 < \lambda \leqslant \overline{\mathsf{B}}_2$. *Let* $\delta \in (0,1]$. *If*

$$n \geqslant 512\left(\|\theta^\star\|^2R^2 \vee 1\right)\log\frac{2}{\delta}, \quad n \geqslant 24\frac{\overline{\mathsf{B}}_2}{\lambda}\log\frac{8\overline{\mathsf{B}}_2}{\lambda\delta}, \quad n \geqslant 16\triangle^2R^2\frac{\overline{\mathsf{df}}_\lambda \vee (\overline{\mathsf{B}}_1^2/\overline{\mathsf{B}}_2)}{\lambda}\ \log\frac{2}{\delta},$$

then with probability at least $1-2\delta$*,*

$$L(\widehat{\theta}_\lambda^\star) - L(\theta^\star) \leqslant \mathsf{C}_{\text{var}}\frac{\overline{\mathsf{df}}_\lambda \vee (\overline{\mathsf{B}}_1^2/\overline{\mathsf{B}}_2)}{n}\ \log\frac{2}{\delta}+\mathsf{C}_{\text{bias}}\lambda\|\theta^\star\|^2,$$

where $\triangle, \mathsf{C}_{\text{bias}}, \mathsf{C}_{\text{var}}$ *are defined in Definition 5.*

Proof.1) Recall the analytical decomposition in Thm. 5. For any $\lambda > 0$ and $n \in \mathbb{N}$, if $\frac{R}{\sqrt{\lambda}}\widehat{\mathsf{Var}}_\lambda \leqslant \frac{1}{2}$,

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leqslant K_{\text{var}}\, \widehat{\mathsf{Var}}^2_\lambda + \lambda\|\theta^\star\|^2,$$

where K_{var} is defined in Definition 5.
2) Now apply Lemma 3 for a given $k \geqslant 1$. If

$$n \geqslant 24\frac{\overline{\mathsf{B}}_2}{\lambda}\log\frac{8\overline{\mathsf{B}}_2}{\lambda\delta}, \quad n \geqslant 2k^2\log\frac{2}{\delta},$$

then with probability at least $1 - 2\delta$,

$$\widehat{\mathsf{Var}}_\lambda \leqslant \triangle\sqrt{\frac{\overline{\mathsf{df}}_\lambda \vee (\overline{\mathsf{B}}^2_1/\overline{\mathsf{B}}_2)\ \log\frac{2}{\delta}}{n}} + \frac{4}{k}\sqrt{\lambda}\|\theta^\star\|,$$

where $\triangle$ is a constant defined in Definition 5.

In order to satisfy the condition to have the analytical decomposition, namely $\frac{R}{\sqrt{\lambda}}\widehat{\mathsf{Var}}_\lambda \leqslant \frac{1}{2}$, it is therefore sufficient to have

$$\triangle\, R\sqrt{\frac{\overline{\mathsf{df}}_\lambda \vee (\overline{\mathsf{B}}^2_1/\overline{\mathsf{B}}_2)\ \log\frac{2}{\delta}}{\lambda n}} \leqslant \frac{1}{4}, \qquad \frac{4}{k}R\|\theta^\star\| \leqslant \frac{1}{4}.$$

3) Thus, if we choose $k = 16(R\|\theta^\star\| \vee 1)$, we have both $k \geqslant 1$ and the second condition in the previous equation. Moreover, the condition $n \geqslant 2k^2\log\frac{2}{\delta}$ becomes $n \geqslant 512(R^2\|\theta^\star\|^2 \vee 1)\log\frac{2}{\delta}$. Hence, under the conditions of this theorem, we can apply the analytical decomposition :

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leqslant K_{\text{var}}\, \widehat{\mathsf{Var}}^2_\lambda + \lambda\|\theta^\star\|^2 \leqslant 2K_{\text{var}}\triangle^2\,\frac{\overline{\mathsf{df}}_\lambda \vee (\overline{\mathsf{B}}^2_1/\overline{\mathsf{B}}_2)\ \log\frac{2}{\delta}}{n} + \left(1 + K_{\text{var}}\frac{32}{k^2}\right)\lambda\|\theta^\star\|^2.$$

In the last inequality, we have used $(a+b)^2 \leqslant 2a^2 + 2b^2$ to separate the terms coming from $\widehat{\mathsf{Var}}^2_\lambda$.

Finally, using the fact that $k \geqslant 16$ and hence that $\frac{32}{k^2} \leqslant \frac{1}{8}$, we get the constants in the theorem.

Proof.of Thm. 2 Since $\forall\lambda > 0,\ \overline{\mathsf{df}}_\lambda \leqslant \frac{\overline{\mathsf{B}}^2_1}{\lambda}$, and since $\lambda \leqslant \overline{\mathsf{B}}_2$, $\overline{\mathsf{df}}_\lambda \vee \overline{\mathsf{B}}^2_1/\overline{\mathsf{B}}_2 \leqslant \frac{\overline{\mathsf{B}}^2_1}{\lambda}$. From Definition 5, we get that $\triangle \leqslant 4$, $\mathsf{C}_{\text{bias}} \leqslant 2$, $\mathsf{C}_{\text{var}} \leqslant 84$. Thus, we can use these bounds in Thm. 6 to obtain the result.

D Main result, refined analysis

In subsection Appendix D.1 we split the excess risk in terms of bias and variance, that will be controlled in Appendix D.3, the final result is Thm. 4 in Appendix D.4, while in Appendix F a version with explicit dependence in λ, n is reported.

Constants First, we introduce three constants that will be crucial for the final bound.

Definition 6.

$$\mathsf{B}^\star_1 = \mathsf{B}_1(\theta^\star), \qquad \mathsf{B}^\star_2 = \mathsf{B}_2(\theta^\star), \qquad \mathsf{Q}^* = \mathsf{B}^\star_1/\sqrt{\mathsf{B}^\star_2}.$$

In the following sections, we also will use the following functions of t_λ and $\widetilde{\mathsf{t}}_\lambda$ which we will treat as constants (see Proposition 7).

Definition 7.

$$K_{\text{bias}}(\mathsf{t}_\lambda) = 2\frac{\psi(\mathsf{t}_\lambda + \log 2)}{\underline{\phi}(\mathsf{t}_\lambda)^2} \leqslant 2e^{3\mathsf{t}_\lambda}, \qquad K_{\text{var}}(\mathsf{t}_\lambda) = 2\frac{\psi(\mathsf{t}_\lambda + \log 2)e^{\mathsf{t}_\lambda}}{\underline{\phi}(\log 2)^2} \leqslant 8e^{2\mathsf{t}_\lambda},$$

$$\square_1(\mathsf{t}_\lambda) = e^{\mathsf{t}_\lambda/2}, \qquad \square_2(\mathsf{t}_\lambda) = e^{\mathsf{t}_\lambda/2}\left(1 + e^{\mathsf{t}_\lambda}\right) \leqslant 2e^{3\mathsf{t}_\lambda/2}$$

$$\mathsf{C}_{\text{bias}} = \psi(\mathsf{t}_\lambda + \log 2)\left(\frac{2}{\underline{\phi}(\mathsf{t}_\lambda)} + \frac{e^{\mathsf{t}_\lambda}}{\underline{\phi}(\log 2)^2}\right) \leqslant 6e^{2\mathsf{t}_\lambda}, \qquad \mathsf{C}_{\text{var}} = \frac{64\psi(\mathsf{t}_\lambda + \log 2)e^{2\mathsf{t}_\lambda}}{\underline{\phi}(\log 2)^2} \leqslant 256e^{3\mathsf{t}_\lambda}$$

$$\triangle_1 = 576\square_1^2\square_2^2(1/2 \vee \widetilde{\mathsf{t}}_\lambda)^2 \leqslant 2304e^{4\mathsf{t}_\lambda}(\widetilde{\mathsf{t}}_\lambda \vee 1/2)^2, \qquad \triangle_2 = 256\square_1^4 \leqslant 256e^{2\mathsf{t}_\lambda}.$$

Note that theses functions are all increasing in t_λ and $\widetilde{\mathsf{t}}_\lambda$, and are lower bounded by strictly positive constants.

For the second bounds, we use the fact that $\psi(t) \leqslant \frac{e^t}{2}$ and $1/\underline{\phi}(t) \leqslant e^t$ to bound all the quantities using only exponentials of t_λ.

A priori, these constants will depend on λ. However, we can always bound t_λ and $\widetilde{\mathsf{t}}_\lambda$ in the following way.

Lemma 4. *Recall the definitions of* $\mathsf{t}_\lambda := \mathsf{t}(\theta^\star_\lambda - \theta^\star)$ *and* $\widetilde{\mathsf{t}}_\lambda := \frac{\mathsf{Bias}_\lambda}{\mathsf{r}_\lambda(\theta^\star)}$. *We have the following cases.*

- *If* $\widetilde{\mathsf{t}}_\lambda \leqslant \frac{1}{2}$, *then* $\mathsf{t}_\lambda \leqslant \log 2$,
- *else,* $\widetilde{\mathsf{t}}_\lambda \leqslant R\|\theta^\star\|$ *and* $\mathsf{t}_\lambda \leqslant 2R\|\theta^\star\|$.

Proof. The first point is a direct application of Eq. (32). One can obtain the second by noting that $\mathsf{t}(\theta^\star_\lambda - \theta^\star) \leqslant R\|\theta^\star_\lambda - \theta^\star\|$. Since $\|\theta^\star_\lambda\| \leqslant \|\theta^\star\|$, we have the bound on t_λ. For the bound on $\widetilde{\mathsf{t}}_\lambda$, since $\mathsf{Bias}_\lambda \leqslant \sqrt{\lambda}\|\theta^\star\|$ and $\frac{1}{\mathsf{r}_\lambda(\theta^\star)} \leqslant \frac{R}{\sqrt{\lambda}}$, we have the wanted bound.

Hence, we can always bound the constants in Definition 7 by constants independant of λ.

Proposition 7. *If* $\widetilde{\mathsf{t}}_\lambda \leqslant 1/2$, *then* $\mathsf{t}_\lambda \leqslant \log 2$ *and*

$$K_{\text{bias}}(\mathsf{t}_\lambda) \leqslant 4, \quad K_{\text{var}}(\mathsf{t}_\lambda) \leqslant 7, \quad \square_1(\mathsf{t}_\lambda) \leqslant 2, \quad \square_2(\mathsf{t}_\lambda) \leqslant 5$$

$$\triangle_1(\mathsf{t}_\lambda, \widetilde{\mathsf{t}}_\lambda) \leqslant 5184, \quad \triangle_2(\mathsf{t}_\lambda) \leqslant 1024, \quad \mathsf{C}_{\text{bias}} \leqslant 6, \quad \mathsf{C}_{\text{var}} \leqslant 414.$$

Else,

$$K_{\text{bias}}(\mathsf{t}_\lambda) \leqslant 2e^{6R\|\theta^\star\|}, \quad K_{\text{var}}(\mathsf{t}_\lambda) \leqslant 8e^{4R\|\theta^\star\|}, \quad \square_1(\mathsf{t}_\lambda) \leqslant e^{R\|\theta^\star\|},$$

$$\square_2(\mathsf{t}_\lambda) \leqslant 2e^{3R\|\theta^\star\|}, \quad \triangle_1(\mathsf{t}_\lambda, \widetilde{\mathsf{t}}_\lambda) \leqslant 2304(R\|\theta^\star\|)^2 e^{8R\|\theta^\star\|}, \quad \triangle_2(\mathsf{t}_\lambda) \leqslant 256e^{4R\|\theta^\star\|},$$

$$\mathsf{C}_{\text{bias}} \leqslant 6e^{4R\|\theta^\star\|}, \quad \mathsf{C}_{\text{var}} \leqslant 256e^{6R\|\theta^\star\|}.$$

Proof. For the first bound, we use the fact that $\mathsf{t}_\lambda \leqslant \log 2$ and plug that in the expressions above as these functions are increasing in t_λ. We compute them numerically from the definition.

For the second set of bounds, we simply inject the bounds for t_λ and $\widetilde{\mathsf{t}}_\lambda$ in the second bounds of Definition 7.

D.1 Analytic decomposition of the risk

In this section, we make use of self-concordance to control certain quantities required to control the variance, with respect to our main quantities Bias_λ, r_λ and df_λ. The excess risk has been already decomposed in Sec. 6.

Theorem 7 (Analytic decomposition)**.** *Let* $\lambda > 0$ *and* K_{bias} *and* K_{var} *be the increasing functions of* t_λ *described in Eq.* (37)*. When* $\widehat{\mathsf{Var}}_\lambda \leqslant \mathsf{r}_\lambda(\theta^\star_\lambda)/2$*, then*

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \quad \leqslant \quad K_{\text{bias}}(\mathsf{t}_\lambda)\ \mathsf{Bias}^2_\lambda \ + \ K_{\text{var}}(\mathsf{t}_\lambda)\ \widehat{\mathsf{Var}}^2_\lambda. \tag{35}$$

Moreover $K_{\text{bias}}(\mathsf{t}_\lambda), K_{\text{var}}(\mathsf{t}_\lambda) \leqslant 7$ *if* $\mathsf{Bias}_\lambda \leqslant \frac{1}{2}\mathsf{r}_\lambda(\theta^\star)$*, otherwise* $K_{\text{bias}}(\mathsf{t}_\lambda), K_{\text{var}}(\mathsf{t}_\lambda) \leqslant 8e^{6\|\theta^\star\|\ R}$ *(see Proposition 7 in Appendix D for more precise bounds).*

Proof. Since $\theta^\star$ exists by Assumption 5, using Eq. (13), applied with $\mu = \rho$ and $\lambda = 0$, we have $L(\theta) - L(\theta^\star) \leqslant \psi(\mathsf{t}(\theta - \theta^\star))\|\theta - \theta^\star\|^2_{\mathbf{H}(\theta^\star)}$, for any $\theta \in \mathcal{H}$. By setting $\theta = \widehat{\theta}^\star_\lambda$, we obtain

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leqslant \psi(\mathsf{t}(\widehat{\theta}^\star_\lambda - \theta^\star))\|\widehat{\theta}^\star_\lambda - \theta^\star\|^2_{\mathbf{H}(\theta^\star)}.$$

Using the fact that $\mathbf{H}(\theta^\star) \preceq \mathbf{H}(\theta^\star) + \lambda I =: \mathbf{H}_\lambda(\theta^\star)$, by adding and subtracting $\theta^\star_\lambda$, we have

$$\|\theta^\star_\lambda - \theta^\star\|_{\mathbf{H}(\theta^\star)} \leqslant \|\theta^\star_\lambda - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)} \leqslant \|\theta^\star_\lambda - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)} + \|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star)},$$

and analogously since $\mathsf{t}(\cdot)$ is a (semi)norm, $\mathsf{t}(\widehat{\theta}^\star_\lambda - \theta^\star) \leqslant \mathsf{t}_\lambda + \mathsf{t}(\widehat{\theta}^\star_\lambda - \theta^\star)$, so

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \quad \leqslant \quad \psi(\mathsf{t}_\lambda + \mathsf{t}(\widehat{\theta}^\star_\lambda - \theta^\star_\lambda))\ (\|\theta^\star_\lambda - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)} + \|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star)})^2.$$

By applying Eq. (12) with $\mu = \rho$ and $\theta = \theta^\star$, we have $\mathbf{H}_\lambda(\theta^\star) \preceq e^{\mathsf{t}_\lambda}\mathbf{H}_\lambda(\theta^\star_\lambda)$ and so

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \quad \leqslant \quad \psi(\mathsf{t}_\lambda + \mathsf{t}(\widehat{\theta}^\star_\lambda - \theta^\star_\lambda))\ (\|\theta^\star_\lambda - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)} + e^{\mathsf{t}_\lambda/2}\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)})^2. \tag{36}$$

The terms t_λ and $\|\theta^\star_\lambda - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)}$ are related to the *bias terms*, while the terms $\mathsf{t}(\widehat{\theta}^\star_\lambda - \theta^\star_\lambda)$ and $\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)}$ are related to the *variance term*.

Bounding the bias terms. Recall the definition of the bias $\mathsf{Bias}_\lambda = \|\nabla L_\lambda(\theta^\star)\|_{\mathbf{H}^{-1}_\lambda(\theta^\star)}$. We bound $\mathsf{t}_\lambda = \mathsf{t}(\theta^\star_\lambda - \theta^\star)$, by Lemma 1 and the term $\|\theta^\star - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star)}$ by applying Eq. (14) with $\mu = \rho$ and $\theta = \theta^\star$

$$\|\theta^\star - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star)} \quad \leqslant \quad 1/\underline{\phi}(\mathsf{t}_\lambda)\ \|\nabla L_\lambda(\theta^\star)\|_{\mathbf{H}^{-1}_\lambda(\theta^\star)} \quad = \quad 1/\underline{\phi}(\mathsf{t}_\lambda)\ \mathsf{Bias}_\lambda.$$

Bounding the variance terms. First we bound the term $\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)} := \|\mathbf{H}_\lambda(\theta^\star_\lambda)^{1/2}(\widehat{\theta}^\star_\lambda - \theta^\star_\lambda)\|$, by multiplying and dividing for $\widehat{\mathbf{H}}_\lambda(\theta^\star_\lambda)^{-1/2}$, we have

$$\begin{aligned}\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)} &= \|\mathbf{H}_\lambda(\theta^\star_\lambda)^{1/2}\widehat{\mathbf{H}}_\lambda(\theta^\star_\lambda)^{-1/2}\widehat{\mathbf{H}}_\lambda(\theta^\star_\lambda)^{1/2}(\widehat{\theta}^\star_\lambda - \theta^\star_\lambda)\| \\ &\leqslant \|\mathbf{H}_\lambda(\theta^\star_\lambda)^{1/2}\widehat{\mathbf{H}}_\lambda(\theta^\star_\lambda)^{-1/2}\|\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star_\lambda)}.\end{aligned}$$

Applying Eq. (14) with $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$ and $\theta = \widehat{\theta}^\star_\lambda$, since $L_{\mu,\lambda} = \widehat{L}_\lambda$ for the given choice of μ, we have

$$\|\theta^\star_\lambda - \widehat{\theta}^\star_\lambda\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star_\lambda)} \quad \leqslant \quad \|\nabla \widehat{L}_\lambda(\theta^\star_\lambda)\|_{\widehat{\mathbf{H}}^{-1}_\lambda(\theta^\star_\lambda)}\ /\ \underline{\phi}(\mathsf{t}(\theta^\star_\lambda - \widehat{\theta}^\star_\lambda))$$

and since $\|\nabla \widehat{L}_\lambda(\theta^\star_\lambda)\|_{\widehat{\mathbf{H}}^{-1}_\lambda(\theta^\star_\lambda)} := \|\widehat{\mathbf{H}}^{-1/2}_\lambda(\theta^\star_\lambda)\nabla \widehat{L}_\lambda(\theta^\star_\lambda)\|$, by multiplying and dividing by $\mathbf{H}_\lambda(\theta^\star_\lambda)$, we have:

$$\begin{aligned}\|\widehat{\mathbf{H}}^{-1/2}_\lambda(\theta^\star_\lambda)\nabla \widehat{L}_\lambda(\theta^\star_\lambda)\| &= \|\widehat{\mathbf{H}}^{-1/2}_\lambda(\theta^\star_\lambda)\mathbf{H}_\lambda(\theta^\star_\lambda)^{1/2}\mathbf{H}_\lambda(\theta^\star_\lambda)^{-1/2}\nabla \widehat{L}_\lambda(\theta^\star_\lambda)\| \\ &\leqslant \|\widehat{\mathbf{H}}^{-1/2}_\lambda(\theta^\star_\lambda)\mathbf{H}_\lambda(\theta^\star_\lambda)^{1/2}\|\|\nabla \widehat{L}_\lambda(\theta^\star_\lambda)\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star_\lambda)}.\end{aligned}$$

Then

$$\|\theta_\lambda^\star - \widehat{\theta}_\lambda^\star\|_{\mathbf{H}_\lambda(\theta_\lambda^\star)} \leqslant \frac{1}{\underline{\phi}(\mathsf{t}(\theta_\lambda^\star - \widehat{\theta}_\lambda^\star))} \|\mathbf{H}_\lambda^{1/2}(\theta_\lambda^\star)\mathbf{H}_\lambda{}^{-1/2}(\theta_\lambda^\star)\|^2 \, \|\nabla \widehat{L}_\lambda(\theta_\lambda^\star)\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta_\lambda^\star)} = \frac{\widehat{\mathsf{Var}}_\lambda}{\underline{\phi}(\mathsf{t}(\theta_\lambda^\star - \widehat{\theta}_\lambda^\star))}.$$

To conclude this part of the proof we need to bound $\mathsf{t}(\widehat{\theta}_\lambda^\star - \theta_\lambda^\star)$. Since we require $\widehat{\mathsf{Var}}_\lambda / \mathsf{r}_\lambda(\theta_\lambda^\star) \leqslant 1/2$, by Proposition 5 we have $\mathsf{t}(\widehat{\theta}_\lambda^\star - \theta_\lambda^\star) \leqslant \log 2$.

Gathering the terms. By gathering the results of the previous paragraphs

$$L(\widehat{\theta}_\lambda^\star) - L(\theta^\star) \leqslant \psi(\mathsf{t}_\lambda + \log 2) \; (\, 1/\underline{\phi}(\mathsf{t}_\lambda) \, \mathsf{Bias}_\lambda \; + \; e^{\mathsf{t}_\lambda/2}/\underline{\phi}(\log 2) \, \widehat{\mathsf{Var}}_\lambda \,)^2$$

Using the fact that $(a+b)^2 \leqslant 2a^2 + 2b^2$, we have the desired result, with

$$K_{\text{bias}}(\mathsf{t}_\lambda) = 2\psi(\mathsf{t}_\lambda + \log 2)/\underline{\phi}(\mathsf{t}_\lambda)^2, \;\; K_{\text{var}}(\mathsf{t}_\lambda) = 2\psi(\mathsf{t}_\lambda + \log 2)e^{\mathsf{t}_\lambda}/\underline{\phi}(\log 2)^2. \tag{37}$$

which are bounded in Definition 7 and Proposition 7 of Appendix D.

D.2 Analytic bounds for terms related to the variance

In this lemma, we aim to control the essential supremum and the variance of the random vector $\mathbf{H}_\lambda{}^{-1/2}(\theta_\lambda^\star)\nabla \ell_z^\lambda(\theta_\lambda^\star)$ relating it to quantities at $\theta^\star$. The results will be used to control the variance via Bernstein concentration inequalities, so we are going to control its essential supremum and its variance.

Lemma 5 (Control of $\mathbf{H}_\lambda(\theta_\lambda^\star)^{-1/2}\nabla \ell_z^\lambda(\theta_\lambda^\star)$). *For any $0 < \lambda \leqslant \mathsf{B}_2^\star$, we have*

1. *A bound on the essential supremum:*

$$\sup_{z \in \operatorname{supp}(\rho)} \|\nabla \ell_z(\theta_\lambda^\star)\|_{\mathbf{H}_\lambda{}^{-1}(\theta_\lambda^\star)} \leqslant \square_1 \frac{\mathsf{B}_1^\star}{\sqrt{\lambda}} + 2\square_2 \frac{\mathsf{B}_2^\star}{\lambda}\mathsf{Bias}_\lambda.$$

2. *A bound on the variance*

$$\mathbb{E}\left[\|\nabla \ell_z^\lambda(\theta_\lambda^\star)\|^2_{\mathbf{H}_\lambda{}^{-1}(\theta_\lambda^\star)}\right]^{1/2} \leqslant \square_1 \sqrt{\mathsf{df}_\lambda} + \sqrt{2}\square_2 \sqrt{\frac{\mathsf{B}_2^\star}{\lambda}}\mathsf{Bias}_\lambda,$$

where $\square_1, \square_2$ are increasing functions of t_λ : $\square_1(\mathsf{t}_\lambda) = e^{\mathsf{t}_\lambda/2}$ $\square_2(\mathsf{t}_\lambda) = e^{\mathsf{t}_\lambda/2}(1 + e^{\mathsf{t}_\lambda})$.

Proof. Start by noting that if $\lambda \leqslant \mathsf{B}_2^\star$, then $\sup_{z\in\operatorname{supp}(\rho)} \|\mathbf{H}_\lambda{}^{-1/2}(\theta^\star)\nabla^2 \ell_z^\lambda(\theta^\star)^{1/2}\|^2 \leqslant 1 + \frac{\mathsf{B}_2^\star}{\lambda} \leqslant 2\frac{\mathsf{B}_2^\star}{\lambda}$. Moreover, note that for any vector $h \in \mathcal{H}$, multiplying and dividing by $\nabla^2 \ell_z(\theta^\star)^{1/2}$,

$$\begin{aligned}
\|h\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)} &:= \|\mathbf{H}_\lambda{}^{-1/2}(\theta^\star)\, h\| = \|\mathbf{H}_\lambda{}^{-1/2}(\theta^\star)\; \nabla^2 \ell_z(\theta^\star)^{1/2} \; \nabla^2 \ell_z(\theta^\star)^{-1/2} \; h\| \\
&\leqslant \|\mathbf{H}_\lambda{}^{-1/2}(\theta^\star)\; \nabla^2 \ell_z(\theta^\star)^{1/2}\| \; \|\nabla^2 \ell_z(\theta^\star)^{-1/2}\; h\| \\
&\leqslant \sqrt{\frac{2\mathsf{B}_2^\star}{\lambda}} \|\nabla^2 \ell_z(\theta^\star)^{-1/2}\; h\| \\
&= \sqrt{\frac{2\mathsf{B}_2^\star}{\lambda}} \|h\|_{\nabla^2 \ell_z(\theta^\star)^{-1}} \;\; ,
\end{aligned}$$

where the last bound is mentioned at the beginning of the proof. Similarly, we can show

$$\|h\|_{\nabla^2 \ell_z(\theta^\star)} \leqslant \sqrt{\frac{2\mathsf{B}_2^\star}{\lambda}}\, \|h\|_{\mathbf{H}_\lambda(\theta^\star)}, \qquad \|h\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)} \leqslant \sqrt{\frac{2\mathsf{B}_2^\star}{\lambda}} \|h\|_{\nabla^2 \ell_z(\theta^\star)^{-1}} \;\; . \tag{38}$$

Essential supremum. Let $z \in \mathrm{supp}(\rho)$. First note that using Eq. (27), we have

$$\|\nabla \ell_z^\lambda(\theta_\lambda^\star)\|_{\mathbf{H}_\lambda{}^{-1}(\theta_\lambda^\star)} \leqslant e^{\mathsf{t}_\lambda/2} \|\nabla \ell_z^\lambda(\theta_\lambda^\star)\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)}.$$

Now bound

$$\|\nabla \ell_z^\lambda(\theta_\lambda^\star)\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)} \leqslant \|\nabla \ell_z^\lambda(\theta_\lambda^\star) - \nabla \ell_z^\lambda(\theta^\star)\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)} + \|\nabla \ell_z^\lambda(\theta^\star)\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)}.$$

Since $\nabla \ell_z^\lambda(\theta^\star) = \nabla \ell_z(\theta^\star) + \lambda \theta^\star$, the last term is bounded by

$$\mathsf{Bias}_\lambda + \sup_{z \in \mathrm{supp}(\rho)} \|\nabla \ell_z(\theta^\star)\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)} \leqslant \mathsf{Bias}_\lambda + \frac{\mathsf{B}_1^\star}{\sqrt{\lambda}}.$$

For the first term, start by using Eq. (38).

$$\|\nabla \ell_z^\lambda(\theta_\lambda^\star) - \nabla \ell_z^\lambda(\theta^\star)\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)} \leqslant \sqrt{\frac{2\mathsf{B}_2^\star}{\lambda}}\, \|\nabla \ell_z^\lambda(\theta_\lambda^\star) - \nabla \ell_z^\lambda(\theta^\star)\|_{\nabla^2 \ell_z^\lambda(\theta^\star)^{-1}}.$$

Using Eq. (29) on ℓ_z^λ, we find

$$\|\nabla \ell_z^\lambda(\theta_\lambda^\star) - \nabla \ell_z^\lambda(\theta^\star)\|_{\nabla^2 \ell_z^\lambda(\theta^\star)^{-1}} \leqslant \overline{\phi}(\mathsf{t}_\lambda)\ \|\theta_\lambda^\star - \theta^\star\|_{\nabla^2 \ell_z^\lambda(\theta^\star)}.$$

Applying once again Eq. (38), we bound

$$\|\theta_\lambda^\star - \theta^\star\|_{\nabla^2 \ell_z^\lambda(\theta^\star)} \leqslant \sqrt{\frac{2\mathsf{B}_2^\star}{\lambda}}\ \|\theta_\lambda^\star - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)}.$$

Finally, using Eq. (28) on L_λ, we get

$$\|\theta_\lambda^\star - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)} \leqslant \frac{1}{\underline{\phi}(\mathsf{t}_\lambda)} \mathsf{Bias}_\lambda.$$

Hence, putting things together, we get

$$\|\nabla \ell_z^\lambda(\theta_\lambda^\star) - \nabla \ell_z^\lambda(\theta^\star)\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)} \leqslant \frac{2\mathsf{B}_2^\star}{\lambda} \frac{\overline{\phi}(\mathsf{t}_\lambda)}{\underline{\phi}(\mathsf{t}_\lambda)} \mathsf{Bias}_\lambda = \frac{2\mathsf{B}_2^\star}{\lambda} e^{\mathsf{t}_\lambda} \mathsf{Bias}_\lambda.$$

We the combine all our different computation to get the bound.

Variance. We start by using Eq. (27) to show that

$$\mathbb{E}\left[\|\nabla \ell_z^\lambda(\theta_\lambda^\star)\|^2_{\mathbf{H}_\lambda{}^{-1}(\theta_\lambda^\star)}\right]^{1/2} \leqslant e^{\mathsf{t}_\lambda/2} \mathbb{E}\left[\|\nabla \ell_z^\lambda(\theta_\lambda^\star)\|^2_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)}\right]^{1/2}.$$

Then we use the triangle inequality

$$\mathbb{E}\left[\|\nabla \ell_z^\lambda(\theta_\lambda^\star)\|^2_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)}\right]^{1/2} \leqslant \mathbb{E}\left[\|\nabla \ell_z^\lambda(\theta_\lambda^\star) - \nabla \ell_z^\lambda(\theta^\star)\|^2_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)}\right]^{1/2} + \mathbb{E}\left[\|\nabla \ell_z^\lambda(\theta^\star)\|^2_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)}\right]^{1/2}.$$

We can easily bound the last term on the right hand side by $\mathsf{Bias}_\lambda + \mathsf{df}_\lambda$. For the first term, we proceed as in the previous case to obtain

$$\forall z \in \mathrm{supp}(\rho),\ \|\nabla \ell_z^\lambda(\theta_\lambda^\star) - \nabla \ell_z^\lambda(\theta^\star)\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)} \leqslant \sqrt{\frac{2\mathsf{B}_2^\star}{\lambda}} \overline{\phi}(\mathsf{t}_\lambda) \|\theta_\lambda^\star - \theta^\star\|_{\nabla^2 \ell_z^\lambda(\theta^\star)}.$$

Now taking the expectancy of this inequality squared,

$$\mathbb{E}\left[\|\nabla \ell_z^\lambda(\theta_\lambda^\star) - \nabla \ell_z^\lambda(\theta^\star)\|^2_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)}\right]^{1/2} \leqslant \sqrt{\frac{2\mathsf{B}_2^\star}{\lambda}}\overline{\phi}(\mathsf{t}_\lambda)\mathbb{E}\left[\|\theta_\lambda^\star - \theta^\star\|^2_{\nabla^2 \ell_z^\lambda(\theta^\star)}\right]^{1/2}$$
$$= \sqrt{\frac{2\mathsf{B}_2^\star}{\lambda}}\overline{\phi}(\mathsf{t}_\lambda)\|\theta_\lambda^\star - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)},$$

where the last equality comes from $\mathbb{E}\left[\nabla^2 \ell_Z^\lambda(\theta^\star)\right] = \mathbf{H}_\lambda(\theta^\star)$. Now applying Eq. (28) to L_λ, we obtain

$$\|\theta_\lambda^\star - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)} \leqslant \frac{1}{\underline{\phi}(\mathsf{t}_\lambda)}\mathsf{Bias}_\lambda.$$

Regrouping all these bounds, we obtain

$$\mathbb{E}\left[\|\nabla \ell_z^\lambda(\theta_\lambda^\star) - \nabla \ell_z^\lambda(\theta^\star)\|^2_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)}\right]^{1/2} \leqslant e^{\mathsf{t}_\lambda}\sqrt{\frac{2\mathsf{B}_2^\star}{\lambda}}\mathsf{Bias}_\lambda.$$

Hence the final bound is proved, regrouping all our computations.

D.3 Concentration lemmas

Here we concentrate in high probability the quantities obtained in the analytical decomposition. Details on the proof technique are given in Sec. 6 of the paper.

Lemma 6 (Equivalence of empirical and expected Hessian)**.** *Let $\theta \in \mathcal{H}$ and $n \in \mathbb{N}$. For any $\delta \in (0,1]$, $\lambda > 0$, if*

$$n \geqslant 24\frac{\mathsf{B}_2(\theta)}{\lambda}\log\frac{8\mathsf{B}_2(\theta)}{\lambda\delta}, \tag{39}$$

then with probability at least $1-\delta$: $\mathbf{H}_\lambda(\theta) \preceq 2\widehat{\mathbf{H}}_\lambda(\theta)$, or equivalently

$$\|\mathbf{H}_\lambda{}^{1/2}(\theta)\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta)\|^2 \leqslant 2.$$

Proof. By Remark 4 and the definition of $\mathsf{B}_2(\theta)$, the condition we require on n is sufficient to apply Proposition 10, in particular Eq. (51), to $\mathbf{H}_\lambda(\theta), \widehat{\mathbf{H}}_\lambda(\theta)$, for $t = 1/2$, which provides the desired result.

Lemma 7 (Concentration of the empirical gradient)**.** *Let $n \in \mathbb{N}, \delta \in (0,1]$, $0 < \lambda \leqslant \mathsf{B}_2^\star$. For any $k \geqslant 4$, if $n \geqslant k^2\square_2^2\,\frac{\mathsf{B}_2^\star}{\lambda}\log\frac{2}{\delta}$, then with probability at least $1-\delta$, we have*

$$\|\nabla \widehat{L}_\lambda(\theta_\lambda^\star)\|_{\mathbf{H}_\lambda{}^{-1}(\theta_\lambda^\star)} \leqslant \frac{2\sqrt{3}}{k}\mathsf{Bias}_\lambda + 2\square_1\sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2\,\log\frac{2}{\delta}}{n}}. \tag{40}$$

Here, $\square_1, \square_2$ are defined in Lemma 5 in Appendix D and $(\mathsf{Q}^\star)^2 = (\mathsf{B}_1^\star)^2/\mathsf{B}_2^\star$.

Proof.1) First let us concentrate $\mathbf{H}_\lambda(\theta_\lambda^\star)^{-1/2}\nabla\widehat{L}_\lambda(\theta_\lambda^\star)$ using a Bernstein-type inequality.

We can see $\mathbf{H}_\lambda(\theta_\lambda^\star)^{-1/2}\nabla\widehat{L}_\lambda(\theta_\lambda^\star)$ as the mean of n i.i.d. random variables distributed from the law of the vector $\mathbf{H}_\lambda(\theta_\lambda^\star)^{-1/2}\nabla\ell_z(\theta_\lambda^\star)$.

As we have shown in Lemma 5, the essential supremum and variance of this vector is bounded, then we can use Bernstein inequality for random vectors [e.g. Thm. 3.3.4 of 42]: for any $\lambda > 0$, any $n \in \mathbb{N}$ and $\delta \in (0,1]$, with probability at least $1-\delta$, we have

$$\|\mathbf{H}_\lambda(\theta_\lambda^\star)^{-1/2}\nabla\widehat{L}_\lambda(\theta_\lambda^\star)\| \leqslant \frac{2M\log\frac{2}{\delta}}{n} + \sigma\sqrt{\frac{2\log\frac{2}{\delta}}{n}},$$

where $M = \sup_{z \in \mathrm{supp}(\rho)} \|\nabla \ell_z(\theta_\lambda^\star)\|_{\mathbf{H}_\lambda{}^{-1}(\theta_\lambda^\star)}$ and $\sigma = \mathbb{E}\left[\|\nabla \ell_z(\theta_\lambda^\star)\|^2_{\mathbf{H}_\lambda{}^{-1}(\theta_\lambda^\star)}\right]^{1/2}$.

2) Using the bounds obtained in Lemma 5,

$$M \leqslant \square_1 \frac{\mathsf{B}_1^\star}{\sqrt{\lambda}} + 2\square_2 \frac{\mathsf{B}_2^\star}{\lambda}\mathsf{Bias}_\lambda, \qquad \sigma \leqslant \square_1 \sqrt{\mathsf{df}_\lambda} + \sqrt{2}\square_2 \frac{\sqrt{\mathsf{B}_2^\star}}{\sqrt{\lambda}}\mathsf{Bias}_\lambda.$$

3) Injecting these in the Bernstein inequality,

$$\begin{aligned}
\|\mathbf{H}_\lambda(\theta_\lambda^\star)^{-1/2}\nabla \widehat{L}_\lambda(\theta_\lambda^\star)\| \leqslant& \frac{2\left(\square_1\ \mathsf{B}_1^\star/\sqrt{\lambda} + 2\square_2\ (\mathsf{B}_2^\star/\lambda)\mathsf{Bias}_\lambda\right)\ \log\frac{2}{\delta}}{n} \\
&+ \left(\square_1\ \mathsf{df}_\lambda^{1/2} + \sqrt{2}\square_2\ \sqrt{\mathsf{B}_2^\star/\lambda}\mathsf{Bias}_\lambda\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n}} \\
=& \left[\frac{4\square_2\ \mathsf{B}_2^\star\ \log\frac{2}{\delta}}{\lambda n} + \sqrt{\frac{4\square_2^2\ \mathsf{B}_2^\star\ \log\frac{2}{\delta}}{\lambda n}}\right]\mathsf{Bias}_\lambda \\
&+ \sqrt{\frac{2\square_1^2\ \mathsf{df}_\lambda\ \log\frac{2}{\delta}}{n}} + \sqrt{\frac{2\ \mathsf{B}_2^\star\ \log\frac{2}{\delta}}{\lambda n}}\sqrt{\frac{2\square_1^2\ (\mathsf{B}_1^\star)^2/\mathsf{B}_2^\star\ \log\frac{2}{\delta}}{n}}.
\end{aligned}$$

In the last inequality, we have regrouped the terms with a factor Bias_λ and we have separated the first term of the decomposition in the following way :

$$\frac{2\square_1 \mathsf{B}_1^\star \log\frac{2}{\delta}}{\sqrt{\lambda}n} = \sqrt{\frac{2\ \mathsf{B}_2^\star\ \log\frac{2}{\delta}}{\lambda n}}\sqrt{\frac{2\square_1^2\ (\mathsf{Q}^\star)^2\ \log\frac{2}{\delta}}{n}}.$$

Hence, we can bound the second line of the last inequality:

$$\sqrt{\frac{2\square_1^2 \mathsf{df}_\lambda \log\frac{2}{\delta}}{n}} + \sqrt{\frac{2\mathsf{B}_2^\star \log\frac{2}{\delta}}{\lambda n}}\sqrt{\frac{2\square_1^2 (\mathsf{Q}^\star)^2 \log\frac{2}{\delta}}{n}} \leqslant \left(1 + \sqrt{\frac{2\mathsf{B}_2^\star \log\frac{2}{\delta}}{\lambda n}}\right)\sqrt{\frac{2\square_1^2 \mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2 \log\frac{2}{\delta}}{n}}.$$

Thus, if we assume that $n \geqslant k^2 \square_2^2 \frac{\mathsf{B}_2^\star}{\lambda} \log\frac{2}{\delta}$,

$$\|\mathbf{H}_\lambda(\theta_\lambda^\star)^{-1/2}\nabla \widehat{L}_\lambda(\theta_\lambda^\star)\| \leqslant \left(\frac{4}{k^2} + \frac{2}{k}\right)\mathsf{Bias}_\lambda + \left(1 + \frac{\sqrt{2}}{k}\right)\sqrt{\frac{2\square_1^2\ \mathsf{df}_\lambda \vee (\mathsf{B}_1^\star)^2/\mathsf{B}_2^\star\ \log\frac{2}{\delta}}{n}}.$$

In particular, for $k \geqslant 4$,

$$\|\mathbf{H}_\lambda(\theta_\lambda^\star)^{-1/2}\nabla \widehat{L}_\lambda(\theta_\lambda^\star)\| \leqslant \frac{3}{k}\mathsf{Bias}_\lambda + 2\square_1\sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2\ \log\frac{2}{\delta}}{n}}.$$

Lemma 8 (control of $\widehat{\mathsf{Var}}_\lambda$)**.** *Let $n \in \mathbb{N}$, $\delta \in (0,1]$ and $0 < \lambda \leqslant \mathsf{B}_2^\star$. Assume that for a certain $k \geqslant 5$,*

$$n \geqslant k^2 \square_2^2 \frac{\mathsf{B}_2^\star}{\lambda} \log\frac{8\square_1^2 \mathsf{B}_2^\star}{\lambda\delta}.$$

Then with probability at least $1 - 2\delta$, we have

$$\widehat{\mathsf{Var}}_\lambda \leqslant \frac{6}{k}\mathsf{Bias}_\lambda + 4\square_1\sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2\ \log\frac{2}{\delta}}{n}}.$$

Here, $\square_1, \square_2$ are defined in Lemma 5

Proof.

- First we apply Lemma 6 to $\theta = \theta^\star_\lambda$. Since $\mathsf{B}_2(\theta^\star_\lambda) \leqslant e^{\mathsf{t}_\lambda}\ \mathsf{B}^\star_2 = \square_1^2\ \mathsf{B}^\star_2$, we see that the condition
$$n \geqslant 24\frac{\mathsf{B}_2(\theta^\star_\lambda)}{\lambda} \log \frac{8\mathsf{B}_2(\theta^\star_\lambda)}{\lambda\delta}$$
is satisfied if
$$n \geqslant 24\square_1^2\frac{\mathsf{B}^\star_2}{\lambda} \log \frac{8\square_1^2\mathsf{B}^\star_2}{\lambda\delta}.$$
Because $k \geqslant 5$ and $\square_2 \geqslant \square_1$, and we see that the assumption of this lemma imply the conditions above and hence Lemma 6 is satisfied. In particular, $\|\mathbf{H}_\lambda(\theta^\star_\lambda)^{1/2}\widehat{\mathbf{H}}_\lambda(\theta^\star_\lambda)^{-1/2}\|^2 \leqslant 2$.

- Note that the condition of this proposition also imply the conditions of Lemma 7, because $\lambda \leqslant \mathsf{B}^\star_2$ and $\square_1 \geqslant 1$ imply $\frac{\square_1^2\mathsf{B}^\star_2}{\lambda\delta} \geqslant \frac{1}{\delta}$.

D.4 Final results

First, we find conditions on n such that the hypothesis $\widehat{\mathsf{Var}}_\lambda \leqslant \frac{\mathsf{r}_\lambda(\theta^\star_\lambda)}{2}$ is satisfied.

Lemma 9. *Let $n \in \mathbb{N}$, $\delta \in (0,1]$, $0 < \lambda \leqslant \mathsf{B}^\star_2$ and*
$$n \geqslant \triangle_1\frac{\mathsf{B}^\star_2}{\lambda} \log \frac{8\square_1^2\mathsf{B}^\star_2}{\lambda\delta}, \quad n \geqslant \triangle_2\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2}{\mathsf{r}_\lambda(\theta^\star)^2} \log \frac{2}{\delta},$$
then with probability at least $1-2\delta$
$$\frac{\widehat{\mathsf{Var}}_\lambda}{\mathsf{r}_\lambda(\theta^\star_\lambda)} \leqslant \square_1\frac{\widehat{\mathsf{Var}}_\lambda}{\mathsf{r}_\lambda(\theta^\star)} \leqslant \frac{1}{2},$$
where $\square_1$, $\triangle_1, \triangle_2$ are constants defined in Definition 7.

Proof. Recall that $\widetilde{\mathsf{t}}_\lambda = \frac{\mathsf{Bias}_\lambda}{\mathsf{r}_\lambda(\theta^\star)}$.
Using Lemma 8, we see that under the conditions of this lemma, we have
$$\square_1\frac{\widehat{\mathsf{Var}}_\lambda}{\mathsf{r}_\lambda(\theta^\star)} \leqslant \frac{6\square_1\ \mathsf{Bias}_\lambda}{k\ \mathsf{r}_\lambda(\theta^\star)} + 4\square_1^2\sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2\ \log\frac{2}{\delta}}{n\ \mathsf{r}_\lambda(\theta^\star)^2}}.$$

Thus, taking $k = 24\square_1(1/2 \vee \widetilde{\mathsf{t}}_\lambda)$ and $n \geqslant 256\square_1^4\frac{\mathsf{df}_\lambda \vee \mathsf{Q}^\star}{\mathsf{r}_\lambda(\theta^\star)^2}\ \log\frac{2}{\delta}$, both terms in the sum are bounded by $1/4$ hence the result.

Note that here, we have defined
$$\triangle_1 = 576\square_1^2\square_2^2(1/2 \vee \widetilde{\mathsf{t}}_\lambda)^2, \quad \triangle_2 = 256\square_1^4,$$
hence the constants in the definition above.

Proof.of Thm. 4 First we recall that $\triangle_1, \triangle_2, \square_1, \mathsf{C}_{\text{bias}}$ and C_{var} are defined in Definition 7, and bounded in Proposition 7.

First note that, given the requirements on n, by Lemma 9, we have $\widehat{\mathsf{Var}}_\lambda \leqslant \frac{\mathsf{r}_\lambda(\theta^\star_\lambda)}{2}$ with probability at least $1-2\delta$. Thus, we are in a position to apply Thm. 7 :
$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leqslant K_{\text{bias}}\ \mathsf{Bias}^2_\lambda + K_{\text{var}}\ \widehat{\mathsf{Var}}^2_\lambda,$$

with $K_{\text{bias}}, K_{\text{var}}$ defined in the proof of the theorem. Note that in the proof of Lemma 9, we have taken $k = 24\square_1(1/2 \vee \widetilde{\mathsf{t}}_\lambda) \geqslant 12$. Hence, using Lemma 8, we find

$$\widehat{\mathsf{Var}}_\lambda \leqslant \frac{1}{2}\mathsf{Bias}_\lambda + 4\square_1 \sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2 \ \log\frac{2}{\delta}}{n}}.$$

Hence,

$$\widehat{\mathsf{Var}}_\lambda^2 \leqslant \frac{1}{2}\mathsf{Bias}_\lambda^2 + 32\square_1^2 \frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2 \ \log\frac{2}{\delta}}{n},$$

which yields the wanted result with $\mathsf{C}_{\text{bias}} = K_{\text{bias}} + \frac{1}{2}K_{\text{var}}$ and $\mathsf{C}_{\text{var}} = 32\square_1^2 K_{\text{var}}$.

Proof.of Thm. 3

We get this theorem as a corollary of Thm. 4. Indeed, $\forall \lambda \leqslant \mathsf{B}_2^\star$, $\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^* \leqslant \frac{(\mathsf{B}_1^\star)^2}{\lambda}$, hence the result.

E Explicit bounds for the simplified case

In this section, assume that Assumptions 1, 3 to 5 and 8 hold.

Define the following constant N :

$$N = 36A^2 \log^2\left(6A^2\frac{1}{\delta}\right) \quad \vee \quad 256\frac{1}{A^2} \ \log\frac{2}{\delta} \quad \vee \quad 512\left(\|\theta^\star\|^2 R^2 \vee 1\right) \log\frac{2}{\delta}, \tag{41}$$

where $A = \frac{\overline{\mathsf{B}}_2}{\overline{\mathsf{B}}_1}$.

We have the following slow rates theorem.

Theorem 8 (Quantitative slow rates result). *Let $n \in \mathbb{N}$. Let $\delta \in (0, 1]$. Setting*

$$\lambda = 16((R \vee 1)\overline{\mathsf{B}}_1) \ \frac{1}{\sqrt{n}} \ \log^{1/2}\frac{2}{\delta},$$

if $n \geqslant N$, with probability at least $1 - 2\delta$,

$$L(\widehat{\theta}_\lambda^\star) - L(\theta^\star) \leqslant 48 \ \ \max(R, 1)\max(\|\theta^\star\|^2, 1)\overline{\mathsf{B}}_1\frac{1}{\sqrt{n}} \ \log^{1/2}\frac{2}{\delta}, \tag{42}$$

and $N = O\left(\text{poly}(\overline{\mathsf{B}}_1, \overline{\mathsf{B}}_2, R\|\theta^\star\|)\right)$ is given explicitly in Eq. (41). Here, poly *denotes a certain rational function of the inputs.*

Proof. Note that $\overline{\mathsf{df}}_\lambda \leqslant \frac{\overline{\mathsf{B}}_1^2}{\lambda}$. Hence, if $\lambda \leqslant \overline{\mathsf{B}}_2$, then $\overline{\mathsf{df}}_\lambda \vee (\overline{\mathsf{B}}_1^2/\overline{\mathsf{B}}_2) \leqslant \frac{\overline{\mathsf{B}}_1^2}{\lambda}$.

1) Let us reformulate Thm. 6. Let $n \in \mathbb{N}$ and $0 < \lambda \leqslant \overline{\mathsf{B}}_2$. Let $\delta \in (0, 1]$. If

$$n \geqslant 512\left(\|\theta^\star\|^2 R^2 \vee 1\right) \log\frac{2}{\delta}, \quad n \geqslant 24\frac{\overline{\mathsf{B}}_2}{\lambda} \log\frac{8\overline{\mathsf{B}}_2}{\lambda\delta}, \quad n \geqslant 16\triangle^2 \ \frac{R^2\overline{\mathsf{B}}_1^2}{\lambda^2} \ \log\frac{2}{\delta},$$

then with probability at least $1 - 2\delta$,

$$L(\widehat{\theta}_\lambda^\star) - L(\theta^\star) \leqslant \mathsf{C}_{\text{var}} \ \frac{\overline{\mathsf{B}}_1^2}{\lambda n} \ \log\frac{2}{\delta} + \mathsf{C}_{\text{bias}}\lambda\|\theta^\star\|^2,$$

where $\triangle, \mathsf{C}_{\text{bias}}, \mathsf{C}_{\text{var}}$ are defined in Definition 5.

2) Now setting $\lambda = 16R\overline{\mathsf{B}}_1 \log^{1/2}\frac{2}{\delta}\frac{1}{n^{1/2}}$, we see that the inequality

$$n \geqslant 16\triangle^2 \frac{R^2\overline{\mathsf{B}}_1^2}{\lambda^2} \log\frac{2}{\delta}$$

is automatically satisfied since $\triangle \leqslant 4$. Hence, if

$$n \geqslant 512\left(\|\theta^\star\|^2 R^2 \vee 1\right)\log\frac{2}{\delta}, \quad n \geqslant 24\frac{\overline{\mathsf{B}}_2}{\lambda}\log\frac{8\overline{\mathsf{B}}_2}{\lambda\delta}, \quad 0 < \lambda \leqslant \overline{\mathsf{B}}_2,$$

then

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leqslant \frac{\mathsf{C}_{\text{var}}}{256}\frac{1}{R^2}\lambda + \mathsf{C}_{\text{bias}}\lambda\|\theta^\star\|^2 \leqslant \left(\frac{\mathsf{C}_{\text{var}}}{256} + \mathsf{C}_{\text{bias}}\right)\max(\frac{1}{R^2}, \|\theta^\star\|^2)\lambda.$$

Since by Definition 5, $\mathsf{C}_{\text{var}} \leqslant 84$ and $\mathsf{C}_{\text{bias}} \leqslant 2$, we get

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leqslant 3\max(\frac{1}{R^2}, \|\theta^\star\|^2)\lambda.$$

3) Having our fixed $\lambda = 16\frac{\overline{\mathsf{B}}_1 R\ \log^{1/2}\frac{2}{\delta}}{n^{1/2}}$, let us look for conditions for which

$$n \geqslant 512\left(\|\theta^\star\|^2 R^2 \vee 1\right)\log\frac{2}{\delta}, \quad n \geqslant 24\frac{\overline{\mathsf{B}}_2}{\lambda}\log\frac{8\overline{\mathsf{B}}_2}{\lambda\delta}, \quad 0 < \lambda \leqslant \overline{\mathsf{B}}_2,$$

are satisfied.

To deal with $n \geqslant 24\frac{\overline{\mathsf{B}}_2}{\lambda}\log\frac{8\overline{\mathsf{B}}_2}{\lambda\delta}$, bound

$$\frac{\overline{\mathsf{B}}_2}{\lambda} \leqslant \frac{1}{16}\frac{\overline{\mathsf{B}}_2}{R\overline{\mathsf{B}}_1\ \log^{1/2}\frac{2}{\delta}}\, n^{1/2} \leqslant \frac{1}{8}\frac{\overline{\mathsf{B}}_2}{R\overline{\mathsf{B}}_1}\, n^{1/2},$$

where we have used the fact that $\log^{1/2}\frac{2}{\delta} \geqslant \frac{1}{2}$. apply Lemma 14 with $a_1 = 3, a_2 = 1, A = \frac{\overline{\mathsf{B}}_2}{R\overline{\mathsf{B}}_1}$ to get the following condition:

$$n \geqslant 4a_1^2 A^2 \log^2\left(\frac{2a_1a_2A^2}{\delta}\right),$$

which we express as

$$n \geqslant 36A^2\log^2\left(6A^2\frac{1}{\delta}\right).$$

To deal with the bound $\lambda < \overline{\mathsf{B}}_2$, we need only apply the definition to obtain

$$n \geqslant 256\frac{R^2\overline{\mathsf{B}}_1^2}{\overline{\mathsf{B}}_2^2}\ \log\frac{2}{\delta}.$$

Thus, we can concentrate all these bounds as $n \geqslant N$ where

$$N = 36A^2\log^2\left(6A^2\frac{1}{\delta}\right) \quad \vee \quad 256\frac{1}{A^2}\ \log\frac{2}{\delta} \quad \vee \quad 512\left(\|\theta^\star\|^2R^2 \vee 1\right)\log\frac{2}{\delta},$$

where $A = \frac{\overline{\mathsf{B}}_2}{R\overline{\mathsf{B}}_1}$.

4) Since R is only an upper bound, we can replace R by $R \vee 1$. In this case, we see that $A \leqslant \frac{\overline{\mathsf{B}}_2}{\overline{\mathsf{B}}_1}$ and $\max(\frac{1}{R\vee 1}, (R\vee 1)\|\theta^\star\|^2) \leqslant (R\vee 1)(\|\theta^\star\| \vee 1)^2$ hence the final bounds.

F Explicit bounds for the refined case

In this part, we continue to assume Assumptions 1, 3 to 5 and 8. We present a classification of distributions ρ and show that we can achieve better rates than the classical slow rates.

Definition 8 (class of distributions)**.** *Let* $\alpha \in [1, +\infty]$ *and* $r \in [0, 1/2]$.
We denote with $\mathcal{P}_{\alpha,r}$ *the set of probability distributions* ρ *such that there exists* $\mathsf{L}, \mathsf{Q} \geqslant 0$,

- $\mathsf{Bias}_\lambda \leqslant \mathsf{L}\, \lambda^{\frac{1+2r}{2}}$
- $\mathsf{df}_\lambda \leqslant \mathsf{Q}^2\, \lambda^{-1/\alpha}$,

where this holds for any $0 < \lambda \leqslant 1$. *For simplicity, if* $\alpha = +\infty$, *we assume that* $\mathsf{Q} \geqslant \mathsf{Q}^\star$.

Note that given our assumptions, we always have

$$\rho \in \mathcal{P}_{1,0}, \quad \mathsf{L} = \|\theta^\star\|,\ \mathsf{Q} = \mathsf{B}_1^\star. \tag{43}$$

We also define

$$\lambda_1 = \left(\frac{\mathsf{Q}}{\mathsf{Q}^\star}\right)^{2\alpha} \wedge 1, \tag{44}$$

such that

$$\forall \lambda \leqslant \lambda_1,\ \mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2 \leqslant \frac{\mathsf{Q}^2}{\lambda^{1/\alpha}}.$$

Interpretation of the classes

- The bias term Bias_λ characterizes the regularity of the objective $\theta^\star$. In a sense, if r is big, then this means $\theta^\star$ is very regular and will be easier to estimate. The following results reformulates this intuition.

  Remark 1 (source condition)**.** *Assume there exists* $0 \leqslant r \leqslant 1/2$ *and* $v \in \mathcal{H}$ *such that*

  $$P_{\mathbf{H}(\theta^\star)}\theta^\star = \mathbf{H}(\theta^\star)^r v.$$

  Then we have

  $$\forall \lambda > 0,\ \mathsf{Bias}_\lambda \leqslant \mathsf{L}\, \lambda^{\frac{1+2r}{2}}, \quad \mathsf{L} = \|\mathbf{H}(\theta^\star)^{-r}\theta^\star\|.$$

- The effective dimension df_λ characterizes the size of the space $\mathcal{H}$ with respect to the problem. The higher α, the smaller the space. If $\mathcal{H}$ is finite dimensional for instance, $\alpha = +\infty$.

We will give explicit bounds for the performance of $\widehat{\theta}^\star_\lambda$ depending on which class ρ belongs to, i.e., as a function of α, r.

Well -behaved problems $\mathsf{r}_\lambda(\theta^\star)$ has a limiting role. However, as soon as we have some sort of regularity, this role is no longer limiting, i.e. this quantity does not appear in the final rates and the constants in these rates have no dependence on the problem. This motivates the following definition.

We say that a problem is well behaved if the following equation holds.

$$\forall \delta \in (0, \frac{1}{2}],\ \exists \lambda_0(\delta) \in (0,1],\ \forall 0 < \lambda \leqslant \lambda_0(\delta),\ \frac{\mathsf{L}\lambda^{1/2+r}}{\mathsf{r}_\lambda(\theta^\star)}\ \log\frac{2}{\delta} \leqslant \frac{1}{2}. \tag{45}$$

Remark 2 (well-behaved problems)**.** *Note that Eq. (45) is satisfied if one of the following holds.*

- *If $R = 0$, then the condition holds for $\lambda_0 = 1$.*
- *If $r > 0$, then the condition holds for $\lambda_0 = (2\mathsf{L}R\log\frac{2}{\delta})^{-1/r} \wedge 1$.*
- *If there exists $\mu \in [0,1)$ and $\mathsf{F} \geqslant 0$ such that $\mathsf{r}_\lambda(\theta^\star) \geqslant \frac{1}{\mathsf{F}}\lambda^{\mu/2}$, then this holds for $\lambda_0 = (2R\mathsf{F}\log\frac{2}{\delta})^{-2/(1-\mu+2r)} \wedge 1$.*

Moreover, if Eq. (45) is satisfied, than for any $\lambda \leqslant \lambda_0$, $\mathsf{t}_\lambda \leqslant \log 2$.

Note that the first possible condition corresponds to the case where the loss functions are quadratic in θ (if the loss is the square loss for instance). The second condition corresponds to having a strict source condition, i.e. something strictly better than just $\theta^\star \in \mathcal{H}$. Finally, the third condition corresponds to the fact that the radius r_λ decreases slower than the original bound of $\mathsf{r}_\lambda \geqslant \frac{\lambda^{1/2}}{R}$, and hence it is not limiting.

Note that a priori, using only the assumptions, our problems do not satisfy Eq. (45) (see Eq. (43), and the fact that $\mathsf{r}_\lambda \geqslant \frac{\sqrt{\lambda}}{R}$).

F.1 Quantitative bounds

In this section, for any given pair (α, r) characterizing the regularity and size of the problem, we associate

$$\beta = \frac{1}{1+2r+1/\alpha}, \quad \gamma = \frac{\alpha(1+2r)}{\alpha(1+2r)+1}.$$

In what follows, we define

$$N = \frac{256\mathsf{Q}^2}{\mathsf{L}^2}\left(\mathsf{B}_2^\star \wedge \lambda_0 \wedge \lambda_1\right)^{-1/\beta} \quad \vee \quad \left(1296\frac{1}{1-\beta}A\log\left(5184\frac{1}{1-\beta}A^2\frac{1}{\delta}\right)\right)^{1/(1-\beta)}, \quad (46)$$

where $A = \frac{\mathsf{B}_2^\star \mathsf{L}^{2\beta}}{\mathsf{Q}^{2\beta}}$, λ_0 is given by Eq. (45) and λ_1 is given by Eq. (44) : $\lambda_1 = \frac{\mathsf{Q}^{2\alpha}}{(\mathsf{Q}^\star)^{2\alpha}}$.

Theorem 9 (Quantitative results when Eq. (45) is satisfied and $\alpha < \infty$ or $r > 0$)**.** *Let $\rho \in \mathcal{P}_{\alpha,r}$ and that we have either $\alpha < \infty$ or $r > 0$. Let $\delta \in (0, \frac{1}{2}]$.*
If Eq. (45) is satisfied, and

$$n \geqslant N, \quad \lambda = \left(256\left(\frac{\mathsf{Q}}{\mathsf{L}}\right)^2\frac{1}{n}\right)^\beta,$$

then with probability at least $1-2\delta$,

$$L(\widehat{\theta}_\lambda^\star) - L(\theta^\star) \leqslant 8\,(256)^\gamma\,\left(\mathsf{Q}^\gamma\,\mathsf{L}^{1-\gamma}\right)^2\,\frac{1}{n^\gamma}\log\frac{2}{\delta},$$

where N is defined in Eq. (46).

Proof.
Using the definition of λ_1, as soon as $\lambda \leqslant \lambda_1$ we have $\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2 \leqslant \mathsf{Q}^2\lambda^{-1/\alpha}$.

Let us formulate Thm. 4 using the fact that $\rho \in \mathcal{P}_{\alpha,r}$.

Let $\delta \in (0,1]$, $0 < \lambda \leqslant \mathsf{B}_2^\star \wedge \lambda_1$ and $n \in \mathbb{N}$ such that

$$n \geqslant \triangle_1 \frac{\mathsf{B}_2^\star}{\lambda} \log \frac{8\square_1^2 \mathsf{B}_2^\star}{\lambda\delta}, \quad n \geqslant \triangle_2 \frac{\mathsf{Q}^2}{\lambda^{1/\alpha} \mathsf{r}_\lambda(\theta^\star)^2} \log \frac{2}{\delta},$$

then with probability at least $1-2\delta$

$$L(\widehat{\theta}_\lambda^\star) - L(\theta^\star) \leqslant \mathsf{C}_{\text{bias}}\, \mathsf{L}^2 \lambda^{1+2r} + \mathsf{C}_{\text{var}} \frac{\mathsf{Q}^2}{\lambda^{1/\alpha} n} \log \frac{2}{\delta},$$

where $\mathsf{C}_{\text{bias}}, \mathsf{C}_{\text{var}}$ are defined in Definition 7. Now let us distinguish the two cases of our theorem.

Assume that ρ satisfies Eq. (45) . In this case the proof proceeds as follows. Note that as soon as $\lambda \leqslant \lambda_0$, we have $\frac{\mathsf{Bias}_\lambda}{\mathsf{r}_\lambda(\theta^\star)} \leqslant \frac{1}{2}$ and hence the bounds in Proposition 7 apply.
1) First, we find a simple condition to guarantee

$$\mathsf{r}_\lambda(\theta^\star)^2 \lambda^{1/\alpha} \geqslant \triangle_2\, \mathsf{Q}^2 \frac{1}{n} \log \frac{2}{\delta}.$$

Using the fact that Eq. (45) is satisfied, we see that if $\lambda \leqslant \lambda_0$, then $\mathsf{r}_\lambda \geqslant 2\mathsf{L}\lambda^{1/2+r} \log \frac{2}{\delta}$. Hence, this condition is satisfied if

$$\lambda \leqslant \lambda_0, \quad 4\mathsf{L}^2 \lambda^{1+2r+1/\alpha} \geqslant \triangle_2\, \mathsf{Q}^2 \frac{1}{n}.$$

2) Now fix $\mathsf{C}_\lambda = 256 \geqslant \triangle_2/4$ (see Proposition 7) and fix

$$\lambda^{1+2r+1/\alpha} = \mathsf{C}_\lambda \frac{\mathsf{Q}^2}{\mathsf{L}^2} \frac{1}{n} \iff \lambda = \left(\mathsf{C}_\lambda \frac{\mathsf{Q}^2}{\mathsf{L}^2} \frac{1}{n}\right)^\beta.$$

where $\beta = 1/(1+2r+1/\lambda) \in [1/2, 1)$.

Using our restatement of Thm. 4, we have that with probability at least $1-2\delta$,

$$L(\widehat{\theta}_\lambda^\star) - L(\theta^\star) \leqslant \left(\mathsf{C}_{\text{bias}} + \frac{1}{\mathsf{C}_\lambda}\mathsf{C}_{\text{var}} \log \frac{2}{\delta}\right) \mathsf{L}^2 \lambda^{1+2r} \leqslant \mathsf{K}\, \log \frac{2}{\delta}\, \mathsf{L}^2 \lambda^{1+2r},$$

where we have set $\mathsf{K} = \left(\mathsf{C}_{\text{bias}} + \frac{1}{256}\mathsf{C}_{\text{var}}\right) \leqslant 8$ (see Proposition 7).
This result holds provided

$$0 < \lambda \leqslant \mathsf{B}_2^\star \wedge \lambda_0 \wedge \lambda_1,\; n \geqslant \triangle_1 \frac{\mathsf{B}_2^\star}{\lambda} \log \frac{8\square_1^2 \mathsf{B}_2^\star}{\lambda\delta}. \tag{47}$$

Indeed, we have shown in the previous point that since $\mathsf{C}_\lambda \geqslant \frac{\triangle_2}{4}$, $\mathsf{r}_\lambda(\theta^\star)^2 \lambda^{1/\alpha} \geqslant \triangle_2\, \mathsf{Q}^2 \frac{1}{n} \log \frac{2}{\delta}$.
3) Let us now work to guarantee the conditions in Eq. (47).
First, to guarantee $n \geqslant \triangle_1 \frac{\mathsf{B}_2^\star}{\lambda} \log \frac{8\square_1^2 \mathsf{B}_2^\star}{\lambda\delta}$, bound

$$\frac{\mathsf{B}_2^\star}{\lambda} = \frac{\mathsf{B}_2^\star \mathsf{L}^{2\beta} n^\beta}{\mathsf{C}_\lambda^\beta \mathsf{Q}^{2\beta}\, \log^\beta \frac{2}{\delta}} \leqslant \frac{2}{\mathsf{C}_\lambda^\beta} \frac{\mathsf{B}_2^\star \mathsf{L}^{2\beta}}{\mathsf{Q}^{2\beta}} n^\beta.$$

Then apply Lemma 15 with $a_1 = \frac{2\triangle_1}{\mathsf{C}_\lambda^\beta}$, $a_2 = \frac{16\square_1^2}{\mathsf{C}_\lambda^\beta}$, $A = \frac{\mathsf{B}_2^\star \mathsf{L}^{2\beta}}{\mathsf{Q}^{2\beta}}$. Since $\beta \geqslant 1/2$, using the bounds in Proposition 7, we find $a_1 \leqslant 648$ and $a_2 \leqslant 4$, hence the following sufficient condition:

$$n \geqslant \left(1296\frac{1}{1-\beta}A\log\left(5184\frac{1}{1-\beta}A^2\frac{1}{\delta}\right)\right)^{1/(1-\beta)}.$$

Then, to guarantee the condition

$$\lambda \leqslant \mathsf{B}_2^\star \wedge \lambda_0 \wedge \lambda_1,$$

we simply need

$$n \geqslant \frac{256\mathsf{Q}^2}{\mathsf{L}^2}\left(\mathsf{B}_2^\star \wedge \lambda_0 \wedge \lambda_1\right)^{-1/\beta}.$$

Hence, defining

$$N = \frac{256\mathsf{Q}^2}{\mathsf{L}^2}\left(\mathsf{B}_2^\star \wedge \lambda_0 \wedge \lambda_1\right)^{-1/\beta} \quad \vee \quad \left(1296\frac{1}{1-\beta}A\log\left(5184\frac{1}{1-\beta}A^2\frac{1}{\delta}\right)\right)^{1/(1-\beta)},$$

where $A = \frac{\mathsf{B}_2^\star \mathsf{L}^{2\beta}}{\mathsf{Q}^{2\beta}}$, we see that as soon as $n \geqslant N$, Eq. (47) holds.

We now state the following corollary, for $r > 0$. We define N in the following way:

$$N = \frac{256\mathsf{Q}^2}{\mathsf{L}^2}\left(\mathsf{B}_2^\star \wedge \lambda_0 \wedge \lambda_1\right)^{-1/\beta} \quad \vee \quad \left(1296\frac{1}{1-\beta}A\log\left(5184\frac{1}{1-\beta}A^2\frac{1}{\delta}\right)\right)^{1/(1-\beta)} \tag{48}$$

where $A = \frac{\mathsf{B}_2^\star \mathsf{L}^{2\beta}}{\mathsf{Q}^{2\beta}}$, $\lambda_0 = (2\mathsf{L}R\log\frac{2}{\delta})^{-1/r} \wedge 1$ and $\lambda_1 = \frac{\mathsf{Q}^{2\alpha}}{(\mathsf{Q}^\star)^{2\alpha}}$.

Corollary 4. *Assume $\rho \in \mathcal{P}_{\alpha,r}$ with $r > 0$. Let $\delta \in (0, 0.5]$ and $n \geqslant N$, where N is defined in Eq. (48). For*

$$\lambda = \left(256\left(\frac{\mathsf{Q}}{\mathsf{L}}\right)^2\frac{1}{n}\right)^\beta,$$

with probability at least $1 - 2\delta$,

$$L(\widehat{\theta}_\lambda^\star) - L(\theta^\star) \leqslant 8\,(256)^\gamma\,\left(\mathsf{Q}^\gamma\,\mathsf{L}^{1-\gamma}\right)^2\,\frac{1}{n^\gamma}\log\frac{2}{\delta},$$

Moreover, $N = O\left(\mathrm{poly}\left(\mathsf{B}_1^\star, \mathsf{B}_2^\star, \mathsf{L}, \mathsf{Q}, R, \log\frac{1}{\delta}\right)\right)$, which means that N is bounded by a rational function of the arguments of poly.

Proof.of Cor. 2 We simply apply Cor. 4 for $\alpha = 1$ and $\mathsf{Q} = \mathsf{B}_1^\star$.

G Additional lemmas

G.1 Self-concordance, sufficient conditions to define L and related quantities

In this section, we will consider an arbitrary probability measure μ on $\mathcal{Z}$. We assume that ℓ_z satisfies Assumption 8 with a certain given function φ. Recall that $R^\mu = \sup_{z\in\mathrm{supp}(\mu)}\sup_{g\in\varphi(z)}\|g\|$. In this section, we will also assume that $R^\mu < \infty$.

Lemma 10 (Gronwall lemma)**.** *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a differentiable function such that*

$$\forall t \in \mathbb{R},\ \varphi'(t) \leqslant C\varphi(t).$$

Then

$$\forall (t_0, t_1) \in \mathbb{R}^2,\ \varphi(t_1) \leqslant e^{C|t_1 - t_0|}\varphi(t_0).$$

Lemma 11. *Assume that there exists θ_0 such that* $\sup_{z\in \mathrm{supp}(\mu)} \mathrm{Tr}\left(\nabla^2 \ell_z(\theta_0)\right) < \infty$

- $\sup_{z\in \mathrm{supp}(\mu)} \mathrm{Tr}\left(\nabla^2 \ell_z(\theta)\right) < \infty$ *for any* $\theta \in \mathcal{H}$*;*
- *For any given radius* $T > 0$*, and any* $\|\theta_0\| \leqslant T$*, we have*

$$\forall \|\theta\| \leqslant T, \forall z \in \mathcal{Z},\ \ \mathrm{Tr}\left(\nabla^2 \ell_z(\theta)\right) \leqslant \exp\left(2R^\mu T\right) \mathrm{Tr}\left(\nabla^2 \ell_z(\theta_0)\right) < \infty.$$

Proof.
Let $z \in \mathrm{supp}(\mu)$ be fixed. Using the same reasoning as in the proof of Eq. (27), we can show

$$\forall \theta_0, \theta_1 \in \mathcal{H},\ \nabla^2 \ell_z(\theta_1) \preceq \exp\left(\sup_{g\in\varphi(z)} |g\cdot(\theta_1-\theta_0)|\right) \nabla^2 \ell_z(\theta_0) \preceq \exp\left(R^\mu \|\theta_1-\theta_0\|\right) \nabla^2 \ell_z(\theta_0)$$

Where we have used the fact that $R^\mu = \sup_{z\in\mathrm{supp}(\mu)} \sup_{g\in\varphi(z)} \|g\| < \infty$ Thus, in particular

$$\forall z \in \mathrm{supp}(\mu),\ \forall \theta_0, \theta_1 \in \mathcal{H},\ \ \mathrm{Tr}\left(\nabla^2 \ell_z(\theta_1)\right) \leqslant \exp\left(R^\mu \|\theta_1-\theta_0\|\right) \mathrm{Tr}\left(\nabla^2 \ell_z(\theta_0)\right),$$

which leads to the desired bounds.

Lemma 12. *Assume that there exists θ_0 such that*

$$\sup_{z\in\mathrm{supp}(\mu)} \mathrm{Tr}\left(\nabla^2 \ell_z(\theta_0)\right) < \infty, \qquad \sup_{z\in\mathrm{supp}(\mu)} \|\nabla \ell_z(\theta_0)\| < \infty.$$

Then

- $\sup_{z\in\mathrm{supp}(\mu)} \|\nabla \ell_z(\theta)\| < \infty$ *for any* $\theta \in \mathcal{H}$
- *For any* $T > 0$ *and any* $\|\theta_0\|, \|\theta\| \leqslant T, z \in \mathrm{supp}(\mu)$*,*

$$\begin{aligned}\|\nabla \ell_z(\theta)\| \leqslant \|\nabla \ell_z(\theta_0)\| + 2T\, \mathrm{Tr}\left(\nabla^2 \ell_z(\theta_0)\right) \\ + 4R^\mu\ \psi(2R^\mu T)\ \mathrm{Tr}\left(\nabla^2 \ell_z(\theta_0)\right)\ R^2.\end{aligned}$$

Proof.
Fix $z \in \mathcal{Z}, \theta_0, \theta_1 \in \mathcal{H}$ and $h \in \mathcal{H}$. Let us look at the function

$$f : t \in [0,1] \mapsto \left(\nabla \ell_z(\theta_t) - \nabla \ell_z(\theta_0) - t\nabla^2 \ell_z(\theta_0)(\theta_1-\theta_0)\right)\cdot h.$$

We have $f''(t) = \nabla^3 \ell_z(\theta_t)[\theta_1-\theta_0, \theta_1-\theta_0, h]$. By the self-concordant assumption, we have

$$\begin{aligned}\left|f''(t)\right| &\leqslant \sup_{g\in\varphi(z)} |g\cdot h| \nabla^2 \ell_z(\theta_t)[\theta_1-\theta_0, \theta_1-\theta_0] \\ &\leqslant \sup_{g\in\varphi(z)} |g\cdot h| \exp(t \sup_{g\in\varphi(z)} |g\cdot \theta_1-\theta_0|) \|\theta_1-\theta_0\|^2_{\nabla^2 \ell_z(\theta_0)}.\end{aligned}$$

Integrating this knowing $f'(0) = f(0) = 0$ yields

$$|f(1)| \leqslant \sup_{g\in\varphi(z)} |g\cdot h|\ \psi(\sup_{g\in\varphi(z)} |g\cdot(\theta_1-\theta_0)|) \|\theta_1-\theta_0\|^2_{\nabla^2 \ell_z(\theta_0)}.$$

Hence :

$$\|\nabla \ell_z(\theta_1)-\nabla \ell_z(\theta_0)\| \leqslant \|\nabla^2 \ell_z(\theta_0)\|\ \|\theta_1-\theta_0\| + \|\varphi(z)\|\ \psi(\sup_{g\in\varphi(z)} |g\cdot(\theta_1-\theta_0)|)\ \|\nabla^2 \ell_z(\theta_0)\|\ \|\theta_1-\theta_0\|^2$$

where $\psi(t) = (e^t - t - 1)/t^2$. Then, noting that $\|\nabla^2 \ell_z(\theta)\| \leqslant \mathrm{Tr}(\nabla^2 \ell_z(\theta))$, we have proved our lemma.

Lemma 13. *Assume that there exists θ_0 such that*

$$\sup_{z\in\mathrm{supp}(\mu)} \mathrm{Tr}\left(\nabla^2\ell_z(\theta_0)\right) < \infty, \qquad \sup_{z\in\mathrm{supp}(\mu)} \|\nabla\ell_z(\theta_0)\| < \infty, \qquad \sup_{z\in\mathrm{supp}(\mu)} |\ell_z(\theta_0)| < \infty.$$

Then

- *For any $\theta \in \mathcal{H}$, $\sup_{z\in\mathrm{supp}(\mu)} |\ell_z(\theta)| < \infty$*
- *For any $\theta_0 \in \mathcal{H}$, $T \geqslant \|\theta_0\|, \|\theta\| \leqslant T,\ z \in \mathrm{supp}(\mu)$, we have:*

$$|\ell_z(\theta)| \leqslant |\ell_z(\theta_0)| + 2\|\nabla\ell_z(\theta_0)\|T + \psi(2R^\mu\, T)\ \mathrm{Tr}(\nabla^2\ell_z(\theta_0))\, T^2.$$

Proof. Proceeding as in the proof of Eq. (30), we get

$$\forall z \in \mathcal{Z},\ \forall\theta_0,\theta_1 \in \mathcal{H},\ 0 \leqslant \ell_z(\theta_1)-\ell_z(\theta_0)-\nabla\ell_z(\theta_0)(\theta_1-\theta_0) \leqslant \psi(\sup_{g\in\varphi(z)} |g\cdot(\theta_1-\theta_0)|)\|\theta_1-\theta_0\|^2_{\nabla^2\ell_z(\theta_0)}$$

where $\psi(t) = (e^t - t - 1)/t^2$.

To conclude, we give the following result.

Proposition 8. *Let $\lambda \geqslant 0$. If a probability measure μ and ℓ satisfy Assumptions 3, 4 and 8, the function $L_{\mu,\lambda}(\theta) := \mathbb{E}_\mu\left[\ell_z(\theta)\right]+\lambda\|\theta\|^2$ and $\nabla L_{\mu,\lambda}(\theta), \nabla^2 L_{\mu,\theta}(\theta)$ are well-defined for any $\theta \in \mathcal{H}$, and we can differentiate under the expectation. Moreover,*

$$\forall\theta \in \mathcal{H}, \quad \sup_{z\in\mathrm{supp}(\rho)} |\ell_z(\theta)|, \quad \sup_{z\in\mathrm{supp}(\rho)} \|\nabla\ell_z(\theta)\|, \quad \sup_{z\in\mathrm{supp}(\rho)} \mathrm{Tr}\left(\nabla^2\ell_z(\theta)\right) < \infty.$$

Proof. We combine the results given in Lemmas 11 to 13.

G.2 Bernstein inequalities for operators

We start by proposing a slight modification of Proposition 6 in [28]. First we need to introduce the following quantitity and some notation for Hermitian operators. We denote by $\preceq$ is the partial order between positive semidefinite Hermitian operators. Let A, B be bounded Hermitian operators on $\mathcal{H}$,

$$A \preceq B \iff v\cdot(Av) \leqslant v\cdot(Bv),\ \forall v \in \mathcal{H} \iff B - A \text{ is positive semidefinite}.$$

Let q be a random positive semi-definite operator and let $\mathbf{Q} := \mathbb{E}\left[q\right]$, denote by $\mathcal{F}(\lambda)$ the function of λ defined as

$$\mathcal{F}(\lambda) := \operatorname{ess\,sup} \mathrm{Tr}\left(\mathbf{Q}_\lambda^{-1/2} q \mathbf{Q}_\lambda^{-1/2}\right),$$

where ess sup is the *essential support* of q.

Remark 3. *Note that if $\mathrm{Tr}(q) \leqslant c_0$, for a $c_0 > 0$ almost surely, then $\mathcal{F}(\lambda) \leqslant c_0/\lambda$. Vice versa, if $\mathcal{F}(\lambda_0) < \infty$ for a given $\lambda_0 > 0$, then $\mathrm{Tr}(q) \leqslant (\|Q\| + \lambda_0)\mathcal{F}(\lambda_0)$ almost surely, moreover $\mathcal{F}(\lambda) < \frac{\|\mathbf{Q}\|+\lambda_0}{\|\mathbf{Q}\|+\lambda}\mathcal{F}(\lambda_0)$ for any $\lambda > 0$.*

Proposition 9 (Prop. 6 of [28])**.** *Let $q_1, ..., q_n$ be identically distributed random positive semi-definite operators on a separable Hilbert space $\mathcal{H}$ such that the q are trace class and $\mathbf{Q} = \mathbb{E}\left[q\right]$. Let $\mathbf{Q}_n = \frac{1}{n}\sum_{i=1}^n q_i$ and take $0 < \lambda \leqslant \|\mathbf{Q}\|$ and assume $\mathcal{F}(\lambda) < \infty$. For any $\delta > 0$, the following holds with probability at least $1-\delta$:*

$$\|\mathbf{Q}_\lambda^{-1/2}\left(\mathbf{Q} - \mathbf{Q}_n\right)\mathbf{Q}_\lambda^{-1/2}\| \leqslant \frac{2\beta(1+\mathcal{F}_\infty(\lambda))}{3n} + \sqrt{\frac{2\beta\mathcal{F}_\infty(\lambda)}{n}}, \quad \beta = \log\frac{8\mathcal{F}_\infty(\lambda)}{\delta}$$

Proof. Use Proposition 3 of [28] and proceed as in the proof of Proposition 6 of [28] except that we bound $\mathrm{Tr}(\mathbf{Q}_\lambda^{-1}\mathbf{Q}) \leqslant \mathcal{F}_\infty(\lambda)$ instead of bounding $\mathrm{Tr}(\mathbf{Q}_\lambda^{-1}\mathbf{Q}) \leqslant \frac{\mathrm{Tr}(\mathbf{Q})}{\lambda}$, we find this result.

Here we slightly extend the results of Prop. 8 and Prop. 6 of [28], to extend the range of λ for which the result on the partial order between operators holds, from $0 < \lambda < \|\mathbf{Q}\|$ to $\lambda > 0$.

Proposition 10 (Prop. 8 together with Prop. 6 of [28])**.** *Let $q_1, ..., q_n$ be identically distributed random positive semi-definite operators on a separable Hilbert space $\mathcal{H}$ such that the q are trace class and $\mathbf{Q} = \mathbb{E}\,[q]$. Let $\mathbf{Q}_n = \frac{1}{n}\sum_{i=1}^n q_i$. Let any $\delta \in (0,1]$, $t > 0, 0 < \lambda \leqslant \|\mathbf{Q}\|$ and assume $\mathcal{F}(\lambda) < \infty$, when*

$$n \geqslant 8\mathcal{F}_\infty(\lambda) \log \frac{8\mathcal{F}_\infty(\lambda)}{\delta}\left(\frac{1}{4t^2} + \frac{1}{t}\right) \tag{49}$$

then the following holds with probability at least $1-\delta$:

$$\|\mathbf{Q}_\lambda^{-1/2}\left(\mathbf{Q} - \mathbf{Q}_n\right)\mathbf{Q}_\lambda^{-1/2}\| \leqslant t. \tag{50}$$

Moreover let $\lambda > 0, \delta \in (0,1]$ and Eq. (49) is satisfied for $t \leqslant 1/2$, then the following holds with probability at least $1-\delta$,

$$\mathbf{Q}_\lambda \preceq 2\mathbf{Q}_{n,\lambda}, \quad \iff \quad \|\mathbf{Q}_{n,\lambda}^{-1/2}\mathbf{Q}_\lambda^{1/2}\|^2 \leqslant 2. \tag{51}$$

Finally, let $\lambda > 0, \delta \in (0,1]$, Eq. (49) is satisfied for $t \leqslant 1/2$ and

$$n \geqslant 16\frac{c_0^2}{\|\mathbf{Q}\|^2}\log\frac{2}{\delta},$$

with $c_0 = ess \sup \mathrm{Tr}(q)$, then the following holds with probability at least $1-\delta$,

$$\mathbf{Q}_{n,\lambda} \preceq \frac{3}{2}\mathbf{Q}_\lambda, \quad \iff \quad \|\mathbf{Q}_{n,\lambda}^{1/2}\mathbf{Q}_\lambda^{-1/2}\|^2 \leqslant 3/2. \tag{52}$$

Proof.

Point 1) Let $\delta \in (0,1]$ and $0 < \lambda \leqslant \mathbf{Q}$. Using Proposition 9, we have that with probability at least $1-\delta$,

$$\|\mathbf{Q}_\lambda^{-1/2}\left(\mathbf{Q} - \mathbf{Q}_n\right)\mathbf{Q}_\lambda^{-1/2}\| \leqslant \frac{2\beta(1+\mathcal{F}_\infty(\lambda))}{3n} + \sqrt{\frac{2\beta\mathcal{F}_\infty(\lambda)}{n}}, \quad \beta = \log\frac{8\mathcal{F}_\infty(\lambda)}{\delta}.$$

Now note that if $\lambda \leqslant \|\mathbf{Q}\|$, we have

$$\frac{1}{2} \leqslant \frac{\|\mathbf{Q}\|}{\|\mathbf{Q}\| + \lambda} = \|\mathbf{Q}_\lambda^{-1}\mathbf{Q}\| \leqslant \mathrm{Tr}\left(\mathbf{Q}_\lambda^{-1}\mathbf{Q}\right) \leqslant \mathcal{F}_\infty(\lambda).$$

Thus we can bound $1 + \mathcal{F}_\infty(\lambda) \leqslant 3\mathcal{F}_\infty(\lambda)$, and we rewrite the previous bound

$$\|\mathbf{Q}_\lambda^{-1/2}\left(\mathbf{Q} - \mathbf{Q}_n\right)\mathbf{Q}_\lambda^{-1/2}\| \leqslant \frac{2\beta\mathcal{F}_\infty(\lambda)}{n} + \sqrt{\frac{2\beta\mathcal{F}_\infty(\lambda)}{n}}, \quad \beta = \log\frac{8\mathcal{F}_\infty(\lambda)}{\delta}.$$

Point 2) Now let $t > 0$, $\delta \in (0,1]$ and $0 < \lambda \leqslant \|\mathbf{Q}\|$. If

$$n \geqslant 8\mathcal{F}_\infty(\lambda)\beta\left(\frac{1}{4t^2} + \frac{1}{t}\right),$$

then

$$\|\mathbf{Q}_\lambda^{-1/2}\left(\mathbf{Q} - \mathbf{Q}_n\right)\mathbf{Q}_\lambda^{-1/2}\| \leqslant t.$$

Indeed, assume we want to find $n_0 > 0$ for which for all $n \geqslant n_0$, $\frac{A}{n} + \sqrt{\frac{B}{n}} \leqslant \frac{1}{2}$ where $A, B \geqslant 0$. setting $x = \sqrt{n}$, this is equivalent to finding x_0 such that $\forall x \geqslant x_0$, $\frac{x^2}{2} - \sqrt{B}x - A \geqslant 0$. A sufficient condition for this is that $x \geqslant \sqrt{B} + \sqrt{B + 2A}$. Thus, since $A, B \geqslant 0$, the condition $x \geqslant 2\sqrt{B + 2A}$ is sufficient, hence the condition $n \geqslant 4(B + 2A)$. Then we apply this to the following A and B to obtain the condition.

$$A = \frac{\beta \mathcal{F}_\infty(\lambda)}{t}, \; B = \frac{\beta \mathcal{F}_\infty(\lambda)}{2t^2}.$$

Point 3) When $\lambda > \|\mathbf{Q}\|$, the result is obtained noting that

$$\|\mathbf{Q}_\lambda^{1/2} \mathbf{Q}_{n,\lambda}^{-1/2}\|^2 \leqslant \frac{\|\mathbf{Q}\| + \lambda}{\lambda} = 1 + \frac{\|\mathbf{Q}\|}{\lambda} \leqslant 2.$$

When, on the other hand $0 < \lambda \leqslant \|\mathbf{Q}\|$, the final result is obtained by applying Prop. 6 and Prop. 8 of [28], or equivalently applying Eq. (50), with $t = 1/2$, for which the following holds with probability $1 - \delta$: $\|\mathbf{Q}_\lambda^{-1/2} (\mathbf{Q} - \mathbf{Q}_n) \mathbf{Q}_\lambda^{-1/2}\| \leqslant t$ and noting that,

$$\|\mathbf{Q}_\lambda^{1/2} \mathbf{Q}_{n,\lambda}^{-1/2}\|^2 \leqslant \frac{1}{1 - \|\mathbf{Q}_\lambda^{-1/2} (\mathbf{Q} - \mathbf{Q}_n) \mathbf{Q}_\lambda^{-1/2}\|} \leqslant 2.$$

To conclude this point, we recall that, given two Hermitian operators A, B and $t > 0$, the inequality $A \preceq tB$ is equivalent to $B^{-1/2}AB^{-1/2} \preceq tI$, when B is invertible. Since $B^{-1/2}AB^{-1/2}$ and tI are commutative, then $B^{-1/2}AB^{-1/2} \preceq tI$ is equivalent to $v \cdot (B^{-1/2}AB^{-1/2}v) \leqslant t\|v\|^2$ for any $v \in \mathcal{H}$, which in turn is equivalent to $\|B^{-1/2}AB^{-1/2}\| \leqslant t$. So

$$\|A^{1/2}B^{-1/2}\|^2 \leqslant t \iff A \preceq tB.$$

Point 4) First note that

$$\|\mathbf{Q}_\lambda^{-1/2} \mathbf{Q}_{n,\lambda}^{1/2}\|^2 \leqslant 1 + \|\mathbf{Q}_\lambda^{-1/2} (\mathbf{Q} - \mathbf{Q}_n) \mathbf{Q}_\lambda^{-1/2}\|. \tag{53}$$

When $0 < \lambda \leqslant \|\mathbf{Q}\|$, by applying Eq. (50) with $t = 1/2$, we have with probability $1 - \delta$: $\|\mathbf{Q}_\lambda^{-1/2} (\mathbf{Q} - \mathbf{Q}_n) \mathbf{Q}_\lambda^{-1/2}\| \leqslant t$, moreover by Eq. (53) we have

$$\|\mathbf{Q}_\lambda^{-1/2} \mathbf{Q}_{n,\lambda}^{1/2}\|^2 \leqslant 1 + t \leqslant 3/2.$$

When instead $\lambda > \|\mathbf{Q}\|$, we consider the following decomposition

$$\|\mathbf{Q}_\lambda^{-1/2} (\mathbf{Q} - \mathbf{Q}_n) \mathbf{Q}_\lambda^{-1/2}\| \leqslant \frac{1}{\lambda}\|\mathbf{Q} - \mathbf{Q}_n\| \leqslant \frac{1}{\lambda}\|\mathbf{Q} - \mathbf{Q}_n\|_{HS},$$

where we denote by $\|\cdot\|_{HS}$, the Hilbert-Schmidt norm (i.e. $\|A\|_{HS}^2 = \operatorname{Tr}(A^*A)$) and $\|\mathbf{Q} - \mathbf{Q}_n\|_{HS}$ is well defined since both $\mathbf{Q}, \mathbf{Q}_n$ are trace class. Now since the space of Hilbert-Schmidt operators on a separable Hilbert space is itself a separable Hilbert space and q are bounded almost surely by $c_0 := \operatorname{ess\,sup} \operatorname{Tr}(q)$, we can concentrate $\|\mathbf{Q} - \mathbf{Q}_n\|_{HS}$ via Bernstein inequality for random vectors [e.g. Thm. 3.3.4 of 42], obtaining with probability at least $1 - \delta$

$$\|\mathbf{Q} - \mathbf{Q}_n\|_{HS} \leqslant \frac{2c_0 \log \frac{2}{\delta}}{n} + \sqrt{\frac{2c_0^2 \log \frac{2}{\delta}}{n}} \leqslant \|\mathbf{Q}\|/2,$$

where the last step is due to the fact that we require $n \geqslant 16c_0^2(\log \frac{2}{\delta})/\|\mathbf{Q}\|^2$, and the fact that by construction $\|\mathbf{Q}\| \leqslant B$. Then,

$$\|\mathbf{Q}_\lambda^{-1/2} \mathbf{Q}_{n,\lambda}^{1/2}\|^2 \leqslant 1 + \frac{\|\mathbf{Q}\|}{2\lambda} \leqslant 3/2.$$

The final result on $\preceq$ is obtained as for Point 5.

Remark 4. *Let* $\mathrm{Tr}(q) \leqslant c_0$ *almost surely, for a* $c_0 > 0$*. Then* $\mathcal{F}(\lambda) \leqslant c_0/\lambda$*. So Eq.* (49) *is satisfied when*

$$n \geqslant \frac{8c_0}{\lambda} \log \frac{8c_0}{\lambda\,\delta} \left(\frac{1}{4t^2} + \frac{1}{t} \right),$$

since $\mathcal{F}(\lambda) \leqslant c_0/\lambda$ *as observed in Remark 3. In particular, when* $t = 1/2$*, Eq.* (49) *is satisfied when*

$$n \geqslant \frac{24c_0}{\lambda} \log \frac{8c_0}{\lambda\,\delta}.$$

G.3 Sufficient conditions to bound n in order to guarantee $n \geqslant C_1 n^p \log \frac{C_2 n^p}{\delta}$

Lemma 14. *Let* $a_1, a_2, A \geqslant 0$ *and* $\delta > 0$*. If*

$$n \geqslant 4a_1^2 A^2 \log^2 \left(\frac{2a_1 a_2 A^2}{\delta} \right),$$

then $n \geqslant a_1 A n^{1/2} \log \frac{a_2 A n^{1/2}}{\delta}$.

Proof. Indeed, note that

$$n \geqslant a_1 A n^{1/2} \log \frac{a_2 A n^{1/2}}{\delta} \iff \frac{a_1 A}{n^{1/2}} \log \frac{a_2 A n^{1/2}}{\delta} \leqslant 1.$$

Now use the fact that for $A, B \geqslant 0$, $k \geqslant 2A\log(2AB)$ implies $\frac{A}{k}\log(Bk) \leqslant 1$. Indeed, $\log(Bk) = \log(2AB) + \log\frac{Bk}{2AB} = \log(2AB) + \log\frac{k}{2A} \leqslant \log(2AB) + \frac{k}{2A}$. Hence, multiplying by $\frac{A}{k}$, we get the result.

We apply this to $A = a_1 A$, $B = \frac{a_2 A}{\delta}$ and $k = n^{1/2}$ to get the bound.

Lemma 15. *Let* $a_1, a_2, A \geqslant 0$ *and* $\delta > 0$*. Let* $p \in [\frac{1}{2}, 1)$*. If*

$$n^{1-p} \geqslant 2\frac{1}{1-p} a_1 A \log \left(2a_1 (a_2 \vee 1) \frac{1}{1-p} A^2 \frac{1}{\delta} \right),$$

then

$$n \geqslant a_1 A n^p \log \frac{a_2 A n^p}{\delta}.$$

Proof.1) Let $C_1, C_2 \geqslant 0$, and $p \in [0, 1)$. Then

$$n \geqslant C_1 n^p \log(C_2 n^p) \iff \frac{C_1 \frac{p}{1-p}}{n^{1-p}} \log \left(C_2^{(1-p)/p} n^{1-p} \right) \leqslant 1.$$

Now use the fact that for $A, B \geqslant 0$, $k \geqslant 2A\log(2AB)$ implies $\frac{A}{k}\log(Bk) \leqslant 1$ (see proof of Lemma 14).

Thus, $n^{1-p} \geqslant 2C_1 \frac{p}{1-p} \log \left(2C_1 \frac{p}{1-p} C_2^{(1-p)/p} \right)$ is a sufficient condition.

2) Now taking $C_1 = a_1 A$ and $C_2 = \frac{a_2 A}{\delta}$, we find that

$$n^{1-p} \geqslant 2\frac{p}{1-p} a_1 A \log \left(2a_1 a_2^{(1-p)/p} \frac{p}{1-p} A^{1/p} (\frac{1}{\delta})^{(1-p)/p} \right).$$

Since $0.5 \leqslant p \leqslant 1$, we see that $\frac{1-p}{p} \leqslant 1$ and $\frac{1}{p} \leqslant 2$ and thus we get our final sufficient condition.

$$n^{1-p} \geqslant 2\frac{1}{1-p} a_1 A \log \left(2a_1 (a_2 \vee 1) \frac{1}{1-p} A^2 \frac{1}{\delta} \right).$$