Beyond Tikhonov: Faster Learning with Self-Concordant Losses via Iterative Regularization

Gaspard Beugnot
Inria*
gaspard.beugnot@inria.fr

Julien Mairal
Inria†
julien.mairal@inria.fr

Alessandro Rudi
Inria*
alessandro.rudi@inria.fr

Abstract

The theory of spectral filtering is a remarkable tool to understand the statistical properties of learning with kernels. For least squares, it allows to derive various regularization schemes that yield faster convergence rates of the excess risk than with Tikhonov regularization. This is typically achieved by leveraging classical assumptions called source and capacity conditions, which characterize the difficulty of the learning task. In order to understand estimators derived from other loss functions, Marteau-Ferey et al. [1] have extended the theory of Tikhonov regularization to generalized self concordant loss functions (GSC), which contain, *e.g.*, the logistic loss. In this paper, we go a step further and show that fast and optimal rates can be achieved for GSC by using the iterated Tikhonov regularization scheme, which is intrinsically related to the proximal point method in optimization, and overcomes the limitation of the classical Tikhonov regularization.

1 Introduction

We consider the problem of supervised learning where we want to find a prediction function θ mapping an input point x living in a set $\mathcal{X}$ to a label y in $\mathcal{Y}$. In this paper, we assume that θ lives in a separable Hilbert space $\mathcal{H}$ and is learned from a set of observations $(x_i, y_i)_{i=1,\dots,n}$ that are i.i.d. samples drawn from an unknown probability distribution ρ on $\mathcal{X} \times \mathcal{Y}$. The goal is to find θ that minimizes the expected risk L, which is defined below along with the empirical risk $\hat{L}$:

$$L(\theta) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, \theta(x)) \mathrm{d}\rho(x, y), \qquad \widehat{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \theta(x_i)), \tag{1}$$

where ℓ is a suitable loss function comparing true labels with predictions. This paper aims for upper bounds on the excess risk for a specific estimator $\widehat{\theta}$. That is, we assume that the minimum of the expected risk is attained for some $\theta^\star$ in $\mathcal{H}$, and we want to derive *probabilistic upper bounds on the excess risk*:

$$\mathbb{P}\left[L(\widehat{\theta}) - L(\theta^\star) > C_1 n^{-\gamma} \log \tfrac{2}{\delta}\right] \leq \delta, \tag{2}$$

given some value δ in $(0, 1)$, where C_1 is a positive constant, and $\widehat{\theta}$ is an estimator built from the n observations. The quantity $O(n^{-\gamma})$ denotes the rate of convergence of the estimator $\hat{\theta}$. A classical "slow" rate with $\gamma = 1/2$ is typically achieved by many estimators and is in fact optimal if only mild assumptions are made about the data distribution ρ. Even though optimal, this rate is nevertheless a worst case and faster rates with $\gamma > 1/2$ can be achieved both in theory and in practice, by making additional assumptions about the difficulty of the learning task. Originally introduced in the

*Inria, École normale supérieure, CNRS, PSL Research University, 75005 Paris, France
†Inria, Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

literature of inverse problems, the so-called *source* and *capacity* conditions have been shown to be appropriate for this purpose, leading to statistical analysis with fast rates of convergence [1, 2, 3]. The optimality of results of the form (2) is characterized by comparing them with lower bounds that are available for various sets of data distributions ρ [3]. Matching upper bounds with lower bounds ensures that the estimator $\widehat{\theta}$ is *optimal*, in the sense that no information is lost in the process of exploiting the data samples to compute $\widehat{\theta}$, for the given set of distributions.

In this search for optimal estimators, most of the attention has been devoted to minimizers of some function of the empirical risk $\widehat{L}$, which is defined in (Eq. (1)). Then, the key challenge is to *regularize* $\widehat{L}$ in order to achieve better generalization properties. The most widely used scheme is probably Tikhonov regularization; other examples when $\mathcal{H}$ is a RKHS include truncated regression [4], or early stopping in gradient descent algorithms [5, 6]. When the loss ℓ is set to least squares, it can be shown that minimizing the excess risk amounts to solving an ill-posed inverse problem [7], which led to the remarkable theory of *spectral filtering*. A large class of regularization schemes can indeed be seen as a filtering process applied to the training labels y_i after regularizing the spectrum of the kernel matrix [2, 8]. Interestingly, this theory has highlighted the fact that not all regularization schemes are equal: some of them obtain fast learning rates in (2) on "easy" problem (a thorough definition is given in Section 2) while others cannot leverage this additional regularity to improve the learning rate.

Such a general analysis for least squares is made possible by the fact that a closed-form expression of the estimator is available. When considering different loss function ℓ, the estimator $\widehat{\theta}$ is unfortunately only implicitly available as the solution of an optimization problem involving $\widehat{L}$. A step to extend least squares results to more general loss functions has been achieved by Marteau-Ferey et al. [1], who provide bounds on the form (2) for Tikhonov estimator on generalized self concordant (GSC) functions. GSC functions are three-times-differentiable functions whose third derivative is bounded by the second-derivative. In practice, they were introduced to conduct a general analysis of the Newton method in optimization [9, 10], and adapted in [11] to encompass a larger class of loss function. It includes notably the logistic regression loss, which is widely used for classification.

While Tikhonov yields fast rates of convergence in several data regimes, it is known to be unable to adapt to the whole range of learning task difficulties. More precisely, it suffers from a "saturation" effect [2], meaning that when the learning task becomes simpler, the learning rate stops improving and is suboptimal. Our paper addresses this limitation for GSC functions by considering instead the iterated Tikhonov regularization (IT) scheme. In the context of least squares, this approach consists of successively fitting the residuals. For more general loss functions, it is equivalent to performing a few steps of the proximal point method in optimization [12]. Our main result is a probabilistic upper bound on the excess risk, which is optimal given usual source and capacity conditions assumptions on the learning task, thus addressing the limitations of the classical Tikhonov regularization.

2 Background and Preliminaries

2.1 Definitions: Estimator and Loss Function

Let $\mathcal{X}$ be a Borel input space, $\mathcal{Y}$ be a vector-valued output spaces, and ρ a probability distribution on $\mathcal{X} \times \mathcal{Y}$. We consider $\mathcal{H}$ to be a separable Hilbert space of functions from $\mathcal{X}$ to $\mathcal{Y}$. Given a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, we aim at minimizing the expected loss, while we only have access to the empirical loss – both are defined in Eq. (1). Our work provides an upper bound on the excess risk of the iterated Tikhonov estimator. For the basic case of least squares with $\mathcal{Y} = \mathbb{R}$, it is usually defined as a procedure that refits the residuals, see, e.g., §5.4 in [2]. Starting with $\widehat{\theta}^0_\lambda = 0$, it consists of the sequence

$$\widehat{\theta}^t_\lambda = \widehat{\theta}^{t-1}_\lambda + \operatorname*{arg\,min}_{\theta \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \left(y_i - \widehat{\theta}^{t-1}_\lambda(x_i) - \theta(x_i) \right)^2 + \frac{\lambda}{2} \|\theta\|^2 \right\}. \tag{3}$$

To extend this regularization to other loss function, we make the change of variable $\theta' = \widehat{\theta}^{t-1}_\lambda + \theta$ in the equation above, which yields the proximal point algorithm [12].

Definition 1 (Iterated Tikhonov estimator a.k.a. proximal point algorithm). *We define the iterated Tikhonov estimator with the following sequence. Given $\lambda > 0$ and $\widehat{\theta}_\lambda^0 = 0$,*

$$\widehat{\theta}_\lambda^{t+1} = \mathsf{prox}_{\widehat{L}/\lambda}(\widehat{\theta}_\lambda^t) \stackrel{\text{def.}}{=} \operatorname*{arg\,min}_{\theta \in \mathcal{H}} \left\{ \widehat{L}(\theta) + \frac{\lambda}{2} \left\| \theta - \widehat{\theta}_\lambda^t \right\|^2 \right\}, \tag{4}$$

where $\mathsf{prox}_{\widehat{L}/\lambda}$ denotes the proximal operator of the empirical risk $\widehat{L}$ rescaled by $1/\lambda$.

Remark 1. In practice, the proximal operator is only computed approximately by using an optimization algorithm. Nevertheless, the benefits in terms of statistical accuracy of the iterated Tikhonov scheme are robust to inexact solutions, as long as the accuracy for solving the sub-problems (Eq. (4)) is high enough. We discuss this point in Section 3.2.

Remark 2. It is easy to show that the sequence of the proximal point algorithm always converges to a minimizer of the unregularized empirical risk, which is of course not what we are interested in. Instead, we consider and analyze the procedure with a fixed small number of steps t and show later that optimal learning rates can be obtained by choosing an appropriate parameter λ.

Remark 3. When the loss is a function of a residual $y - \theta(x)$—assuming $\mathcal{Y}$ to be a vector space—as in the least square case, we recover the classical definition consisting of refitting the residual, and with $t = 1$, we recover Tikhonov.

Interestingly, our definition makes the estimator compatible with other loss functions, such as the logistic loss. More precisely, the main assumption we make on the loss is to be *generalized self concordant*. We follow the definition of [1], which is a special case of 2-self concordance introduced in [13]:

Definition 2 (Generalized self-concordance). *For any $z = x, y \in \mathcal{X} \times \mathcal{Y}$, the function $\ell_z : \mathcal{H} \to \mathbb{R}$ defined as $\ell_z(\theta) = \ell(y, \theta(x))$ is convex and three times differentiable. Besides, there exists a set $\phi(z) \subseteq \mathcal{H}$ s.t:*

$$\forall \theta, h, k \in \mathcal{H}, \quad \left| \nabla^3 \ell_z(\theta)\,[h,k,k] \right| \leq \sup_{g \in \phi(z)} |k \cdot g|\, \nabla^2 \ell_z(\theta)\,[k,k]\,. \tag{5}$$

The brackets indicate that the vectors h, k and k are applied to the 3-dimensional tensor $\nabla^3 \ell_z(\theta)$. The definition seems technical at first sight, but intuitively, this assumption allows to upper bound the deviation between the objective function and its local quadratic approximation. This enables a simple analysis of the Newton method for optimization, making it easy to quantify the basin of quadratic convergence [14]. On top of this, it has the benefit of encompassing a large class of loss functions, such as the logistic loss: see Example 1 in [1] for values of $\phi(z)$ with usual losses. We provide some intuition on GSC loss functions in Remark 6 in Appendix C.1.

In order to ensure the existence of the loss and its derivatives everywhere, we also need the following technical assumptions also introduced in [1], which are reasonable in practice. This ensures that both L and $\widehat{L}$ are generalized self concordant too.

Assumption 1 (Technical assumptions). *There exists R s.t $\sup_{g \in \phi(z)} \|g\| \leq R$ almost surely for z drawn from the distribution ρ and $|\ell_z(0)|\,, \|\nabla \ell_z(0)\|\,, \operatorname{Tr} \nabla^2 \ell_z(0)$ are almost surely bounded.*

The following assumption is usual in excess risk analysis [1, 15]. In our proof strategy, all the quantities are vectors and operators in $\mathcal{H}$, which makes the analysis simpler. Weakening this assumption (*e.g.* assuming that $\theta^\star \in \mathcal{L}_2(\mathcal{X})$) would require finding an equivalent of the covariance operator for GSC loss function, which constitute an interesting future direction.

Assumption 2 (Existence of a minimizer). *There exists $\theta^\star$ in $\mathcal{H}$ s.t $L(\theta^\star) = \inf_{\theta \in \mathcal{H}} L(\theta)$.*

Finally, following [1] we also define the *expected Hessian* and the *regularized expected Hessian* as

$$\forall \theta \in \mathcal{H},\ \lambda > 0, \quad \mathbf{H}(\theta) = \mathbb{E}_{z \sim \rho}\left[\nabla^2 \ell_z(\theta) \right], \quad \mathbf{H}_\lambda(\theta) = \mathbf{H}(\theta) + \lambda \mathbf{I},$$

and we introduce the degrees of freedom, also known as the effective dimension of the problem:

Definition 3 (Degrees of freedom). *The degrees of freedom is defined as:*

$$\forall \lambda, \quad \mathsf{df}_\lambda = \mathbb{E}_{z \sim \rho}\left[\|\nabla \ell_z(\theta^\star)\|^2_{\mathbf{H}_\lambda^{-1}(\theta^\star)} \right].$$

where we denote by $\|\theta\|_A = \left\| A^{1/2} \theta \right\|$, with $\theta \in \mathcal{H}$, the norm induced by a positive definite operator A on $\mathcal{H}$.

Remark 4. The intuition about this definition is not straightforward. To better understand why this quantity is a key to characterize the amount of regularization in a learning problem, it is useful to consider the specific case of the square loss with kernels. In such a case, $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) and $\theta(x) = \theta^\top \Phi(x)$, where $\Phi : \mathcal{X} \to \mathcal{H}$ is the kernel mapping. Then, the Hessian is constant everywhere and equal to the *covariance operator* $T = \mathbb{E}_{x\sim\rho_x}\left[\Phi(x) \otimes \Phi(x)\right]$ where ρ_x is the marginal of ρ. Consequently, the degrees of freedom (also known as *effective dimension*) is a spectral function of T which may be written as $\mathsf{df}_\lambda = \operatorname{Tr} T T_\lambda^{-1}$. This is the classical quantity which appears on the bias/variance decomposition of the excess risk, with a variance part decaying in $\operatorname{Tr} T T_\lambda^{-1}/n$, see [3].

2.2 Source and Capacity Conditions

We now introduce the hypotheses we make on the learning task, which will allow us to derive fast rates of convergence. They measure the difficulty of the problem and are classical in the context of learning with kernels, see *e.g.* [8, 16, 17]. It is indeed established that given an algorithm which outputs an estimator $\widehat{\theta}$, one can find a probability measure ρ s.t the learning rate of the estimator is arbitrarily low, a result known as the "no-free lunch theorem" [18]. Inspired by the literature of inverse problems, two assumptions were introduced to restrict the space of considered distributions.

Assumption 3 (Source condition). *There exists $r > 0$ and v in $\mathcal{H}$ s. t: $\theta^\star = \mathbf{H}^r(\theta^\star)v$.*

$A \mapsto A^r$ is the usual power for positive definite operators. The source condition should be seen as a smoothness assumption on $\theta^\star$, and for least square, we recover the usual definition of the source condition, that is $\theta^\star = T^r v$, with T the covariance operator we previously defined. Bigger r implies that the optimum can be well approximated by a few eigenvectors. Assuming $r = 0$ simplifies to $\theta^\star \in \mathcal{H}$.

The second assumption characterizes the ill-posedness of the problem:

Assumption 4 (Capacity condition). *There exists $\alpha > 1$, $\mathsf{s}, \mathsf{S} > 0$ s.t $\mathsf{s}\lambda^{-1/\alpha} \leq \mathsf{df}_\lambda \leq \mathsf{S}\lambda^{-1/\alpha}$.*

Again, for the square loss, it turns to a bound on the eigenvalue decay of the covariance operator. If σ_j, e_j is an eigenbasis of T, then $\sigma_j = O(j^{-\alpha})$. Said differently, the bigger α, the fewer directions are needed to approximate well a sample $x \sim \rho_x$ in expectation, and the easier is the learning task. This is an assumption on the input space $\mathcal{X}$ and does not imply anything on the labels $\mathcal{Y}$.

2.3 Previous Results

Our main result considers iterated Tikhonov *with* GSC loss functions. While iterated Tikhonov has been previously analyzed for squared loss by leveraging the theory of spectral filtering (see below), extensions to other loss functions raise several difficulties, which will be detailed in Section 3.

Spectral filters and least squares. As we mentioned earlier, the key insight on regularization with the square loss is that a closed-form expression of the estimator is available. By using the same notation as in Remark 4, the kernel ridge regression estimator can be for instance written

$$\widehat{\theta}_\lambda = \sum_{i=1}^n \beta_i \Phi(x_i) \quad \text{with} \quad \beta = \frac{1}{n} g_\lambda\left(\frac{K}{n}\right) y, \tag{6}$$

where K is the $n \times n$ kernel matrix, $y = (y_i)_{1\leq i\leq n}$ is the vector of training labels and $g_\lambda(K/n) = (K/n + \lambda I)^{-1}$. Note that g_λ is a function acting on the spectrum of K, which makes it a special case of regularization by *spectral filtering*, which may be analyzed for more general functions g_λ. In particular, a key quantity for understanding the regularization effect of a filter g_λ is the so-called *qualification*. Following [2, 8], this quantity is defined below.

Definition 4 (Qualification of a spectral filter). *For any $\lambda > 0$, define $g_\lambda : [0, 1] \to \mathbb{R}$ a filter function. Its qualification is the highest q such that*

$$\forall \nu \leq q, \quad \sup_\sigma \left|1 - \sigma g_\lambda(\sigma)\right| \sigma^\nu \leq \omega_\nu \lambda^\nu, \tag{7}$$

with ω_ν a constant independant of λ.

Under the source and capacity conditions, it is possible to show that the resulting estimator would enjoy an optimal rate in $n^{-\frac{\alpha(1+2r)}{1+\alpha(1+2r)}}$ if $r + 1/2 \leq q$ (where r comes from the source condition). When $r + 1/2 > q$, the rate is instead of order $n^{-\frac{\alpha(1+2q)}{1+\alpha(1+2q)}}$, which is suboptimal, see *e.g.* Thm. 3.4 [17] (set the parameter s to $1/2$). This illustrates the *saturation effect* of some regularization schemes. For example, Tikhonov regularization amounts to filtering with $g_\lambda : \sigma \mapsto (\sigma + \lambda)^{-1}$ and has *qualification* 1, so the parameter r *saturates* at $r = 1/2$. Thus, even if $r \gg 1/2$, the excess risk of $\widehat{\theta}_\lambda$ will decay in $n^{-\frac{\alpha}{1+\alpha}}$, which is suboptimal. Designing estimators with high qualification is key to obtaining fast rates that can adapt to both hard and easy learning tasks.

Iterated Tikhonov with the Square Loss. We can compute the spectral filter function g_λ^t corresponding to t iterations of IT, which yields

$$g_\lambda^t : \sigma \mapsto (\sigma + \lambda)^{-1} \sum_{i=0}^{t-1} \left(\frac{\lambda}{\sigma + \lambda}\right)^i = \sigma^{-1} \left(1 - \left(\frac{\lambda}{\sigma + \lambda}\right)^t\right). \tag{8}$$

Choosing a fixed t and computing the supremum of $\sigma \mapsto |1 - \sigma g_\lambda(\sigma)|\, \sigma^\nu$, we find that IT estimator has qualification t, which is thus better than Tikhonov. IT has been thoroughly studied in the community of inverse problems, dating back to the work of [19]. It was naturally transferred to learning with kernels thanks to the aforementioned connection with inverse problems.

The link we make with the proximal point algorithm has never been studied from a statistical perspective, to the best of our knowledge, even though it has attracted a lot of attention in the optimization literature, notably with accelerated algorithms [20, 21], or variants of the proximal operator on a class of self-concordant loss functions [22]. More attention was devoted to *boosting*, where the penalty λ is fixed but the number of iterations t may go to infinity, necessitating an appropriate stopping rule [23]. Nevertheless, such a work focuses on the least square loss, where the theory of spectral filter can be applied. Finally, the proximal sequence in Eq. (4) can be cast as a constrained optimization problem related to sequential greedy approximation [24].

Tikhonov and Generalized Self Concordant losses. Extending the results obtained with the square loss to more general losses is challenging since there is no closed form available for the resulting estimator, and the theory of spectral filtering does not apply. Nevertheless, the case of Tikhonov regularization for GSC loss functions was treated in [1]. It is shown that the resulting estimator enjoys optimal rate as long as $r \leq 1/2$, meaning that the saturation of Tikhonov regularization is recovered in those settings. We will extend these results to the IT regularization, showing that an improved qualification can be achieved, leading to fast rates for a larger class of learning tasks.

3 Main Result

Our main result establishes an optimal non-asymptotic bias variance decomposition of the excess risk. It is optimal in the sense that choosing an appropriate regularization parameter λ enables to achieve the optimal lower rates of convergence established for least squares.

Theorem 1 (Optimal rates of IT estimator). *Let $\delta \in (0, 1]$, and set $\lambda \in (0, \mathsf{L}_0)$, $n \geq \mathsf{N}$. The following bound on the excess risk holds with probability greater than $1 - 2\delta$:*

$$L(\widehat{\theta}_\lambda^t) - L(\theta^\star) \leq \mathsf{C}_{\text{bias}} \lambda^{2s} + \mathsf{C}_{\text{var}} \frac{\mathsf{df}_\lambda}{n}, \text{ with } s = \min\{r + 1/2, t\}. \tag{9}$$

If we further assume that the capacity condition holds and that the estimator does not saturate, that is $t \geq r + 1/2$, then setting

$$\lambda = \mathsf{C}_{\text{risk}}\, n^{-\frac{\alpha}{1+\alpha(2r+1)}}, \tag{10}$$

makes the following holds with probability greater than $1 - 2\delta$:

$$L(\widehat{\theta}_\lambda^t) - L(\theta^\star) \leq 2\mathsf{C}_{\text{risk}}\, n^{-\frac{\alpha(2r+1)}{1+\alpha(2r+1)}}. \tag{11}$$

The constants $\mathsf{L}_0, \mathsf{N}, \mathsf{C}_{\text{bias}}, \mathsf{C}_{\text{var}}, \mathsf{C}_{\text{risk}}$ are detailed in Theorem 4 in the appendix; they are explicit and depend only on $r, \alpha, \mathsf{S}, R, t, \delta$ and the distribution ρ.

Optimal rates. First, we note that the decay rate of the excess risk is optimal provided $t \geq r+1/2$. It means that, up to constant factors, no estimators trained on n observations can benefit from a better learning rate (in the worse case sense) with the prior considered on ρ, that is source and capacity conditions of parameters r, α. This leads to the second point: we see that IT has qualification $q = t$. When $t = 1$, this is Tikhonov estimator and we recover the result of [1]. This qualification shows in the bound on the bias: if $r \leq t - 1/2$, the bias is optimal in λ^{2r+1}; otherwise, it is suboptimal and decays only in λ^{2t}, which leads to higher excess risk, hence generalization error.

Influence of t. The leading multiplicative constant of the rate $\mathsf{C}_{\mathrm{var}}$ in Eq. (9) depends linearly on the number of steps t, as shown in Eq. (55) in Appendix B.5. Thus, the rate in Eq. (11) is optimal in n when $t = O(r)$. Letting t go to infinity amounts to minimizing the empirical risk, which yield the unregularized estimator: this agrees with our bound on the excess risk, as the constant $\mathsf{C}_{\mathrm{var}}$ would go to infinity in that case.

Source and capacity condition. The source and capacity conditions enable precise bounds on the bias and the variance, respectively. If they do not hold, the bias can only be bounded by $O(\lambda)$, while we can upper bound the degrees of freedom with $O(1/\lambda)$, leading to slow learning rates. If the source condition holds but the capacity condition does not, we then obtain learning rates in $n^{-2s/(2s+1)}, s = \min\{r + 1/2, t\}$, which are also optimal in these settings.

Example: a very easy learning task. Suppose the source condition satisfies $r = 10$ and that the capacity condition does not hold. Then, using Tikhonov estimator [1] amounts to setting $t = 1$. The generalization error would then decay as $n^{-2/3}$. On the other hand, using Iterated Tikhonov estimator with $t = 10$ would make the generalization error decay in $n^{-20/21}$, which is much better.

3.1 Sketch of the proof

The proof, which is fully detailed in the appendix, has the following outline:

- First, we give technical results on generalized self concordant functions;
- Then, we define the intermediate quantity in our bias-variance decomposition;
- Finally, we proceed to bounding the bias and the variance separately, which plugged together give our bound on the excess risk.

To prove the theorem above we build upon the tools from [1] on generalized self concordant functions. The resulting proof covers and simplifies the case of Tikhonov regularization (one step of iterated Tikhonov) and generalizes the rates to $r > 1/2$. We provide also a fine control of the constants, that takes into account the sequential nature of the IT estimator.

Properties of generalized self concordant loss functions Here, we report key properties of GSC loss functions, which are covered in depth in Appendix A. GSC loss functions are convenient to study as they come with a set of bounds on the Hessian, the gradients and the function values. Intuitively, by integrating multiple times the relation between the third and second derivative in the definition from Eq. (5), one can obtain bounds on function values. To introduce them, we first define the following function:

$$\forall \theta \in \mathcal{H}, \quad \mathsf{t}(\theta) = \sup_{z \in \mathsf{Supp}\, \rho} \sup_{g \in \phi(z)} |g \cdot \theta| \,. \tag{12}$$

By integrating three times the bound of the definition, one can show that:

$$L(\widehat{\theta}^t_\lambda) - L(\theta^\star) \leq \Psi\left(\mathsf{t}(\widehat{\theta}^t_\lambda - \theta^\star)\right) \left\|\widehat{\theta}^t_\lambda - \theta^\star\right\|^2_{\mathbf{H}(\theta^\star)}, \quad \Psi : t \mapsto (e^t - t - 1)/t^2. \tag{13}$$

This type of bound first appeared in [11] and was given in this form in [1]. We report it in Proposition 3 in the appendix. For instance, when ℓ is the square loss, $\mathsf{t} = 0$ everywhere and the r.h.s turns to $1/2\|\widehat{\theta}^t_\lambda - \theta^\star\|^2_T$, see [17, 25]. On top of this, we generalize a lower bound on the gradient:

Lemma 1 (Stacking operator on gradient bounds). *Let $\theta, \nu, \xi \in \mathcal{H}$, $\lambda > 0$. If $A : \mathcal{H} \to \mathcal{H}$ commutes with $\mathbf{H}(\xi)$, the following holds:*

$$e^{-\mathsf{t}(\theta-\xi)} \underline{\phi}\left(\mathsf{t}(\nu - \theta)\right) \|A(\nu - \theta)\|_{\mathbf{H}_\lambda(\xi)} \leq \|A(\nabla L_\lambda(\nu) - \nabla L_\lambda(\theta))\|_{\mathbf{H}^{-1}_\lambda(\xi)}, \tag{14}$$

where $\underline{\phi} : t \mapsto (1 - e^{-t})/t$.

Together with Eq. (13), this result is the workhorse of our proof for the upper bound on the excess risk. It is detailed and proven in Appendix D.

Bias-variance decomposition. Thanks to Eq. (13), we can relate the excess risk with the distance between estimates. This is why bounding the excess risk amounts to finding a good bias-variance decomposition. Most of the proof we find for the square loss rely on the quantity

$$\vartheta_\lambda^t = g_\lambda^t(\hat{T})\hat{T}\theta^\star, \tag{15}$$

with $\hat{T} = 1/n \sum_i \Phi(x_i) \otimes \Phi(x_i)$ the empirical covariance operator, obtained by replacing ρ with the empirical distribution in Remark 4. This is basically the estimator trained on *noiseless empirical data* (*i.e.* using $\theta^\star(x_i)$ instead of y_i) [17, 26, 23]. Unfortunately, working with GSC function makes the spectral filtering point of view inapplicable. We need to translate a closed-form expression of the intermediate quantity with filters into the solution of an optimization problem. In our case, we can achieve the optimal bias-variance decomposition with the following quantity:

$$\begin{aligned} \vartheta_\lambda^0 &= \theta^\star, \\ \vartheta_\lambda^{k+1} &= \mathsf{prox}_{\widehat{L}/\lambda}(\vartheta_\lambda^k), \quad k \geq 0. \end{aligned} \tag{16}$$

Consequently, we write

$$\|\widehat{\theta}_\lambda^t - \theta^\star\|_{\mathbf{H}(\theta^\star)} \leq \|\widehat{\theta}_\lambda^t - \vartheta_\lambda^t\|_{\mathbf{H}(\theta^\star)} + \|\vartheta_\lambda^t - \theta^\star\|_{\mathbf{H}(\theta^\star)}. \tag{17}$$

We recover Eq. (15) with the square loss. In [1], a different decomposition is used; we found Eq. (16) to greatly simplify the proof.

Bounding the bias and the variance. The first term in Eq. (17) is the *bias* of the estimator, as it goes to 0 when the regularization λ goes to 0. By applying the lower bound on gradient values – Eq. (14) – with the definition of the proximal operator, one can express $\|\widehat{\theta}_\lambda^t - \vartheta_\lambda^t\|$ function of $\|\widehat{\theta}_\lambda^{t-1} - \vartheta_\lambda^{t-1}\|$. Unfolding the recursion, we obtain Theorem 2 in the appendix. It shows that the bias decreases in $O\left(\lambda^{r+1/2}\right)$ if the qualification is sufficient, *i.e.* $t \geq r + 1/2$. Otherwise, we recover the saturation experienced with least squares: the bias only decreases in $O\left(\lambda^t\right)$. Specific attention is devoted to bounding the prefactor, which is otherwise difficult to manage.

The second term in Eq. (17) is the *variance*, as it goes to 0 when the number of samples n increases. Theorem 3 shows that it decays in $O(\sqrt{\mathsf{df}_\lambda/n})$. It follows closely the work of [17]. However, we cannot use the convenient fact that $\mathsf{df}_\lambda = \operatorname{Tr} \mathbf{H}(\theta^\star)\widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star)$, which is valid for least squares but not in general. Thus, we took specific care in adapting our bounds to the different regimes so as not to impact the learning rate.

Plugging these results together, we obtain the upper bound on the excess risk.

3.2 Optimization

The aim of this section is to extend the result of Theorem 1 to a practical case, where we only have access to an inexact solver for computing the proximal operator. Specifically, let $\epsilon > 0$ be the error (to be defined precisely in Proposition 1) made when approximating $\widehat{\theta}_\lambda^t$ with $\overline{\theta}_\lambda^t$, the quantity we compute numerically. We aim for a bound of the type:

$$L(\overline{\theta}_\lambda^t) - L(\theta^\star) \leq \mathsf{C}_{\text{risk}}\, n^{-\frac{\alpha(2r+1)}{1+\alpha(2r+1)}} + \epsilon.$$

The first term in the right hand side is the *statistical error*, and is optimal following the discussion of Theorem 1. The second term is the *optimization error*, which is the price to pay for approximating $\widehat{\theta}_\lambda^k$ by $\overline{\theta}_\lambda^k$ with tolerance ϵ. The goal is to give a simple optimization rule on the sub-problems to ensure that ε is of the same order as the upper-bound for the noiseless case.

Assuming that we cannot compute the proximal operator in Eq. (4) exactly, we need to evaluate how the error in approximating $\widehat{\theta}_\lambda^1$ propagates to the evaluation of $\widehat{\theta}_\lambda^2$, and so on. As generalized self-concordant functions are well suited to (approximate) second-order optimization scheme, we

assume we use a solver with guarantees on a quantity called *Newton decrement*, such as the one developed in [14]. Starting from $\widehat{\theta}_\lambda^0 = \overline{\theta}_\lambda^0 = 0$, define the following for $k > 0$:

$$\widehat{\theta}_\lambda^k = \arg\min_{\theta \in \mathcal{H}} \widehat{L}_\lambda^{k-1}(\theta) \overset{\text{def.}}{=} \widehat{L}(\theta) + \frac{\lambda}{2}\left\|\theta - \widehat{\theta}_\lambda^{k-1}\right\|^2, \quad \nu_\lambda^k(\theta) \overset{\text{def.}}{=} \left\|\nabla \widehat{L}_\lambda^{k-1}(\theta)\right\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta)}, \tag{18}$$

$$\overline{\theta}_\lambda^k \approx \arg\min_{\theta \in \mathcal{H}} \overline{L}_\lambda^{k-1}(\theta) \overset{\text{def.}}{=} \widehat{L}(\theta) + \frac{\lambda}{2}\left\|\theta - \overline{\theta}_\lambda^{k-1}\right\|^2, \quad \overline{\nu}_\lambda^k(\theta) \overset{\text{def.}}{=} \left\|\nabla \overline{L}_\lambda^{k-1}(\theta)\right\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta)}. \tag{19}$$

$\overline{\theta}_\lambda^k$ approximates the proximal operator evaluated on $\overline{\theta}_\lambda^{k-1}$, and $\overline{L}_\lambda^{t-1}$ is the function we manipulate at step t. If the optimization was carried without error in Eq. (19), we would have $\overline{L}^{t-1} = \widehat{L}^{t-1}$. The quality of the approximation is measured with the Newton decrement of Eq. (19), see *e.g*, Lemma 6 of [14]. We need to enforce a bound on the true Newton decrement in Eq. (18) when we only have access to $\overline{L}_\lambda^{t-1}$. The next proposition gives a simple rule to achieve this.

Proposition 1 (Error propagation with proximal sequence). *Let $\epsilon > 0$ the target precision. Assume that we can solve each sub-problem with precision $\bar{\epsilon}_k$:*

$$\forall k \in \{1, \dots, t\}, \quad \overline{\nu}_\lambda^{k-1}\left(\overline{\theta}_\lambda^k\right) \leq \bar{\epsilon}_k = \epsilon \frac{1.4^{k-t}}{t},$$

and that $\epsilon \leq \sqrt{\lambda}/(2R)$. This suffice to achieve an error ϵ on the target function:

$$\nu_\lambda^{t-1}\left(\overline{\theta}_\lambda^t\right) \leq \epsilon.$$

This is a specialized version of Proposition 7, whose proof is detailed in the appendix. Intuitively, this means that enforcing a geometrically higher precision on the first steps is sufficient to obtain high precision on the final estimate. To compute IT's estimator in practice, one would need to solve t optimization problem with decreasing precision. As second order schemes have double logarithmic complexity w.r.t the precision ϵ, the complexity of computing the proximal sequence of IT with tolerance ϵ would be only (up to logarithm term) t times bigger than estimating Tikhonov estimator with tolerance ϵ. In practice, when learning with kernels, one would use the representer theorem and aim at estimating β in $\mathbb{R}^n$ as in Eq. (6) [27]. This results in an optimization problem with n observations in dimension n, with complexity $O(n^3)$. A practical implementation could use Nyström projection to avoid this cubic computational burden in the number of samples. The statistical effects of such projection are well studied with Tikhonov regularization [14, 15]; their effect on other regularization scheme is an interesting future research direction.

This proposition can be used directly to bound the excess risk with inexact solvers.

Proposition 2 (Upper bound on the excess risk with inexact solvers). *Let $\delta \in (0, 1)$ and assume that the* statistical *assumptions of Theorem 1 hold as well as the* optimization *assumptions of Proposition 1. Then, the following bound on the excess risk holds with probability greater than $1 - 2\delta$:*

$$L(\overline{\theta}_\lambda^t) - L(\theta^\star) \leq 2\mathsf{C}_{\text{risk}}\, n^{-\frac{\alpha(2r+1)}{1+\alpha(2r+1)}} + \mathsf{E}_{1/2}\,\epsilon, \quad s = \min\{r + 1/2, t\} \tag{20}$$

with C_{risk} as in Theorem 1 and $\mathsf{E}_{1/2} \leq 4.3 \cdot 10^3$.

This is a specialized version of Proposition 8 proved in the appendix. The first term is the statistical excess risk, whereas the second term in ϵ is the price we pay for inexact approximation. For the sake of clarity, crude upper bounds were used (notably $\widehat{\mathbf{H}}_\lambda^{-1/2}(\cdot) \leq \mathsf{B}_2^\star/\sqrt{\lambda}$) at the expanse of big constants. They can be expected to be an order of magnitude lower in practice.

Setting t in real application. In classical machine learning settings, we do not have access to the source condition parameter r. The number of proximal steps t can be seen as an hyperparameter, which is chosen by cross-validation. One would run the algorithm and test the resulting error on a validation set for each iteration, and keep doing proximal steps as long as the validation loss improves.

4 Experiments

The purpose of the experiments is to illustrate the saturation effect of the Tikhonov estimator when $r \gg 1/2$, and see how the saturation is overcome by iterated Tikhonov IT. We also show that the statistical rates we derive are achieved both in theory and in practice on synthetic data with well-controlled source and capacity conditions.

Settings. To that end, we use a synthetic binary classification data set for which we know the source and capacity condition parameters r and α by design. Then, we study the performance of $\mathsf{IT}(t)$, $t \in \{1, \dots, 8\}$, trained with the logistic loss, which satisfies Definition 2 about generalized self-concordant functions. Related experiments were conducted in the context of kernel ridge regression with synthetic data in [16], which we follow here. Specifically, we use splines of order α to define a kernel matrix:

$$K(x,z) = \Lambda_\alpha(x,z) = \sum_{k \in \mathbb{Z}} \frac{e^{2i\pi k(x-z)}}{|k|^\alpha},$$

for which a closed form expression is available as soon as α is a positive even integer (see for instance Eq (2.1.7) in [28]). We then use $\mathcal{X} = [0,1]$, ρ_x is the uniform distribution, and $\theta^\star(x) = \Lambda_{(r+1/2)\alpha+1/2}(0,\cdot)$, which may be shown to live in the RKHS $\mathcal{H}$ of K. Then, it is possible to show that the source and capacity assumption are satisfied with value r, α, see [16].

Finally, we design the distribution $\rho_{y|x}$ of the labels such that $\theta^\star$ is indeed the minimizer of the risk over $\mathcal{H}$. This may be ensured if $\theta^\star$ coincides with the minimizer of the risk over the set of measurable functions, which has the following form under mild assumptions (see Eq. (3) in [26]):

$$\theta^\star(x) = \arg\min_z \mathbb{E}_{y|x}\left[\ell(y,z)\right]. \tag{21}$$

The previous relation can be satisfied by choosing $\rho_{y|x}$ accordingly. More precisely, we need

$$\mathcal{Y} = \{-1,1\}, \quad \mathbb{P}(y=1 \mid x) = \left(1 + e^{-\theta^\star(x)}\right)^{-1}, \quad \mathbb{P}(y=-1 \mid x) = \left(1 + e^{\theta^\star(x)}\right)^{-1},$$

which ensures that Eq. (21) holds – see details in Appendix E.3. To our knowledge, this is the first synthetic dataset with given source and capacity condition for classification tasks. For each λ, t, we sample n points uniformly on $[0,1]$, evaluate $\theta^\star$, the observed labels y_i, and $\widehat{\theta}^t_\lambda$. We evaluate the excess risk $L(\widehat{\theta}^t_\lambda) - L(\theta^\star)$ with Monte Carlo sampling. We then report the *lowest excess risk* achieved across the regularization λ, and the *optimal regularization* used to achieve this loss. We plot lines of slope $2s\alpha/1+2s\alpha$ and $\alpha/1+2s\alpha$ respectively, with $s = (r+1/2) \wedge t$ in order to compare the statistical rates achieved in practice and in theory.

Results. Results for the logistic loss are available in Fig. 1 and we also present results with least squares where the noise is Gaussian in Appendix E.2. We set $\alpha = 2$, $r \in \{1/4, 41/4\}$, and we study the performance of Iterated Tikhonov estimators with $t \in \{1, 3, 8\}$. $t = 1$ corresponds to Tikhonov estimator and saturates at $r = 1/2$. $\mathsf{IT}(3)$ and $\mathsf{IT}(8)$ saturates at $r = 5/2$ and $r = 15/2$ respectively. Consequently, all estimators have optimal rates on the difficult task with $r = 1/4$; however, only IT exploits the additional regularity of the easy task, with $r = 41/4$. This experimentally shows that better sample complexity can be achieved when the learning task is easier and t is high, matching the rates predicted in Theorem 4, which are $n^{-\alpha(1+2s)/1+\alpha(1+2s)}$, with $s = \min\{r, t-1/2\}$. Learning rates were estimated with an ordinary least square regression in log-log scale, and are given in Table 1, where they are compared with the theoretical values. To conclude, we observe a slight improvement in absolute value of the excess risk in the range $r \ll t$, suggesting that IT is useful even when the learning task is hard. This could be because of lower constants for high t: *e.g.* we show that $\mathsf{C}_{\mathrm{bias}}$ decays in $1/t^r$ when $t \geq r + 1/2$, see Theorem 2 in the appendix. We report in the appendix additional experimental results such as plots with the chosen regularization λ as a function of n, and plots on the ratio between the excess risk of $\mathsf{IT}(t)$ and Tikhonov, to show that the former is consistently better than the latter on easy tasks.

5 Conclusion

This paper studies a well-known regularization scheme for least square, and extend it for the first time to other loss functions, which notably contain the logistic loss used for classification. We prove

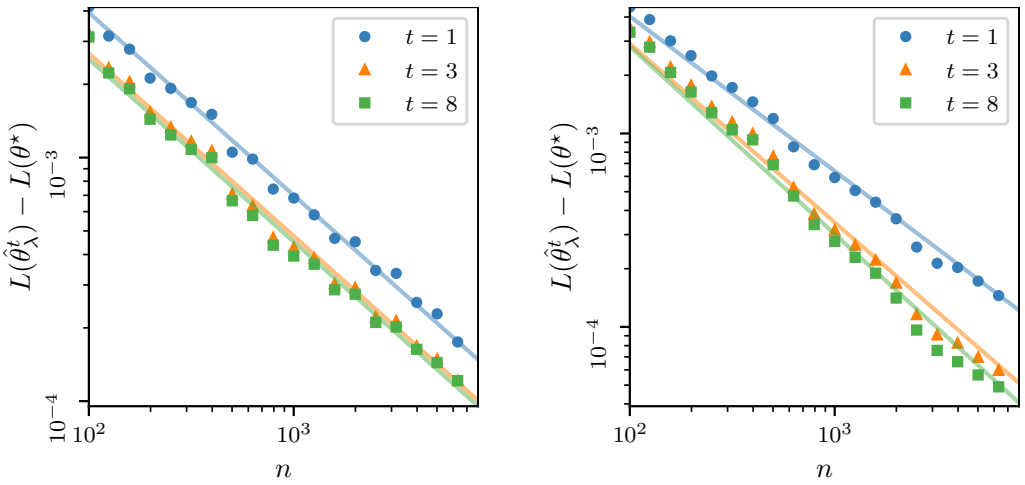

Figure 1: Excess risk for various Iterated Tikhonov estimators as a function of n. **Colors**: $t = 1$ (Tikhonov) estimator is shown in blue; $t = 3, 8$ in green, orange. **Left**: from a difficult problem, $r = 1/4, \alpha = 2$. **Right**: easy problem, $r = 41/4, \alpha = 2$. Plain lines are predicted by theory, with slope $-\alpha(1+2s)/1+\alpha(1+2s)$, $s = \min\{r, t - 1/2\}$ (see main text). All plots are averaged over 100 runs of the optimization procedure with different initialization.

Table 1: Learning rate coefficients for capacity condition $\alpha = 2$ and various source condition assumption r. We estimate γ with ordinary least square with the model $L(\widehat{\theta}^t_\lambda) - L(\theta^\star) \propto n^{-\gamma}$. We display the coefficient we expect in theory, and the one we estimate.

	r	0.25	3.25	10.25
$t = 1$	Theory	0.75	0.80	0.80
	Estimation	0.71	0.73	0.72
$t = 3$	Theory	0.75	0.92	0.92
	Estimation	0.75	0.83	0.87
$t = 8$	Theory	0.75	0.94	0.97
	Estimation	0.79	0.95	0.98

that Iterated Tikhonov, corresponding to proximal point iterations, has optimal learning rates and higher qualification than Tikhonov, and as such could outperform it on easy tasks. We extend the scope of the theory of learning with generalized self concordant loss functions beyond standard Tikhonov regularization, which fills a gap in the previous theory, showing that it is possible to be fully adaptive to the regularity of the learning problem, without saturation effects. On top of this, we gave sufficient conditions to compute the estimator in practice, which is nontrivial by its sequential nature. Interesting research directions include related regularization schemes, such as boosting, but also implementations of the iterated Tikhonov procedure with sketching techniques as Nyström projections. The goal is to derive algorithms that are both optimal, in terms of statistical guarantees, and with reduced computational complexity, which is an aspect we will address in future work.

Acknowledgments

A.R. acknowleges support of the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). A.R. acknowledges support of the European Research Council (grant REAL 947908). J. Mairal was supported by the ERC grant number 714381 (SOLARIS project) and by ANR 3IA MIAI@Grenoble Alpes, (ANR19-P3IA-0003).

References

[1] U. Marteau-Ferey, D. Ostrovskii, F. Bach, and A. Rudi. Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. In *Conference on Learning Theory (COLT)*, 2019.

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

[3] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

[4] B. Schölkopf and A. Smola. Support vector machines and kernel algorithms. *Encyclopedia of Biostatistics*, 04 2002.

[5] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

[6] Y. Averyanov and A. Celisse. Early stopping and polynomial smoothing in regression with reproducing kernels. *arXiv preprint arXiv:2007.06827*, 2020.

[7] E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, and F. Odone. Learning from examples as an inverse problem. *Journal of Machine Learning Research (JMLR)*, 6(30):883–904, 2005.

[8] F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

[9] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[10] Y. Nesterov and A. Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

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

[12] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[13] Tianxiao Sun and Quoc Tran-Dinh. Generalized self-concordant functions: A recipe for newton-type methods, 2018.

[14] U. Marteau-Ferey, F. Bach, and A. Rudi. Globally convergent newton methods for ill-conditioned generalized self-concordant losses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[15] A. Rudi, R. Camoriano, and L. Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[16] A. Rudi and L. Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[17] G. Blanchard and N. Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18, 2016.

[18] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer, 2006.

[19] J. Thomas King and Chillingworth D. Approximation of generalized inverses by iterated regularization. *Numerical Functional Analysis and Optimization*, 1(5):499–513, 1979.

[20] H. Lin, J. Mairal, and Z. Harchaoui. Catalyst acceleration for first-order convex optimization: from theory to practice. *Journal of Machine Learning Research (JMLR)*, 18(1):7854–7907, 2018.

[21] A. Kulunchakov and J. Mairal. Estimate sequences for stochastic composite optimization: Variance reduction, acceleration, and robustness to noise. *Journal of Machine Learning Research (JMLR)*, 21(155):1–52, 2020.

[22] Yair Carmon, Arun Jambulapati, Qijia Jiang, Yujia Jin, Yin Tat Lee, Aaron Sidford, and Kevin Tian. Acceleration with a ball optimization oracle. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[23] S. Lin, Y. Lei, and D. Zhou. Boosted kernel ridge regression: Optimal learning rates and early stopping. *Journal of Machine Learning Research (JMLR)*, 20(46):1–36, 2019.

[24] Tong Zhang. Sequential greedy approximation for certain convex optimization problems. *IEEE Transactions on Information Theory*, 49(3):682–691, 2003.

[25] A. Dieuleveut, N. Flammarion, and F. Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *Journal of Machine Learning Research (JMLR)*, 18(101):1–51, 2017.

[26] C. Ciliberto, L. Rosasco, and A. Rudi. A general framework for consistent structured prediction with implicit loss embeddings. *Journal of Machine Learning Research (JMLR)*, 21(98):1–67, 2020.

[27] B. Schölkopf, R. Herbrich, and A. Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426, 2001.

[28] G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990.

[29] R. Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

[30] J. Fujii, M. Fujii, T. Furuta, and R. Nakamoto. Norm inequalities equivalent to heinz inequality. In *Proceedings of the American Mathematical Society*, 1993.

[31] J. Mairal. Cyanure: An open-source toolbox for empirical risk minimization for Python, C++, and soon more, 2019.

Appendix

Table of Contents

A Settings, notations and assumptions

Given a separable Hilbert space $\mathcal{H}$, $\|\cdot\|$ denotes the norm in $\mathcal{H}$. For any operator A on $\mathcal{H}$, $\|A\|$ denotes its operator norm, and $\operatorname{Tr} A$ its trace norm. If A is a p.d operator, we denote by $\|\cdot\|_A = \left\|A^{1/2}\cdot\right\|$ the norm induced by A. We denote $\|A\|_{HS}$ the Hilbert Schmidt norm of A. We use the short-hand notation

$$A_\lambda = A + \lambda \mathbf{I},$$

where $\mathbf{I}$ is the identity. We denote by $a \wedge b$ the minimum of $\{a, b\}$, and $a \vee b$ its maximum.

A.1 Settings and technical assumptions

The settings in this subsection are the same as in [1]. We report them for completeness.

Let $\mathcal{X}$ a Borel input space, $\mathcal{Y}$ be a vector-valued output spaces, and ρ a probability distribution on $\mathcal{X} \times \mathcal{Y}$. We consider $\mathcal{H}$ to be a separable Hilbert space of functions from $\mathcal{X}$ to $\mathcal{Y}$. We consider a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ for measuring the fit between predictions and true labels. Given n observations $(x_1, y_1), \ldots, (x_n, y_n)$ i.i.d according to ρ, the goal is to build a measurable function $\widehat{\theta}$, which minimizes the expected loss

$$L(\widehat{\theta}) = \mathbb{E}_{x,y\sim\rho}\left[\ell(y, \widehat{\theta}(x))\right].$$

In this paper, we evaluate the quality of the estimator with probabilistic upper bounds on the *excess risk*

$$L(\widehat{\theta}) - \inf_{\theta\in\mathcal{H}} L(\theta) \leq K n^{-\gamma},$$

with probability greater than $1-\delta$. The rate of decay γ is referred to as the *learning rate* of the estimator. Our main assumption on the loss function is to be generalized self-concordant (GSC).

Assumption 5 (Generalized Self-Concordance). *For any $z = x, y \in \mathcal{X} \times \mathcal{Y}$, the function $\ell_z : \mathcal{H} \to \mathbb{R}$ defined as $\ell_z(\theta) = \ell(y, \theta(x))$ for $\theta \in \mathcal{H}$ is convex and three times differentiable. Besides, there exists a set $\phi(z) \subset \mathcal{H}$ s.t*

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

Next, we introduce the following quantities.

Definition 5 (Useful quantities). *Let $\theta \in \mathcal{H}$. The following quantities are independant of the random variable $z \sim \rho$, either by taking the supremum over the support of ρ or by considering the expectation. Define:*

- *uniform bounds on the derivatives:*

$$\mathsf{B}_1(\theta) = \sup_{z \in \mathsf{Supp}\, \rho} \left\|\nabla \ell_z(\theta)\right\|, \quad \mathsf{B}_2(\theta) = \sup_{z \in \mathsf{Supp}\, \rho} \operatorname{Tr} \nabla^2 \ell_z(\theta);$$

- *the Hessian of the expected and empirical loss:*

$$\mathbf{H}(\theta) = \nabla^2 \, \mathbb{E}\left[\ell_z(\theta)\right], \quad \widehat{\mathbf{H}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla^2 \ell_{z_i}(\theta);$$

- *the function t, s.t:*

$$\mathsf{t}(\theta) = \sup_{z \in \mathsf{Supp}\, \rho} \sup_{g \in \phi(z)} |\theta \cdot g| \, .$$

We make technical assumption to ensure that the loss function and its derivatives are well defined everywhere and that we can exchange expectation and derivative.

Assumption 6 (Technical assumptions). *There exists R s.t $\sup_{g \in \phi(z)} \|g\| \leq R$ almost surely; $|\ell_z(0)|, \|\nabla \ell_z(0)\|, \operatorname{Tr} \mathbf{H}(\nabla^2 \ell_z(0))$ are almost surely bounded.*

Using Prop. 2 of [1], we have that $\mathsf{B}_1(\theta), \mathsf{B}_2(\theta), L(\theta), \nabla L(\theta), \mathbf{H}(\theta)$ exist for all $\theta \in \mathcal{H}$, and

$$\nabla L(\theta) = \mathbb{E}\left[\nabla \ell_z(\theta)\right], \quad \mathbf{H}(\theta) = \mathbb{E}\left[\nabla^2 \ell_z(\theta)\right].$$

Finally, $\mathbf{H}(\theta)$ is trace-class, that is its trace is finite for any $\theta \in \mathcal{H}$. The same properties hold when considering $\hat{\rho}$ instead of ρ, that is for the quantities $\widehat{L}(\theta), \nabla \widehat{L}(\theta)$ and $\widehat{\mathbf{H}}(\theta)$.

We make three key assumptions to obtain our learning rate.

Assumption 7 (Existence of a minimizer). *There exists a minimizer of L in $\mathcal{H}$. There is $\theta^\star \in \mathcal{H}$ s.t*

$$L(\theta^\star) = \inf_{\theta \in \mathcal{H}} L(\theta^\star).$$

Assumption 8 (Source condition). *There exists $r > 0$ and $v \in \mathcal{H}$ s. t*

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

The third assumption qualifies the ill-posedness of the problem:

Assumption 9 (Capacity condition). *There exists $\alpha > 1$, $\mathsf{s}, \mathsf{S} > 0$ s.t*

$$\mathsf{s} \lambda^{-1/\alpha} \leq \mathsf{df}_\lambda \leq \mathsf{S} \lambda^{-1/\alpha}.$$

To understand the source and capacity condition, one must pay attention to the counterpart of the covariance operator for GSC loss function, that is the expected hessian at optimality. It is denoted with $\mathbf{H}(\theta^\star)$ throughout the paper. The source and capacity conditions are assumptions on the eigen-decomposition of this operator. To better quanitfy these assumptions, take σ_j, e_j an eigenbasis of $\mathbf{H}(\theta^\star)$, with $\sigma_j > \sigma_{j+1}$.

The source condition is a smoothness assumption on $\theta^\star$. It amounts to assuming that the eigendecomposition of $\theta^\star$ on the basis of the Hessian decays faster than its spectrum. Indeed, rewriting Assumption 8 we obtain

$$\|v\|^2 = \sum_{j\geq 1} \sigma_j^{-2r} \langle \theta^\star, e_j \rangle^2 < +\infty.$$

Assuming $r = 0$ simplifies to $\theta^\star \in \mathcal{H}$. Bigger r implies that the optimum can be well approximated by the first few eigenvectors (as $(\sigma_j^{-2r})_j$ goes quickly to infinity).

Similarly, the capacity condition is an assumption on the decay of the spectrum of the Hessian. Specifically, it assumes that the spectrum decays polynomially, i.e $\sigma_j \sim j^{-\alpha}$. As this operator is compact, we have $\alpha > 1$ for the $\sum_j j^{-\alpha}$ to be summable. Bigger α gives easier input space $\mathcal{X}$.

See Section 2 in the main body of the paper for a discussion on the significance of these assumptions.

A.2 Basic results on GSC loss functions

Here, we present Prop. 4 of [1], which we then extend with an additional lemma.

Proposition 3 (Properties of GSC functions). *Let $\theta, \nu \in \mathcal{H}$, $\lambda \geq 0$. The following properties hold:*

$$\widehat{\mathbf{H}}_\lambda(\theta) \preceq e^{\mathsf{t}(\theta-\nu)} \widehat{\mathbf{H}}_\lambda(\nu) \tag{22}$$

$$\left\| \nabla \widehat{L}_\lambda(\theta) - \widehat{L}_\lambda(\nu) \right\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta)} \leq \|\theta - \nu\|_{\widehat{\mathbf{H}}_\lambda(\theta)} \, \overline{\phi}\left(\mathsf{t}(\theta - \nu)\right) \tag{23}$$

$$L_\lambda(\theta) - L_\lambda(\nu) - \nabla L_\lambda(\nu) \cdot (\theta - \nu) \leq \Psi\left(\mathsf{t}(\theta - \nu)\right) \|\theta - \nu\|^2_{\mathbf{H}_\lambda(\theta)} \tag{24}$$

where $\underline{\phi} : t \mapsto (1 - e^{-t})/t$ and $\Psi : t \mapsto (e^t - t - 1)/t^2$. Moreover, if $\nu, \xi \in \mathcal{H}$, $A : \mathcal{H} \to \mathcal{H}$ commutes with $\mathbf{H}(\xi)$, then the following holds:

$$e^{-\mathsf{t}(\theta-\xi)} \underline{\phi}\left(\mathsf{t}(\nu - \theta)\right) \|A(\nu - \theta)\|_{\mathbf{H}_\lambda(\xi)} \leq \|A(\nabla L_\lambda(\nu) - \nabla L_\lambda(\theta))\|_{\mathbf{H}_\lambda^{-1}(\xi)} \tag{25}$$

We slightly modify the lower bound gradient, which is crucial for obtaining higher qualification with IT.

Lemma 2 (Stacking operator on gradient bounds). *Let $\theta, \nu, \xi \in \mathcal{H}$, $\lambda > 0$. If $A : \mathcal{H} \to \mathcal{H}$ commutes with $\mathbf{H}(\xi)$, the following holds:*

$$e^{-\mathsf{t}(\theta-\xi)} \underline{\phi}\left(\mathsf{t}(\nu - \theta)\right) \|A(\nu - \theta)\|_{\mathbf{H}_\lambda(\xi)} \leq \|A(\nabla L_\lambda(\nu) - \nabla L_\lambda(\theta))\|_{\mathbf{H}_\lambda^{-1}(\xi)} .$$

Proof. Defining $v_s = \theta + s(\nu - \theta)$ for $s \in \{0, 1\}$, we have:

$$A^2 \left(\nabla L_\lambda(\nu) - \nabla L_\lambda(\theta)\right) = A^2 \int_0^1 \mathbf{H}_\lambda(v_s) \left(\nu - \theta\right) \mathrm{d}s,$$

$$\text{which implies} \quad \left\langle A^2 \left(\nabla L_\lambda(\nu) - \nabla L_\lambda(\theta)\right), \nu - \theta \right\rangle = A^2 \int_0^1 \left\langle \mathbf{H}_\lambda(v_s) \left(\nu - \theta\right), \nu - \theta \right\rangle \mathrm{d}s.$$

We may then use the lower bound on the Hessian from Eq. (22),

$$\mathbf{H}_\lambda(v_s) \succeq \mathbf{H}_\lambda(\xi) e^{-\mathsf{t}(v_s - \xi)} \succeq \mathbf{H}_\lambda(\xi) e^{-\mathsf{t}(\theta-\xi)} e^{-s\,\mathsf{t}(\nu-\theta)},$$

where the second inequality comes from t satisfying the triangle inequality. Plugging this in the previous equation and using the fact that $\mathbf{H}(\xi)$ and A commute, we have that:

$$\begin{aligned}\int_0^1 \left\langle A^2 \mathbf{H}_\lambda(v_s) \left(\nu - \theta\right), \nu - \theta \right\rangle \mathrm{d}s &\geq e^{-\mathsf{t}(\theta-\xi)} \int_0^1 e^{-s\,\mathsf{t}(\nu-\theta)} \mathrm{d}s \ \left\langle \mathbf{H}_\lambda(\xi) A\left(\nu - \theta\right), A\left(\nu - \theta\right) \right\rangle \\ &= e^{-\mathsf{t}(\theta-\xi)} \underline{\phi}\left(\mathsf{t}(\nu - \theta)\right) \left\langle \mathbf{H}_\lambda(\xi) A(\nu - \theta), A(\nu - \theta) \right\rangle,\end{aligned}$$

which gives the lower bound

$$e^{-\mathsf{t}(\theta-\xi)} \underline{\phi}\left(\mathsf{t}(\nu - \theta)\right) \|A(\nu - \theta)\|^2_{\mathbf{H}_\lambda(\xi)} \leq \left\langle A^2 \left(\nabla L_\lambda(\nu) - \nabla L_\lambda(\theta)\right), \nu - \theta \right\rangle. \tag{26}$$

On the other hand, with Cauchy Schwartz inequality, we obtain:

$$\left\langle A^2 (\nabla L_\lambda(\nu) - \nabla L_\lambda(\theta)), \nu - \theta \right\rangle \leq \|A(\nabla L_\lambda(\nu) - \nabla L_\lambda(\theta))\|_{\mathbf{H}_\lambda^{-1}(\xi)} \|A(\nu - \theta)\|_{\mathbf{H}_\lambda(\xi)} . \tag{27}$$

Combining the inequalities Eqs. (26) and (27) and dividing by $\|A(\nu - \theta)\|_{\mathbf{H}_\lambda(\xi)}$, we obtain the result needed. □

B Proof of Theorem 1

B.1 Error decomposition

Thanks to Eq. (24), the excess risk is bounded by the distance between estimate in $\mathbf{H}(\theta^\star)$ norm with

$$L_\lambda(\widehat{\theta}^t_\lambda) - L_\lambda(\theta^\star) \leq \Psi\left(\mathsf{t}(\widehat{\theta}^t_\lambda - \theta^\star)\right) \left\|\widehat{\theta}^t_\lambda - \theta^\star\right\|^2_{\mathbf{H}_\lambda(\theta^\star)}.$$

In order to compute $\|\widehat{\theta}^t_\lambda - \theta^\star\|_{\mathbf{H}(\theta^\star)}$, we need to go through an intermediate quantity ϑ. In the context of least squares and spectral filters, such quantity is usually defined to be

$$\vartheta = g_\lambda(\hat{T})\hat{S}^*\hat{S}\theta^\star, \tag{28}$$

where:

- $\hat{T} = \hat{S}^*\hat{S}$ is the *empirical covariance operator*, equal to $\sum_{i=1}^n \Psi(x_i) \otimes \Psi(x_i)$ when $\mathcal{H}$ is a RKHS with feature map Ψ (see Remark 4);
- $\hat{S} : \mathcal{H} \to \mathbb{R}^n$ is the *sampling operator*, with $\hat{S}\theta = 1/\sqrt{n}(\theta(x_i), \dots, \theta(x_n))$;
- Its dual is $\hat{S}^* : \mathbb{R}^n \to \mathcal{H}$, with $\hat{S}^* y = 1/\sqrt{n}\sum_{i=1}^n y_i \Phi(x_i)$;

see [17] for details. Thus, the quantity in Eq. (15) can be seen as the estimator trained on the *empirical noiseless distribution*, where we use $\hat{S}\theta^\star$ instead of $y = (y_i)_{1\leq i\leq n}$. It is optimal in the sense that its bias $\|\vartheta - \theta^\star\|_{\hat{T}}$ will be of the order of $\lambda^{r+1/2}$ and its variance $\left\|\widehat{\theta}^t_\lambda - \vartheta\right\|_{\hat{T}}$ of the order of df_λ / n, leading to the optimal rates for least squares [3].

Expressing the quantity above as a proximal sequence is the key insight of the proof. It turns out that the following quantity obtains the same optimal decomposition.

Definition 6 (Error decomposition). *Define the following quantity:*

$$\begin{aligned}\vartheta^0_\lambda &= \theta^\star \\ \vartheta^{k+1}_\lambda &= \mathsf{prox}_{\widehat{L}/\lambda}(\vartheta^k_\lambda), \quad k \geq 0\end{aligned}$$

Remark 5. In fact, the estimator above, when expressed with filters, has its (bias, variance) equals to the (variance, bias) of the estimator of Eq. (28). It is easy to change the intermediate quantity of Definition 6 to match, but it introduces unnecessary burden with the notations.

The purpose of next sections is to bound

$$\left\|\widehat{\theta}^t_\lambda - \theta^\star\right\|_{\mathbf{H}(\theta^\star)} \leq \left\|\widehat{\theta}^t_\lambda - \vartheta^t_\lambda\right\|_{\mathbf{H}(\theta^\star)} + \left\|\vartheta^t_\lambda - \theta^\star\right\|_{\mathbf{H}(\theta^\star)}. \tag{29}$$

The first term will be the *bias* of the estimator (decreases with λ/t) while the second one will be the *variance* (decreases with t/λ and n). The intermediate quantity of Definition 6 being very close to the one of Eq. (28) used in [17], it is natural that the proof follows similarly.

B.2 Bounding the bias

Here, we proceed in bounding the bias, that is the quantity $\|\widehat{\theta}^t_\lambda - \vartheta^t_\lambda\|_{\mathbf{H}(\theta^\star)}$.

Theorem 2 (Improved qualification of Iterated Tikhonov estimator). *Let $\delta \in (0, 1]$. Recall the source condition of parameter $r, \|v\|$. Define the following conditions on the number of samples:*

$$\begin{aligned}\mathsf{H}_1 &: n \geq 24\frac{\mathsf{B}_2^\star}{\lambda}\log\frac{16\mathsf{B}_2^\star}{\lambda\delta}, \\ \mathsf{H}_{1b} &: n \geq 8\frac{\mathsf{B}_2^{\star 2}}{\lambda^2}\log^2\frac{4}{\delta}, \\ \mathsf{H}_2 &: n \geq 2\left[1 \vee \left(\frac{2\mathsf{B}_2^\star(t-1/2)^r}{\lambda^{s-1/2}}\right)^2\right]\log\frac{4}{\delta},\end{aligned}$$

Now assume:

$$\begin{aligned} \mathsf{H}_1 \quad & \textit{if } r \le 1/2, \\ \mathsf{H}_1 + \mathsf{H}_{1b} \quad & \textit{if } 1/2 < r \le 1, \\ \mathsf{H}_1 + \mathsf{H}_2 \quad & \textit{if } r > 1. \end{aligned}$$

Then, with probability greater than $1-\delta$*:*

$$\left\| \widehat{\theta}_\lambda^t - \vartheta_\lambda^t \right\|_{\mathbf{H}(\theta^\star)} \le \sqrt{2}\mathsf{T}(r,t)\mathsf{P}_\lambda^t \lambda^s, \tag{30}$$

with $s = (r+1/2) \wedge t$*,*

$$\mathsf{T}(r,t) = \begin{cases} \|v\| \left(1 \vee (\mathsf{B}_2^\star + \lambda)\right) 2^r & \textit{if } r \le 1, \\ \|v\| \frac{w(r)+r}{(t-1/2)^r} & \textit{if } r > 1 \textit{ and } r + 1/2 < t \\ \|v\| \frac{w(r)}{(t-1/2)^r} + \mathsf{B}_2^{\star r - t + 1/2} & \textit{if } r > 1 \textit{ and } r + 1/2 \ge t, \end{cases} \tag{31}$$

$w(r) = r 2^{\lfloor r \rfloor + 1} \mathsf{B}_2^{\star r}$*, and:*

$$\mathsf{P}_\lambda^t \overset{\text{def.}}{=} \prod_{k=1}^t \underline{\phi}^{-1}\left(\hat{\mathsf{t}}(\widehat{\theta}_\lambda^k - \vartheta_\lambda^k)\right) e^{\hat{\mathsf{t}}(\vartheta_\lambda^k - \theta^\star)}.$$

This term is the optimal bias for LS with the usual excess risk decomposition. The saturation effect is explicit: we go from a bias decay in λ^t when $t \le r+1/2$ to λ^r when the source condition saturates IT's regularization. That is, IT's estimator has a qualification of t, in the sense that it can exploits source condition up to $r = t - 1/2$. If $r > t - 1/2$, the estimator saturates and the learning rate becomes suboptimal.

Proof. This proof simply relies on the upper bound on gradients enabled by GSC functions. We will use Lemma 2 for that purpose. Also, we will use the definition of a proximal sequence; that is, we have that

$$\forall k \le t, \quad \nabla \widehat{L}(\widehat{\theta}_\lambda^k) + \lambda(\widehat{\theta}_\lambda^k - \widehat{\theta}_\lambda^{k-1}) = 0,$$

which is just another way of saying that we perform implicit gradient steps of size $1/\lambda$.

Changing the norm. We first change the norm we operate on:

$$\begin{aligned} \left\| \widehat{\theta}_\lambda^t - \vartheta_\lambda^t \right\|_{\mathbf{H}(\theta^\star)} &\le \left\| \widehat{\mathbf{H}}_\lambda^{-1/2}(\theta^\star) \mathbf{H}^{1/2}(\theta^\star) \right\| \left\| \widehat{\theta}_\lambda^t - \vartheta_\lambda^t \right\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star)} \\ &\le \left\| \widehat{\mathbf{H}}_\lambda^{-1/2}(\theta^\star) \mathbf{H}_\lambda^{1/2}(\theta^\star) \right\| \left\| \widehat{\theta}_\lambda^t - \vartheta_\lambda^t \right\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star)}. \end{aligned}$$

We bound the operator norm using Proposition 9 in Appendix D, with $\mathcal{F}_\lambda = \mathsf{B}_2^\star/\lambda$. We obtain:

$$\mathsf{H}_1 : n \ge 24 \frac{\mathsf{B}_2^\star}{\lambda} \log \frac{8\mathsf{B}_2^\star}{\lambda\delta} \implies \left\| \widehat{\mathbf{H}}_\lambda^{-1/2}(\theta^\star) \mathbf{H}_\lambda^{1/2}(\theta^\star) \right\| \le \sqrt{2}. \tag{32}$$

We now proceed in bounding the distance between estimates, that is the quantity $\|\widehat{\theta}_\lambda^t - \vartheta_\lambda^t\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star)}$. We denote

$$s = (r+1/2) \wedge t. \tag{33}$$

Upper bound on gradients. Use Lemma 2 on $\widehat{L}_\lambda$ to have:

$$\begin{aligned} \left\| \widehat{\theta}_\lambda^t - \vartheta_\lambda^t \right\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star)} &\le \underline{\phi}^{-1}\left(\hat{\mathsf{t}}(\widehat{\theta}_\lambda^t - \vartheta_\lambda^t)\right) e^{\hat{\mathsf{t}}(\vartheta_\lambda^t - \theta^\star)} \left\| \nabla \widehat{L}_\lambda(\widehat{\theta}_\lambda^t) - \nabla \widehat{L}_\lambda(\vartheta_\lambda^t) \right\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star)} \\ &= \underline{\phi}^{-1}\left(\hat{\mathsf{t}}(\widehat{\theta}_\lambda^t - \vartheta_\lambda^t)\right) e^{\hat{\mathsf{t}}(\vartheta_\lambda^t - \theta^\star)} \left\| \lambda(\widehat{\theta}_\lambda^{t-1} - \vartheta_\lambda^{t-1}) \right\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star)} \\ &= \underline{\phi}^{-1}\left(\hat{\mathsf{t}}(\widehat{\theta}_\lambda^t - \vartheta_\lambda^t)\right) e^{\hat{\mathsf{t}}(\vartheta_\lambda^t - \theta^\star)} \left\| \lambda \widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star)(\widehat{\theta}_\lambda^{t-1} - \vartheta_\lambda^{t-1}) \right\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star)}. \end{aligned}$$

Let us detail the recursion. Let $k \leq t$. Then, the following inequality holds, thanks to Lemma 2:

$$\begin{aligned}
\left\|\lambda^k \widehat{\mathbf{H}}_\lambda^{-k}(\theta^\star)(\widehat{\theta}_\lambda^{t-k} - \vartheta_\lambda^{t-k})\right\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star)} &\leq \underline{\phi}^{-1}\left(\hat{\mathsf{t}}(\widehat{\theta}_\lambda^{t-k} - \vartheta_\lambda^{t-k})\right) e^{\hat{\mathsf{t}}(\vartheta_\lambda^{t-k} - \theta^\star)} \\
&\quad \left\|\lambda^k \widehat{\mathbf{H}}_\lambda^{-k}(\theta^\star)\left(\nabla \widehat{L}_\lambda(\widehat{\theta}_\lambda^{t-k}) - \nabla \widehat{L}_\lambda(\vartheta_\lambda^{t-k})\right)\right\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star)} \\
&= \underline{\phi}^{-1}\left(\hat{\mathsf{t}}(\widehat{\theta}_\lambda^{t-k} - \vartheta_\lambda^{t-k})\right) e^{\hat{\mathsf{t}}(\vartheta_\lambda^{t-k} - \theta^\star)} \\
&\quad \left\|\lambda^{k+1} \widehat{\mathbf{H}}_\lambda^{-k}(\theta^\star)\left(\widehat{\theta}_\lambda^{t-(k+1)} - \vartheta_\lambda^{t-(k+1)}\right)\right\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star)} \\
&= \underline{\phi}^{-1}\left(\hat{\mathsf{t}}(\widehat{\theta}_\lambda^{t-k} - \vartheta_\lambda^{t-k})\right) e^{\hat{\mathsf{t}}(\vartheta_\lambda^{t-k} - \theta^\star)} \\
&\quad \left\|\lambda^{k+1} \widehat{\mathbf{H}}_\lambda^{-(k+1)}(\theta^\star)\left(\widehat{\theta}_\lambda^{t-(k+1)} - \vartheta_\lambda^{t-(k+1)}\right)\right\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star)}.
\end{aligned}$$

Thus, unfolding the recursion, we obtain:

$$\begin{aligned}
\left\|\widehat{\theta}_\lambda^t - \vartheta_\lambda^t\right\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star)} &\leq \mathsf{P}_\lambda^t \left\|\lambda^t \widehat{\mathbf{H}}_\lambda^{-t}(\theta^\star)\theta^\star\right\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star)}, \\
\text{with} \quad \mathsf{P}_\lambda^t &\stackrel{\text{def.}}{=} \prod_{k=1}^{t} \underline{\phi}^{-1}\left(\hat{\mathsf{t}}(\widehat{\theta}_\lambda^k - \vartheta_\lambda^k)\right) e^{\hat{\mathsf{t}}(\vartheta_\lambda^k - \theta^\star)}.
\end{aligned} \tag{34}$$

We now use the source condition on $\theta^\star$. Recall that it gives

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

for some $v \in \mathcal{H}$. Thus, we have:

$$\left\|\lambda^t \widehat{\mathbf{H}}_\lambda^{-t}(\theta^\star)\theta^\star\right\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star)} = \left\|\lambda^t \widehat{\mathbf{H}}_\lambda^{-(t-1/2)}(\theta^\star)\mathbf{H}^r(\theta^\star)v\right\| \tag{35}$$

$$\leq \left\|\lambda^t \widehat{\mathbf{H}}_\lambda^{-(t-1/2)}(\theta^\star)\mathbf{H}^r(\theta^\star)\right\| \|v\| \tag{36}$$

We need to distinguish between $r \leq 1$ and $r > 1$ to bound the operator norm

$$\left\|\lambda^t \widehat{\mathbf{H}}_\lambda^{-(t-1/2)}(\theta^\star)\mathbf{H}^r(\theta^\star)\right\|.$$

Case $r \leq 1$**.** We use the following decomposition:

$$\left\|\lambda^t \widehat{\mathbf{H}}_\lambda^{-(t-1/2)}(\theta^\star)\mathbf{H}^r(\theta^\star)\right\| \leq \left\|\lambda^t \widehat{\mathbf{H}}_\lambda^{-(t-1/2)}(\theta^\star)\widehat{\mathbf{H}}_\lambda^r(\theta^\star)\right\| \left\|\widehat{\mathbf{H}}_\lambda^{-r}(\theta^\star)\mathbf{H}^r(\theta^\star)\right\|.$$

The first term is bounded like this:

$$\begin{aligned}
\left\|\lambda^t \widehat{\mathbf{H}}_\lambda^{-(t-1/2)+r}(\theta^\star)\right\| &\leq \sup_{\widehat{\sigma}_{\min} < \sigma \leq \mathsf{B}_2^\star} \frac{\lambda^t}{(\sigma+\lambda)^{t-1/2-r}} \\
&\leq \lambda^s \begin{cases} 1 & \text{if } r + 1/2 < t \\ \mathsf{B}_2^\star + \lambda & \text{if } t = 1 \text{ and } r > 1/2. \end{cases}
\end{aligned}$$

This illustrates that Tikhonov regularization ($t = 1$) saturates at $r = 1/2$.

For the second term, write

$$\left\|\widehat{\mathbf{H}}_\lambda^{-r}(\theta^\star)\mathbf{H}^r(\theta^\star)\right\| \leq \left\|\widehat{\mathbf{H}}_\lambda^{-r}(\theta^\star)\mathbf{H}_\lambda^r(\theta^\star)\right\|$$

Then, use the Hermitian inequalities of Eq. (67) in Lemma 4, then use the concentration inequalities of Proposition 9. Both can be found in Appendix D. In details:

- If $r \leq 1/2$, use then the concentration inequality of Eq. (64):

$$\begin{aligned}
\left\|\widehat{\mathbf{H}}_\lambda^{-r}(\theta^\star)\mathbf{H}_\lambda^r(\theta^\star)\right\| &\leq \left\|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta^\star)\mathbf{H}_\lambda^{1/2}(\theta^\star)\right\|^{2r} \\
&\leq 2^{r/2} \text{ if } \mathsf{H}_1.
\end{aligned}$$

  with confidence $1 - \delta$.

- If $r > 1/2$, use the concentration inequality of Eq. (65):

$$\left\|\widehat{\mathbf{H}}_\lambda^{-r}(\theta^\star)\mathbf{H}_\lambda^r(\theta^\star)\right\| \leq \left\|\widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star)\mathbf{H}_\lambda(\theta^\star)\right\|^r \leq 2^r \text{ if } \mathsf{H}_{1b} : n \geq 8\frac{\mathsf{B}_2^{\star 2}}{\lambda^2}\log^2\frac{2}{\delta}.$$

All in all, after simplification, the bound on the operator norm when $r \leq 1$ reads

$$\left\|\lambda^t\widehat{\mathbf{H}}_\lambda^{-(t-1/2)}(\theta^\star)\mathbf{H}^r(\theta^\star)\right\| \leq \lambda^s(1 \vee (\mathsf{B}_2^\star + \lambda))2^r \quad \text{if } \begin{cases}\mathsf{H}_1 & \text{when } r \leq 1/2\\ \mathsf{H}_{1b} & \text{when } r > 1/2,\end{cases} \tag{37}$$

with confidence $1-\delta$. We now turn to the case $r > 1$.

Case $r > 1$. We tackle this case with a different decomposition:

$$\left\|\lambda^t\widehat{\mathbf{H}}_\lambda^{-(t-1/2)}(\theta^\star)\mathbf{H}^r(\theta^\star)\right\| \leq \left\|\lambda^t\widehat{\mathbf{H}}_\lambda^{-(t-1/2)}(\theta^\star)\widehat{\mathbf{H}}^r(\theta^\star)\right\| + \left\|\lambda^t\widehat{\mathbf{H}}_\lambda^{-(t-1/2)}(\theta^\star)(\mathbf{H}^r(\theta^\star) - \widehat{\mathbf{H}}^r(\theta^\star))\right\|$$

Looking at the first term; recalling that $\widehat{\mathbf{H}}(\theta^\star) \leq \mathsf{B}_2^\star$, we have:

$$\begin{aligned}\left\|\lambda^t\widehat{\mathbf{H}}_\lambda^{-(t-1/2)}(\theta^\star)\widehat{\mathbf{H}}^r(\theta^\star)\right\| &\leq \sqrt{\lambda}\sup_{0<\sigma\leq\mathsf{B}_2^\star}\left(\frac{\lambda}{\lambda+\sigma}\right)^{t-1/2}\sigma^r\\ &\leq \lambda^s\begin{cases}\frac{r}{(t-1/2)^r} & \text{if } \ r+1/2<t\\ \frac{\mathsf{B}_2^{\star r}}{\left(\mathsf{B}_2^{\star r}+\lambda\right)^{t-1/2}} & \text{otherwise}\end{cases}\\ &\leq \lambda^s\begin{cases}\frac{r}{(t-1/2)^r} & \text{if } \ r+1/2<t\\ \mathsf{B}_2^{\star r-t+1/2} & \text{otherwise}\end{cases}\end{aligned}$$

where we used the computation of Lemma 5. The second term can be upper bounded as follows:

$$\begin{aligned}\left\|\lambda^t\widehat{\mathbf{H}}_\lambda^{-(t-1/2)}(\theta^\star)(\mathbf{H}^r(\theta^\star) - \widehat{\mathbf{H}}^r(\theta^\star))\right\| &\leq \left\|\lambda^t\widehat{\mathbf{H}}_\lambda^{-(t-1/2)}(\theta^\star)\right\|\left\|\mathbf{H}^r(\theta^\star) - \widehat{\mathbf{H}}^r(\theta^\star)\right\|\\ &\leq w(r)\sqrt{\lambda}\left\|\mathbf{H}(\theta^\star) - \widehat{\mathbf{H}}(\theta^\star)\right\|\\ &\leq w(r)\frac{\lambda^s}{(t-1/2)^r} \text{ if } \mathsf{H}_2 : n \geq 2\left(1 \vee \left(\frac{2\mathsf{B}_2^\star(t-1/2)^r}{\lambda^{s-1/2}}\right)^2\right)\log\frac{2}{\delta}\end{aligned}$$

with confidence $1-\delta$. We applied Eq. (68) in Lemma 4 on the second inequality, and Eq. (66) in Proposition 9 for the last inequality, both of which can be found in Appendix D. We used:

$$w(r) = r2^{\lfloor r\rfloor+1}\mathsf{B}_2^{\star r}. \tag{38}$$

Thus, the bound on the operator norm when $r > 1$ reads:

$$\left\|\lambda^t\widehat{\mathbf{H}}_\lambda^{-(t-1/2)}(\theta^\star)\mathbf{H}^r(\theta^\star)\right\| \leq \lambda^s\begin{cases}\frac{w(r)+r}{(t-1/2)^r} & \text{if } \ r+1/2<t\\ \frac{w(r)}{(t-1/2)^r} + \mathsf{B}_2^{\star r-t+1/2} & \text{otherwise}\end{cases} \quad \text{if } \mathsf{H}_2, \tag{39}$$

with confidence $1-\delta$.

Gluing things together. We proceed to the conclusion. Define the following conditions:

$$\begin{aligned}\mathsf{H}_1 &: n \geq 24\frac{\mathsf{B}_2^\star}{\lambda}\log\frac{16\mathsf{B}_2^\star}{\lambda\delta},\\ \mathsf{H}_{1b} &: n \geq 8\frac{\mathsf{B}_2^{\star 2}}{\lambda^2}\log^2\frac{4}{\delta},\\ \mathsf{H}_2 &: n \geq 2\left[1 \vee \left(\frac{2\mathsf{B}_2^\star(t-1/2)^r}{\lambda^{s-1/2}}\right)^2\right]\log\frac{4}{\delta},\end{aligned}$$

where we replace δ by $\delta/2$ in order to have bounds with confidence $1-\delta/2$, so that the overall bound holds with confidence $1-\delta$ (in fact, $1-\delta/2$ in the first case). Now assume the following:

$$\begin{aligned} \mathsf{H}_1 \quad & \text{if } r \leq 1/2, \\ \mathsf{H}_1 + \mathsf{H}_{1b} \quad & \text{if } 1/2 < r \leq 1, \\ \mathsf{H}_1 + \mathsf{H}_2 \quad & \text{if } 1 < r. \end{aligned}$$

Then, we can chain the inequalities of Eqs. (32), (36), (37) and (39). We obtain:

$$\left\|\widehat{\theta}_\lambda^t - \vartheta_\lambda^t\right\|_{\mathbf{H}(\theta^\star)} \leq \sqrt{2}\,\|v\|\,\mathsf{P}_\lambda^t \lambda^s \begin{cases} (1 \vee (\mathsf{B}_2^\star + \lambda))2^r & \text{if } r \leq 1, \\ \frac{w(r)+r}{(t-1/2)^r} & \text{if } r > 1 \text{ and } r + 1/2 < t, \\ \frac{w(r)}{(t-1/2)^r} + \mathsf{B}_2^{\star r-t+1/2} & \text{if } r > 1 \text{ and } r + 1/2 \geq t, \end{cases} \tag{40}$$

with confidence $1-\delta$. □

B.3 Bounding the variance

After bounding the bias, we study the variance term: $\|\vartheta_\lambda^t - \theta^\star\|_{\mathbf{H}(\theta^\star)}$.

Theorem 3 (Optimal variance of Iterated Tikhonov estimator). *Let $\delta \in (0,1]$. Recall the definition of the degrees of freedom df_λ. Define the following conditions on the number of samples:*

$$\begin{aligned} \mathsf{H}_1 &: n \geq 24 \frac{\mathsf{B}_2^\star}{\lambda} \log \frac{16 \mathsf{B}_2^\star}{\lambda \delta}, \\ \mathsf{H}_3 &: n \geq 2 \frac{\mathsf{B}_1^{\star 2}}{\lambda\, \mathsf{df}_\lambda} \log \frac{4}{\delta}. \end{aligned}$$

Then, with probability greater than $1-\delta$:

$$\left\|\vartheta_\lambda^t - \theta^\star\right\|_{\mathbf{H}(\theta^\star)} \leq 4\sqrt{2} t \mathsf{R}_\lambda^t \sqrt{\frac{\mathsf{df}_\lambda}{n}} \cdot \sqrt{\log 2/\delta},$$

where we introduced:

$$\mathsf{R}_\lambda^t \overset{\text{def.}}{=} \prod_{k=1}^{t} \underline{\phi}^{-1}\left(\hat{\mathsf{t}}(\vartheta_\lambda^k - \theta^\star)\right).$$

Proof. The proof begins similarly to the study of the bias term (Theorem 2).

Changing the norm. We have the following bound (proof of Theorem 2, Eq. (32)):

$$\mathsf{H}_1 : n \geq 24 \frac{\mathsf{B}_2^\star}{\lambda} \log \frac{8 \mathsf{B}_2^\star}{\lambda \delta} \implies \left\|\vartheta_\lambda^t - \theta^\star\right\|_{\mathbf{H}(\theta^\star)} \leq \sqrt{2}\left\|\vartheta_\lambda^t - \theta^\star\right\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star)}. \tag{41}$$

Upper bounds on gradient. To ease the notation, we denote by $\mathsf{a}_k = \underline{\phi}^{-1}\left(\hat{\mathsf{t}}(\vartheta_\lambda^k - \theta^\star)\right)$. We have, thanks to the lower bound on gradient of Eq. (25):

$$\begin{aligned} \left\|\vartheta_\lambda^t - \theta^\star\right\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star)} &\leq \mathsf{a}_t \left\|\nabla \widehat{L}_\lambda(\vartheta_\lambda^t) - \nabla \widehat{L}_\lambda(\theta^\star)\right\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star)} \\ &= \mathsf{a}_t \left\|\lambda(\vartheta_\lambda^{t-1} - \theta^\star) - \nabla \widehat{L}(\theta^\star)\right\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star)} \\ &= \mathsf{a}_t \left\|\lambda \widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star)(\vartheta_\lambda^{t-1} - \theta^\star)\right\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star)} + \mathsf{a}_t \left\|\nabla \widehat{L}(\theta^\star)\right\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star)} \\ &\leq \mathsf{a}_t \mathsf{a}_{t-1} \left\|\lambda^2 \widehat{\mathbf{H}}_\lambda^{-2}(\theta^\star)(\vartheta_\lambda^{t-2} - \theta^\star)\right\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star)} + \left[\mathsf{a}_t + \mathsf{a}_{t-1}\left\|\lambda \widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star)\right\|\right]\left\|\nabla \widehat{L}(\theta^\star)\right\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star)} \end{aligned}$$

We can unfold the recursion. The first term will disappears thanks to $\vartheta_\lambda^0 = \theta^\star$, and we are left with:

$$\left\|\vartheta_\lambda^t - \theta^\star\right\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star)} \leq \sum_{k=0}^{t-1} \left(\prod_{i=t-k}^{t} \mathsf{a}_i \right) \left\|\lambda^k \widehat{\mathbf{H}}_\lambda^{-k}(\theta^\star)\right\| \left\|\nabla \widehat{L}(\theta^\star)\right\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star)} \tag{42}$$

$$\leq \mathsf{R}_\lambda^t \left\|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta^\star)\mathbf{H}_\lambda^{1/2}(\theta^\star)\right\| \left(\sum_{k=0}^{t-1} \left\|\lambda^k \widehat{\mathbf{H}}_\lambda^{-k}(\theta^\star)\right\| \right) \left\|\nabla \widehat{L}(\theta^\star)\right\|_{\mathbf{H}_\lambda^{-1}(\theta^\star)}, \tag{43}$$

$$\text{where} \quad \mathsf{R}_\lambda^t \overset{\text{def.}}{=} \prod_{k=1}^{t} \underline{\phi}^{-1}\left(\hat{\mathsf{t}}(\vartheta_\lambda^k - \theta^\star)\right). \tag{44}$$

Consider the prefactor of $\|\nabla \widehat{L}(\theta^\star)\|_{\mathbf{H}_\lambda^{-1}(\theta^\star)}$. We will bound $\|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta^\star)\mathbf{H}_\lambda^{1/2}(\theta^\star)\|$ by $\sqrt{2}$ with the same concentration argument as for the bias. The sum is more difficult to deal with. By computing the supremum of $\sigma \mapsto \lambda^k/(\sigma+\lambda)^k$ we would find that the first $\lfloor t/2 \rfloor$ terms have their maximum in 0. We would end up with a bound for the sum of the order of $t/2$. We rather use the simpler, if not optimal, following bound:

$$\sum_{k=0}^{t-1} \left\|\lambda^k \widehat{\mathbf{H}}_\lambda^{-k}(\theta^\star)\right\| \leq t.$$

It is suboptimal, but of the same order of an exact computation of the operator norm. Thus, we now have:

$$\left\|\vartheta_\lambda^t - \theta^\star\right\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star)} \leq \sqrt{2} t \mathsf{R}_\lambda^t \left\|\nabla \widehat{L}(\theta^\star)\right\|_{\mathbf{H}_\lambda^{-1}(\theta^\star)} \quad \text{when } \mathsf{H}_1 \tag{45}$$

Bounding the gradient $\|\nabla \widehat{L}(\theta^\star)\|_{\mathbf{H}_\lambda^{-1}(\theta^\star)}$. We use a plain Bernstein inequality to bound the gradient, as in Proposition 9:

$$\mathbf{H}_\lambda^{-1/2}(\theta^\star)\nabla \widehat{L}(\theta^\star) = \frac{1}{n}\sum_{k=1}^{n} \mathbf{H}_\lambda^{-1/2}(\theta^\star)\nabla \ell_{z_i}(\theta^\star).$$

We have

$$\sup_{z \in \mathsf{Supp}\,\rho} \|\nabla \ell_z(\theta^\star)\|_{\mathbf{H}_\lambda^{-1}(\theta^\star)} \leq \frac{\mathsf{B}_1^\star}{\sqrt{\lambda}},$$

$$\text{and} \quad \mathbb{E}_{z\sim\rho}\left[\|\nabla \ell_z(\theta^\star)\|_{\mathbf{H}_\lambda^{-1}(\theta^\star)}\right]^2 \overset{\text{def.}}{=} \mathsf{df}_\lambda.$$

With confidence $1-\delta$, we now have

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

We simplify this equation. Assuming

$$\mathsf{H}_3 : n \geq 2\frac{\mathsf{B}_1^{\star 2}}{\lambda\,\mathsf{df}_\lambda}\log\frac{2}{\delta}$$

we get the bound:

$$\left\|\nabla \widehat{L}(\theta^\star)\right\|_{\mathbf{H}_\lambda^{-1}(\theta^\star)} \leq 2\sqrt{2}\sqrt{\mathsf{df}_\lambda \frac{2\log 2/\delta}{n}}. \tag{46}$$

Gluing things together. All in all, we can glue together the inequalities in Eqs. (41), (45) and (46). We obtain:

$$\left\|\vartheta_\lambda^t - \theta^\star\right\|_{\mathbf{H}(\theta^\star)} \leq 4\sqrt{2} t \mathsf{R}_\lambda^t \sqrt{\frac{\mathsf{df}_\lambda}{n}} \cdot \sqrt{\log 2/\delta} \quad \text{when } \mathsf{H}_1 + \mathsf{H}_3$$

with confidence $1-2\delta$. We obtain the statement of the theorem by replacing δ with $\delta/2$, so that the result holds with confidence $1-\delta$. □

B.4 Conditions for non-exponentials prefactors

The prefactors P_λ^t and R_λ^t are hard to bound; they can depend exponentially on $\|\theta^\star\|$ in the worst case [1]. The purpose of this section is to give sufficient conditions on the number of samples n for those quantities to turn constant. The key quantity to compare to is the *Dikin radius* [1, 14].

Definition 7 (Dikin radius). *For $\theta \in \mathcal{H}$ and $\lambda > 0$, define $\mathsf{r}_\lambda(\theta)$ s.t*

$$\frac{1}{\mathsf{r}_\lambda(\theta)} = \sup_{z \in \mathsf{Supp}\,\rho} \sup_{g \in \phi(z)} \|g\|_{\mathbf{H}_\lambda^{-1}(\theta)}\,. \tag{47}$$

The inverse of the Dikin radius can be upper bounded by $R/\sqrt{\lambda}$. However, we prefer keeping bounds in $\mathsf{r}_\lambda^\star$. Indeed, they take into account the geometry of the loss function around the optimum, and are thus much more precise.

Note that in the following, we might be content with the *empirical* Dikin radius $\widehat{\mathsf{r}_\lambda(\theta)}$, ie. replacing ρ by $\hat{\rho}$ in the previous definition. So as not to ladden the notations and have something independant of the sampling, we use the fact that $\mathsf{Supp}\,\hat{\rho} \subset \mathsf{Supp}\,\rho$ to ensure that:

$$\frac{1}{\widehat{\mathsf{r}_\lambda(\theta)}} \leq \frac{1}{\mathsf{r}_\lambda(\theta)} \quad \text{and} \quad \hat{\mathsf{t}}(\cdot) \leq \mathsf{t}(\cdot).$$

Finally, we will use the following notation:

$$\mathsf{r}_\lambda^\star \overset{\text{def.}}{=} \mathsf{r}_\lambda(\theta^\star). \tag{48}$$

B.4.1 Pefactor of the variance

We first proceed with the prefactor of the variance R_λ^t.

Proposition 4 (Constant prefactor for the variance). *The following condition:*

$$\mathsf{H}_4 : n \geq 8(et)^2 \left(4 \vee C^2 t^2\right) \frac{\mathsf{df}_\lambda}{\mathsf{r}_\lambda^\star} \log 2/\delta, \tag{49}$$

where $C \leq 0.8$ is a constant, is sufficient to guarantee that

$$\mathsf{R}_\lambda^t \overset{\text{def.}}{=} \prod_{k=1}^{t} \underline{\phi}^{-1}\left(\hat{\mathsf{t}}(\vartheta_\lambda^k - \theta^\star)\right) \leq e. \tag{50}$$

Proof.

A first bound. Note that:

$$\begin{aligned}
\mathsf{t}(\vartheta_\lambda^t - \theta^\star) &= \sup_{z \in \mathsf{Supp}\,\rho} \sup_{g \in \phi(z)} \left|g \cdot (\vartheta_\lambda^t - \theta^\star)\right| \\
&\leq \sup_{z \in \mathsf{Supp}\,\rho} \sup_{g \in \phi(z)} \|g\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star)} \left\|\vartheta_\lambda^t - \theta^\star\right\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star)},
\end{aligned}$$

which gives us a bound we will use multiple times:

$$\mathsf{t}(\vartheta_\lambda^t - \theta^\star) \leq \frac{\|\vartheta_\lambda^t - \theta^\star\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star)}}{\mathsf{r}_\lambda^\star}. \tag{51}$$

We simply used the definition of the Dikin radius in Eq. (47).

We now use an upper bound of the numerator, available in the proof of Theorem 3:

$$\begin{aligned}
&\mathsf{t}(\vartheta_\lambda^t - \theta^\star) \leq \mathsf{R}_\lambda^t \left[2\sqrt{2}t\sqrt{\log 2/\delta}\right] \sqrt{\frac{\mathsf{df}_\lambda}{n\mathsf{r}_\lambda^\star}} \\
\iff \quad &\mathsf{t}(\vartheta_\lambda^t - \theta^\star)\underline{\phi}\left(\mathsf{t}(\vartheta_\lambda^t - \theta^\star)\right) \leq \mathsf{R}_\lambda^{t-1} \left[2\sqrt{2}t\sqrt{\log 2/\delta}\right] \sqrt{\frac{\mathsf{df}_\lambda}{n\mathsf{r}_\lambda^\star}} \overset{\text{def.}}{=} X_{t-1}.
\end{aligned}$$

Now, using the fact that $x\underline{\phi}(x) = 1 - e^{-x}$, we get that

$$\begin{aligned}
\mathsf{t}(\vartheta_\lambda^t - \theta^\star) &\leq -\log(1 - X_{t-1}) \\
\mathsf{a}_t \overset{\text{def.}}{=} \underline{\phi}^{-1}\left(\mathsf{t}(\vartheta_\lambda^t - \theta^\star)\right) &\leq -X_{t-1}^{-1}\log(1 - X_{t-1}) \overset{\text{def.}}{=} h(X_{t-1}).
\end{aligned} \tag{52}$$

Recursion hypotheses. The idea is to ensure:

1. $X_{k-1} \le 1/2$ so that
$$h(X_{k-1}) \le 1 + CX_{k-1}$$
with C a numeric constant s.t $h(1/2) = 1 + C/2$, which implies that $C \le 0.8$. We are simply upper bounding h which is convex on $[0, 1/2]$.

2. $\mathsf{a}_k \le 1 + 1/t$ for all $k \le t$, so that we can have:
$$\begin{aligned}\mathsf{R}_\lambda^t &= \prod_{k=1}^t \mathsf{a}_k = \exp \sum_{k=1}^t \log(\mathsf{a}_k) \\ &\le \exp \sum_{k=1}^t \log(1 + 1/t) \le e.\end{aligned}$$

Recursion. Set $k = 1$. Then $\mathsf{R}_\lambda^0 = 1$ and to have
$$X_0 \le 1/2 \quad \text{that is} \quad \left[2\sqrt{2}t\sqrt{\log 2/\delta}\right]\sqrt{\frac{\mathsf{df}_\lambda}{n\mathsf{r}_\lambda^\star}} \le \frac{1}{2},$$
it is sufficient to have
$$n \ge N_0 \overset{\text{def.}}{=} 32t^2 \frac{\mathsf{df}_\lambda}{\mathsf{r}_\lambda^\star} \log 2/\delta.$$
We want to enforce
$$\mathsf{a}_1 \le 1 + 1/t.$$
A sufficient condition is
$$\begin{aligned}h(X_0) \le 1 + CX_0 \le 1 + 1/t \iff & X_0 \le 1/tC \\ \impliedby & \; n \ge N_0' \overset{\text{def.}}{=} 8t^4C^2 \frac{\mathsf{df}_\lambda}{\mathsf{r}_\lambda^\star} \log 2/\delta.\end{aligned}$$
Now, let $k < n$. Assume the two conditions hold at step $k - 1$. Then, $\mathsf{R}_\lambda^{k-1} \le e$ and
$$n \ge N_{k-1} \overset{\text{def.}}{=} 32(et)^2 \frac{\mathsf{df}_\lambda}{\mathsf{r}_\lambda^\star} \log 2/\delta \quad \text{implies} \quad X_{k-1} \le \frac{1}{2}.$$
Likewise,
$$n \ge N_{k-1}' \overset{\text{def.}}{=} 8t^4(Ce)^2 \frac{\mathsf{df}_\lambda}{\mathsf{r}_\lambda^\star} \log 2/\delta$$
gives
$$X_{k-1} \le 1/tC, \quad \text{so that} \quad \mathsf{a}_k \le 1 + 1/k.$$

Conclusion. All in all, requiring
$$\mathsf{H}_4 : n \ge 8(et)^2 \left(4 \vee C^2t^2\right) \frac{\mathsf{df}_\lambda}{\mathsf{r}_\lambda^\star} \log 2/\delta$$
is sufficient to have $\mathsf{R}_\lambda^k \le e$, for any $k \le t$. □

B.4.2 Pefactor of the bias

The prefactor of the bias can be treated similarly. The only difficulty comes from the large number of subcases. Remember from Theorem 2 that we have, with appropriate hypotheses,

$$\left\|\widehat{\theta}_\lambda^t - \vartheta_\lambda^t\right\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star)} \leq \mathsf{P}_\lambda^t \mathsf{T}(r,t)\lambda^s, \quad \text{with } s = (r+1/2) \wedge t. \tag{53}$$

Proposition 5 (Constant prefactor for the bias). *Assume* H_4 *and*

$$\mathsf{H}_5 : \lambda \leq \mathsf{L} \overset{\text{def.}}{=} \left[e^{t+2}\mathsf{T}(r,t)(2 \wedge Ct)\right]^{-1/(r+1/2\wedge 1)}.$$

Then

$$\mathsf{P}_\lambda^t \leq e^{t+2}.$$

Proof. The proof is almost identical to the proof of Proposition 4. Let us simply point out the differences. We will drop the dependance of T on r, t in the notation for simplicity.

A first bound. Here, we have that:

$$\mathsf{t}\left(\widehat{\theta}_\lambda^k - \vartheta_\lambda^k\right) \leq \frac{\left\|\widehat{\theta}_\lambda^k - \vartheta_\lambda^k\right\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star)}}{\mathsf{r}_\lambda^\star} \leq \frac{\mathsf{P}_\lambda^t \mathsf{T}\lambda^s}{\mathsf{r}_\lambda^\star}, \quad \text{with } s = r + 1/2 \wedge k. \tag{54}$$

We used Eq. (53) in the second inequality. Recall the definition

$$\mathsf{P}_\lambda^t \overset{\text{def.}}{=} \prod_{k=1}^t \underline{\phi}^{-1}\left(\hat{\mathsf{t}}(\widehat{\theta}_\lambda^k - \vartheta_\lambda^k)\right) e^{\hat{\mathsf{t}}(\vartheta_\lambda^k - \theta^\star)}.$$

Thanks to H_4, we have $\mathsf{R}_\lambda^t \leq e$. Specifically, noting that $\underline{\phi}^{-1}(x) \geq x$, we have from the proof of Proposition 4:

$$1 + 1/t \geq \underline{\phi}^{-1}\left(\mathsf{t}(\vartheta_\lambda^k - \theta^\star)\right) \geq \mathsf{t}(\vartheta_\lambda^k - \theta^\star) \implies \prod_{k=1}^t e^{\mathsf{t}(\vartheta_\lambda^k - \theta^\star)} \leq e^{t+1}.$$

Thus, we have that

$$\mathsf{P}_\lambda^k \leq \underbrace{\prod_{i=1}^k \underline{\phi}^{-1}\left(\mathsf{t}(\widehat{\theta}_\lambda^i - \vartheta_\lambda^i)\right)}_{\mathsf{Q}_\lambda^k} e^{t+1}.$$

Dividing both sides in the Eq. (54) with $\underline{\phi}^{-1}\left(\mathsf{t}(\widehat{\theta}_\lambda^k - \vartheta_\lambda^k)\right)$, we obtain that

$$\mathsf{t}\left(\widehat{\theta}_\lambda^k - \vartheta_\lambda^k\right) \underline{\phi}\left(\mathsf{t}(\widehat{\theta}_\lambda^k - \vartheta_\lambda^k)\right) \leq \frac{\mathsf{Q}_\lambda^{k-1}\mathsf{T}e^{t+1}\lambda^s}{\mathsf{r}_\lambda^\star},$$

and we can apply the same reasoning as for the variance. Using $t\underline{\phi}(t) = 1 - e^{-t}$, we have

$$\mathsf{a}_k \overset{\text{def.}}{=} \underline{\phi}^{-1}\left(\mathsf{t}(\widehat{\theta}_\lambda^k - \vartheta_\lambda^k)\right) \leq -X_{k-1}^{-1}\log(1 - X_{k-1})$$
$$X_{k-1} \overset{\text{def.}}{=} \mathsf{Q}_\lambda^{k-1}\mathsf{T}e^{t+1}\lambda^s/\mathsf{r}_\lambda^\star.$$

Recursion. We then do the exact same reasoning to the variance, that is require at each step $X_{k-1} \leq 1/2$ and $\mathsf{a}_k \leq 1 + 1/t$. Here, this amounts to require

$$\lambda \leq \mathsf{L}_s \overset{\text{def.}}{=} \left[e^{t+2}\mathsf{T}(2 \wedge Ct)\right]^{-1/s}.$$

The L_s is increasing with s. So

$$\forall k \leq t, \quad \mathsf{L}_s \leq \mathsf{L}_{r+1/2\wedge 1} \overset{\text{def.}}{=} \mathsf{L}, \quad \text{with } s = r + 1/2 \wedge k.$$

Table 2: Hypotheses needed to bound the bias and the variance, depending on the source condition parameter r.

Source condition	Bias	Variance	Numerical prefactors
$0 < r \leq 1/2$	H_1		
$1/2 < r < 1$	$\mathsf{H}_1 + \mathsf{H}_{1b}$	$\mathsf{H}_1 + \mathsf{H}_3$	$\mathsf{H}_4 + \mathsf{H}_5$
$r \geq 1$	$\mathsf{H}_1 + \mathsf{H}_2$		

Conclusion. Requiring

$$\mathsf{H}_5 : \lambda \leq \mathsf{L} \overset{\text{def.}}{=} \left[e^{t+2}\mathsf{T}(2 \wedge Ct)\right]^{-1/(r+1/2\wedge 1)}$$

is sufficient to ensure $\mathsf{Q}_\lambda^t \leq e$, so that $\mathsf{P}_\lambda^t \leq e^{t+2}$. □

B.5 Optimal rates for IT estimator

The bound on the bias and the variance holds if the number of samples is "high enough". The purpose of next proposition is to merge all these hypotheses together. Precisely, the hypotheses requires in each regime are summed up in Table 2.

Proposition 6 (Satisfying the hypotheses H_{1-5} with bounds on n and λ). *The following relations hold:*

$$n \geq \mathsf{N}_0 \overset{\text{def.}}{=} \frac{2}{\lambda}\left[12\mathsf{B}_2^\star \vee \frac{\mathsf{B}_1^{\star 2}}{\mathsf{df}_\lambda}\right]\log\frac{4}{\delta}\left[1 \vee \frac{4\mathsf{B}_2^\star}{\lambda}\right] \implies \mathsf{H}_1 + \mathsf{H}_3,$$

$$n \geq \mathsf{N}_{1/2} \overset{\text{def.}}{=} \frac{2}{\lambda}\left[12\mathsf{B}_2^\star \vee \frac{\mathsf{B}_1^{\star 2}}{\mathsf{df}_\lambda} \vee \frac{4\mathsf{B}_2^\star}{\lambda}\right]\log^2\frac{4}{\delta}\left[1 \vee \frac{4\mathsf{B}_2^\star}{\lambda}\right] \implies \mathsf{H}_1 + \mathsf{H}_{1b} + \mathsf{H}_3,$$

$$n \geq \mathsf{N}_1 \overset{\text{def.}}{=} \frac{2}{\lambda}\left[12\mathsf{B}_2^\star \vee \frac{\mathsf{B}_1^{\star 2}}{\mathsf{df}_\lambda} \vee \lambda \vee \left(\frac{2\mathsf{B}_2^\star(t-1/2)^r}{\lambda^{r-1/2}}\right)^2\right]\log\frac{4}{\delta}\left[1 \vee \frac{4\mathsf{B}_2^\star}{\lambda}\right] \implies \mathsf{H}_1 + \mathsf{H}_2 + \mathsf{H}_3,$$

$$n \geq \overline{\mathsf{N}} \overset{\text{def.}}{=} 8(et)^2\left(4 \vee C^2t^2\right)\frac{\mathsf{df}_\lambda}{\mathsf{r}_\lambda^\star}\log 2/\delta \implies \mathsf{H}_4,$$

$$\lambda \leq \mathsf{L} \overset{\text{def.}}{=} \left[e^{t+2}\mathsf{T}(r,t)(2 \wedge Ct)\right]^{-1/(r+1/2\wedge 1)} \implies \mathsf{H}_5.$$

Recall that T *is defined in Theorem 2. Moreover, having*

$$\lambda = Kn^{-\frac{\alpha}{1+\alpha(2r+1)}}$$

with K *a constant not depending on* n *make all these conditions possible.*

Proof. The expression of the constant boils down to taking the maximum of each expression. Recall that:

- $\mathsf{H}_1, \mathsf{H}_{1b}, \mathsf{H}_2$ are defined in Theorem 2;
- H_3 is defined in Theorem 3;
- H_4 is defined in Proposition 4;
- H_5 is defined in Proposition 5.

About the fact they are attainable, we need to check that the power of n is smaller than 1, in order that

$$\exists n, \quad n \geq \mathsf{N}(\lambda) \quad \text{and} \quad \lambda = Kn^{-\frac{\alpha}{1+\alpha(2r+1)}},$$

with $\mathsf{N}(\lambda)$ chosen among $\left\{\mathsf{N}_0, \mathsf{N}_{1/2}, \mathsf{N}_1, \overline{\mathsf{N}}\right\}$. In the following, $\sim$ denotes equality up to log factors between two quantities. Recall that $\mathsf{s}\lambda^{-1/\alpha} \leq \mathsf{df}_\lambda \leq \mathsf{S}\lambda^{-1/\alpha}$, and assume

$$\lambda = Kn^{-\frac{\alpha}{1+\alpha(2r+1)}}$$

for some K a positive constant.

- When $r \leq 1/2$, $\mathsf{N}_0 \sim \lambda^{-(1+1/\alpha)} \sim n^{\frac{1+\alpha}{1+\alpha(2r+1)}}$ and $\frac{1+\alpha}{1+\alpha(2r+1)} < 1$.
- When $1/2 < r \leq 1$, $\mathsf{N}_{1/2} \sim \lambda^{-2} \sim n^{\frac{2\alpha}{1+\alpha(2r+1)}}$ and $\frac{2\alpha}{1+\alpha(2r+1)} < 1$ as $\alpha(2r+1) > 2\alpha$.
- Finally, when $r > 1$, $\mathsf{N}_1 \sim \lambda^{-2r} \sim n^{\frac{\alpha(2r)}{1+\alpha(2r+1)}}$ and $\alpha(2r) < 1 + \alpha(2r+1)$.
- For $\overline{\mathsf{N}}$, use the upper bound $\frac{\mathsf{df}_\lambda}{\mathsf{r}^\star_\lambda} \leq \mathsf{S}R\lambda^{-(1/\alpha+1/2)}$. Then, $\overline{\mathsf{N}} \sim n^{\frac{\alpha+2}{2(1+\alpha(2r+1))}}$ and $\frac{\alpha+2}{2(1+\alpha(2r+1))} = 1 - \frac{\alpha(4r+1)}{\alpha(4r+2)+2} \leq 1$.

□

Having bounded the bias and the variance of the estimator, we are now in shape to state our main result.

Theorem 4 (Optimal rates of IT estimator). *Let $\delta \in (0,1]$, $\lambda > 0$ and choose n so that H_3 and the following holds:*

$$\begin{aligned} \mathsf{H}_1 \quad & \text{if } r \leq 1/2, \\ \mathsf{H}_1 + \mathsf{H}_{1b} \quad & \text{if } 1/2 < r \leq 1, \\ \mathsf{H}_1 + \mathsf{H}_2 \quad & \text{if } r > 1. \end{aligned}$$

Then we can bound the excess risk with probability greater than $1-\delta$ as

$$L(\widehat{\theta}^t_\lambda) - L(\theta^\star) \leq \mathsf{C}_{\text{bias}}\lambda^{2s} + \mathsf{C}_{\text{var}}\frac{\mathsf{df}_\lambda}{n}, \quad \text{with } s = (r+1/2) \wedge t.$$

If we further assume that the capacity condition holds and that the estimator does not saturate, that is $t \geq r + 1/2$, then setting

$$\lambda = \left[\left(\frac{\mathsf{C}_{\text{var}}}{\mathsf{C}_{\text{bias}}}\right)^2 \mathsf{S}\right]^{\frac{\alpha}{1+\alpha(2r+1)}} n^{-\frac{\alpha}{1+\alpha(2r+1)}}$$

makes the following holds with confidence 2δ:

$$L(\widehat{\theta}^t_\lambda) - L(\theta^\star) \leq 2\left[\left(\frac{\mathsf{C}_{\text{var}}}{\mathsf{C}_{\text{bias}}}\right)^2 \mathsf{S}\right]^{\frac{\alpha(2r+1)}{1+\alpha(2r+1)}} n^{-\frac{\alpha(2r+1)}{1+\alpha(2r+1)}},$$

where the constants $\mathsf{C}_{\text{bias}}, \mathsf{C}_{\text{var}}$ are bounded by quantities only depending on $r, t, \mathsf{B}^\star_2, \delta$ as soon as hypotheses H_4 and H_5 are satisfied.

Proof.

Decomposition of the risk. We use the decomposition of the risk:

$$\begin{aligned} L(\widehat{\theta}^t_\lambda) - L(\theta^\star) &\leq \Psi\left(\mathsf{t}(\widehat{\theta}^t_\lambda - \theta^\star)\right) \left\|\widehat{\theta}^t_\lambda - \theta^\star\right\|^2_{\mathbf{H}(\theta^\star)} \\ &\leq 2\Psi\left(\mathsf{t}(\widehat{\theta}^t_\lambda - \vartheta^t_\lambda) + \mathsf{t}(\vartheta^t_\lambda - \theta^\star)\right)\left[\left\|\widehat{\theta}^t_\lambda - \vartheta^t_\lambda\right\|^2_{\mathbf{H}(\theta^\star)} + \left\|\vartheta^t_\lambda - \theta^\star\right\|^2_{\mathbf{H}(\theta^\star)}\right], \end{aligned}$$

where we applied Proposition 3, and used that $(a+b)^2 \leq 2(a^2+b^2)$.

Bias and variance prefactors. We introduce the following quantities:

$$\begin{aligned} \mathsf{C}^2_{\text{bias}} &= 2\Psi\left(\mathsf{t}(\widehat{\theta}^t_\lambda - \vartheta^t_\lambda) + \mathsf{t}(\vartheta^t_\lambda - \theta^\star)\right)\mathsf{T}(r,t)\mathsf{P}^t_\lambda, \\ \mathsf{C}^2_{\text{var}} &= 2\Psi\left(\mathsf{t}(\widehat{\theta}^t_\lambda - \vartheta^t_\lambda) + \mathsf{t}(\vartheta^t_\lambda - \theta^\star)\right)\left[4\sqrt{2}t\mathsf{R}^t_\lambda\sqrt{\log 2/\delta}\right], \end{aligned}$$

where $\mathsf{P}_\lambda^t, \mathsf{R}_\lambda^t$ are defined in Theorems 2 and 3 respectively. Then the bound on the excess risk reads:

$$L(\widehat{\theta}_\lambda^t) - L(\theta^\star) \leq \mathsf{C}_{\text{bias}} \begin{cases} \lambda^{2r+1} & \text{if } r + 1/2 \leq t \\ \lambda^{2t} & \text{otherwise} \end{cases} + \mathsf{C}_{\text{var}} \frac{\mathsf{df}_\lambda}{n}$$

with confidence 2δ with the appropriate hypothesis $\mathsf{H}_1, \mathsf{H}_2$ or H_3, depending on r, see Table 2.

Optimal λ. Further assume $t \geq r + 1/2$ and the capacity condition holds with parameters S, α. Then, setting:

$$\lambda^{\frac{1+\alpha(2r+1)}{\alpha}} = \left(\frac{\mathsf{C}_{\text{var}}}{\mathsf{C}_{\text{bias}}}\right)^2 \frac{\mathsf{S}}{n} \iff \lambda = \left[\left(\frac{\mathsf{C}_{\text{var}}}{\mathsf{C}_{\text{bias}}}\right)^2 \mathsf{S}\right]^{\frac{\alpha}{1+\alpha(2r+1)}} n^{-\frac{\alpha}{1+\alpha(2r+1)}},$$

makes the following bound holds with probability $1 - 2\delta$:

$$L(\widehat{\theta}_\lambda^t) - L(\theta^\star) \leq 2 \left[\left(\frac{\mathsf{C}_{\text{var}}}{\mathsf{C}_{\text{bias}}}\right)^2 \mathsf{S}\right]^{\frac{\alpha(2r+1)}{1+\alpha(2r+1)}} n^{-\frac{\alpha(2r+1)}{1+\alpha(2r+1)}}.$$

Explicit prefactors. Assume Hyp. H_4 and H_5 hold, and $\lambda \leq \mathsf{B}_2^\star$. Then the quantities $\mathsf{C}_{\text{bias}}, \mathsf{C}_{\text{var}}$ only depend on r, t up to the term $\Psi\left(\mathsf{t}(\widehat{\theta}_\lambda^t - \vartheta_\lambda^t) + \mathsf{t}(\vartheta_\lambda^t - \theta^\star)\right)$. Noting that:

$$1 + 1/t \geq \underline{\phi}^{-1}(x) \geq x \quad \text{implies} \quad 1 + 1/t \geq x,$$

and Ψ is increasing we have

$$\Psi\left(\mathsf{t}(\widehat{\theta}_\lambda^t - \vartheta_\lambda^t) + \mathsf{t}(\vartheta_\lambda^t - \theta^\star)\right) \leq \Psi(4) \leq 4.$$

In the end $\mathsf{C}_{\text{bias}}, \mathsf{C}_{\text{var}}$ only depend on r, t and the parameters of the problem:

$$\begin{aligned} \mathsf{C}_{\text{bias}}^2 &\leq 8\mathsf{T}(r,t)e^{t+2} \\ \mathsf{C}_{\text{var}}^2 &\leq 32te\sqrt{\log 2/\delta}, \end{aligned} \tag{55}$$

where $\mathsf{T}(r,t)$ was introduced previously in Theorem 2:

$$\mathsf{T}(r,t) = \begin{cases} \|v\| \left(1 \vee (\mathsf{B}_2^\star + \lambda)\right) 2^r & \text{if } r \leq 1, \\ \|v\| \frac{w(r)+r}{(t-1/2)^r} & \text{if } r > 1 \text{ and } r + 1/2 < t, \\ \|v\| \frac{w(r)}{(t-1/2)^r} + \mathsf{B}_2^{\star r-t+1/2} & \text{if } r > 1 \text{ and } r + 1/2 \geq t. \end{cases}$$

Proof of Theorem 1 in the paper. We took the maximum on the lower bounds on the samples to simplify the result in the main body. Simply define:

$$\mathsf{N} = \overline{\mathsf{N}} \vee \begin{cases} \mathsf{N}_0 & \text{if } r \leq 1/2 \\ \mathsf{N}_{1/2} & \text{if } 1/2 < r < 1 \\ \mathsf{N}_1 & \text{otherwise} \end{cases}$$

$$\text{and} \quad \mathsf{C}_{\text{risk}} = \left[\left(\frac{\mathsf{C}_{\text{var}}}{\mathsf{C}_{\text{bias}}}\right)^2 \mathsf{S}\right]^{\frac{\alpha}{1+\alpha(2r+1)}}.$$

Again, we highlight that the observation made in Proposition 6 is key to ensure that these constants are attainable, in the sense that they are not in contradiction with the optimal rate in n. □

C Statistical guarantees with inexact solvers

This section is devoted to finding a rule on the tolerance enforced at each step of the proximal sequence. Given a tolerance ϵ, we look for $\bar{\epsilon}_1, \ldots, \bar{\epsilon}_n$, the tolerance to ensure at each proximal step. It leads to Proposition 1 in the main body of the article.

Important remark on the notation. So as to simplify the notation, we drop the hat $\widehat{\ }$ on the loss function. That is, we simply take a loss function L assumed to be GSC. In practice, this function is of course the empirical loss $\widehat{L}$. We denote with a bar $\bar{}$ the quantity we compute at each step, and whose aim is to approximate the estimator of L.

Tikhonov regularization. For a GSC function L, we define:

$$\begin{aligned}
\theta_\mu^1 = \mathsf{prox}_{L/\mu}(0) = \arg\min_\theta L_\mu(\theta), \qquad & L_\mu(\theta) \stackrel{\text{def.}}{=} L(\theta) + \frac{\mu}{2}\|\theta\|^2 \\
\theta_\mu^{k+1} = \mathsf{prox}_{L/\mu}(\theta_\lambda^k) = \arg\min_\theta L_\mu^{\lambda,k}(\theta), \qquad & L_\mu^{\lambda,k}(\theta) \stackrel{\text{def.}}{=} L(\theta) + \frac{\mu}{2}\left\|\theta - \theta_\lambda^k\right\|^2 \\
\overline{\theta}_\mu^{k+1} = \mathsf{prox}_{L/\mu}(\overline{\theta}_\lambda^k) = \arg\min_\theta \overline{L}_\mu^{\lambda,k}(\theta), \qquad & \overline{L}_\mu^{\lambda,k}(\theta) \stackrel{\text{def.}}{=} L(\theta) + \frac{\mu}{2}\left\|\theta - \overline{\theta}_\lambda^k\right\|
\end{aligned}$$

so that we can refer easily to the function which has to be minimized when evaluating the proximal operator.

C.1 Definitions

We use the following notations for the *Newton decrement*:

- The theoretical quantity writes:

$$\nu_\mu^{\lambda,k}(\theta) = \left\|\nabla L_\mu^{\lambda,k}(\theta)\right\|_{\mathbf{H}_\lambda^{-1}(\theta)} = \left\|\nabla L(\theta) + \mu(\theta - \theta_\lambda^{\lambda,k})\right\|_{\mathbf{H}_\mu^{-1}(\theta)};$$

- The normalized Newton decrement is defined with:

$$\widetilde{\nu}_\lambda^{k-1}(\theta) = \frac{\nu_\lambda^{k-1}(\theta)}{\mathsf{r}_\lambda(\theta)};$$

- The quantity we compute is:

$$\overline{\nu}_\mu^{\lambda,k}(\theta) = \left\|\nabla L_\mu^{\lambda,k}(x)\right\|_{\mathbf{H}_\lambda^{-1}(\theta)} = \left\|\nabla L(\theta) + \mu(\theta - \overline{\theta}_\lambda^{\lambda,k})\right\|_{\mathbf{H}_\mu^{-1}(\theta)}.$$

We also recall some definition and properties. R is defined with

$$R = \sup_{z \in \mathsf{Supp}\,\rho} \sup_{g \in \phi(z)} \|g\| \quad \text{so that} \quad \mathop{\mathsf{r}}_\lambda(\theta) \geq R/\sqrt{\lambda}$$

and $\mathsf{r}_\lambda(\theta)$ is given in Definition 7. The Dikin ellipsoid, as in [14], reads

$$\forall \mathsf{c} \in \mathbb{R}, \quad \mathsf{D}_\lambda^{k-1}(\mathsf{c}) = \left\{\theta \in \mathcal{H}; \widetilde{\nu}_\lambda^{k-1}(\theta) \leq \mathsf{c}\right\}.$$

We provide a short lemma to show how controlling the *normalized* Newton decrement enables to control quantities depending on t.

Lemma 3 (Localization properties with the Newton decrement). *Let $k \leq t$ and $\mathsf{c} > 0$. Assume*

$$\overline{\theta}_\lambda^k \in \mathsf{D}_\lambda^{k-1}(\mathsf{c}), \quad \textit{that is} \quad \widetilde{\nu}_\lambda^{k-1}\left(\overline{\theta}_\lambda^k\right) \leq \mathsf{c}.$$

Then, we have

$$\underline{\phi}^{-1}\left(\mathsf{t}(\overline{\theta}_\lambda^k - \theta_\lambda^k)\right) \leq -\frac{1}{\mathsf{c}}\log(1-\mathsf{c}) \stackrel{\text{def.}}{=} \kappa_{\mathsf{c}}. \tag{56}$$

Proof. The proof combines inequalities we already used, replacing the normalized gradient with the Newton decrement. Recall Eq. (51), which states that

$$\mathsf{t}(\overline{\theta}_\lambda^k - \theta_\lambda^k) \leq \frac{\left\|\overline{\theta}_\lambda^k - \theta_\lambda^k\right\|_{\widehat{\mathbf{H}}_\lambda(\overline{\theta}_\lambda^k)}}{\mathsf{r}_\lambda(\overline{\theta}_\lambda^k)}. \tag{57}$$

Using the lower bound on gradient of Lemma 2 gives

$$\left\|\overline{\theta}_\lambda^k - \theta_\lambda^k\right\|_{\widehat{\mathbf{H}}_\lambda(\overline{\theta}_\lambda^k)} \leq \underline{\phi}^{-1}(\mathsf{t}(\overline{\theta}_\lambda^k - \theta_\lambda^k)) \left\|\nabla L_\lambda^{k-1}(\overline{\theta}_\lambda^k)\right\|_{\mathbf{H}_\lambda^{-1}(\overline{\theta}_\lambda^k)},$$

and using the definition of the Newton decrement in the previous equation gives

$$\left\|\overline{\theta}_\lambda^k - \theta_\lambda^k\right\|_{\widehat{\mathbf{H}}_\lambda(\overline{\theta}_\lambda^k)} \leq \underline{\phi}^{-1}(\mathsf{t}(\overline{\theta}_\lambda^k - \theta_\lambda^k))\nu_\lambda^{k-1}\left(\overline{\theta}_\lambda^k\right). \tag{58}$$

Plugging Eq. (58) in Eq. (57) implies

$$\underline{\phi}(\mathsf{t}(\overline{\theta}_\lambda^k - \theta_\lambda^k))\,\mathsf{t}(\overline{\theta}_\lambda^k - \theta_\lambda^k) \leq \frac{\nu_\lambda^{k-1}\left(\overline{\theta}_\lambda^k\right)}{\mathsf{r}_\lambda(\overline{\theta}_\lambda^k)} \overset{\text{def.}}{=} \widetilde{\nu}_\lambda^{k-1}\left(\overline{\theta}_\lambda^k\right).$$

Use the fact that $\underline{\phi}(x)x = 1-e^{-x}$ combined with the definition of the normalized Newton decrement to simplify both sides of the previous equation. After simplification, we obtain

$$\mathsf{t}(\overline{\theta}_\lambda^k - \theta_\lambda^k) \leq -\log\left(1 - \widetilde{\nu}_\lambda^{k-1}\left(\overline{\theta}_\lambda^k\right)\right).$$

Apply $\underline{\phi}^{-1}$ on both side to have

$$\underline{\phi}^{-1}\left(\mathsf{t}(\overline{\theta}_\lambda^k - \theta_\lambda^k)\right) \leq -\widetilde{\nu}_\lambda^{k-1}\left(\overline{\theta}_\lambda^k\right)^{-1}\log\left(1 - \widetilde{\nu}_\lambda^{k-1}\left(\overline{\theta}_\lambda^k\right)\right),$$

and the conclusion follows with the fact that this is an increasing function of the normalized Newton decrement, which is upper bounded by c. □

This lemma will be useful in the following derivation, and provide some intuition on GSC loss function.

Remark 6. *Intuition for GSC loss function.* The purpose of working with Generalized self-concordant loss functions is to be able to control the deviation of the function with their local quadratic approximation. For $\theta \in \mathcal{H}$, Lemma 3 gives us that we can bound quantities depending on t in the inequalities of GSC loss functions of Proposition 3. When θ is deep into D_λ^{k-1}, then $\mathsf{t} \to 1/2$ and the bounds of Proposition 3 are tight. On the contrary, when θ leaves this ellipsoid, the upper bound diverges exponentially to infinity while the lower bound goes exponentially to 0, making the deviation from the quadratic approximation very loose.

To conclude, a GSC function with high R has small Dikin ellipsoids, and is far from its quadratic approximation. On the contrary, a GSC function with low R will be close to its quadratic approximation; the Dikin ellipsoid is large. The extreme case is obtained when ℓ is the square loss. Then, $\phi = \{0\}$, so $R = 0$, and the Dikin ellipsoid spans the whole space for any $\theta \in \mathcal{H}$. This implies *e.g.* that the lower and upper bound on the gradient matches, making the quadratic approximation tight.

C.2 Error propagation

In this section, we give a sufficient condition for achieving an ϵ error on a sequence of proximal operators. Indeed, we aim at minimizing L_λ^{t-1}, but we do not have access to this function; only to its approximation $\overline{L}_\lambda^{t-1}$. Relating both is the purpose of the next result.

Proposition 7 (Error propagation with proximal sequence). *Let* $\mathsf{c} > 0$. *Assume that you can solve each subproblem with precision* $\bar{\epsilon}_k$ *and that you have a guarantee on the exact normalized decrement:*

$$\forall k \in \{1, \dots, t\}, \quad \begin{cases} \overline{\nu}_\lambda^{k-1}\left(\overline{\theta}_\lambda^k\right) \leq \bar{\epsilon}_k \\ \overline{\theta}_\lambda^k \in \mathsf{D}_\lambda^{k-1}(\mathsf{c}) \iff \widetilde{\nu}_\lambda^{k-1}\left(\overline{\theta}_\lambda^k\right) \leq \mathsf{c} \end{cases}$$

Then requiring:

$$\forall k \in \{1, \dots, t\}, \quad \bar{\epsilon}_k = \epsilon\frac{\kappa_\mathsf{c}^{k-t}}{t}$$

with $\kappa_\mathsf{c} = -1/\mathsf{c}\log(1-\mathsf{c})$ *suffice to achieve an error* ϵ*:*

$$\nu_\lambda^{t-1}\left(\overline{\theta}_\lambda^t\right) \leq \epsilon.$$

We can replace the condition $\overline{\theta}_\lambda^k \in \mathsf{D}_\lambda^{k-1}(\mathsf{c})$ *with* $\epsilon \leq \mathsf{c}\sqrt{\lambda}/R$.

Proof. Let us track the error step by step. Denote by ϵ_k the Newton decrement of the exact function at each step:

$$\forall k, \quad \epsilon_k \overset{\text{def.}}{=} \nu_\lambda^{k-1}\left(\overline{\theta}_\lambda^k\right).$$

Consider the following decomposition at step k:

$$\begin{aligned}
\nu_\lambda^{k-1}\left(\overline{\theta}_\lambda^k\right) &= \left\|\nabla L(\overline{\theta}_\lambda^k) + \lambda\left(\overline{\theta}_\lambda^k - \theta_\lambda^{k-1}\right)\right\|_{\mathbf{H}_\lambda^{-1}(\overline{\theta}_\lambda^k)} \\
&\leq \left\|\nabla L(\overline{\theta}_\lambda^k) + \lambda\left(\overline{\theta}_\lambda^k - \overline{\theta}_\lambda^{k-1}\right)\right\|_{\mathbf{H}_\lambda^{-1}(\overline{\theta}_\lambda^k)} + \left\|\lambda\left(\overline{\theta}_\lambda^{k-1} - \theta_\lambda^{k-1}\right)\right\|_{\mathbf{H}_\lambda^{-1}(\overline{\theta}_\lambda^k)} \\
&\leq \overline{\nu}_\lambda^{k-1}\left(\overline{\theta}_\lambda^k\right) + \lambda\left\|\mathbf{H}_\lambda^{-1/2}(\overline{\theta}_\lambda^k)\mathbf{H}_\lambda^{-1/2}(\theta_\lambda^k)\right\|\left\|\overline{\theta}_\lambda^{k-1} - \theta_\lambda^{k-1}\right\|_{\mathbf{H}_\lambda(\overline{\theta}_\lambda^{k-1})} \\
&\leq \overline{\nu}_\lambda^{k-1}\left(\overline{\theta}_\lambda^k\right) + \underline{\phi}^{-1}\left(\mathsf{t}(\overline{\theta}_\lambda^{k-1} - \theta_\lambda^{k-1})\right)\nu_\lambda^{k-2}\left(\overline{\theta}_\lambda^{k-1}\right)
\end{aligned}$$

In the last inequality we used that $\left\|\mathbf{H}_\lambda^{-1/2}(\overline{\theta}_\lambda^k)\mathbf{H}_\lambda^{-1/2}(\theta_\lambda^k)\right\| \leq 1/\lambda$ and the relation between the distance in Hessian's norm and the Newton decrement of Eq. (58). Introducing the notation with epsilon, the last line is by definition

$$\epsilon_k \leq \bar{\epsilon}_k + \underline{\phi}^{-1}\left(\mathsf{t}(\overline{\theta}_\lambda^{k-1} - \theta_\lambda^{k-1})\right)\epsilon_{k-1}. \tag{59}$$

The first term $\bar{\epsilon}_k$ is the error we can control at each step whereas ϵ_{k-1} is the error of interest which increases with k. Using the fact that $\overline{\theta}_\lambda^{k-1} \in \mathsf{D}_\lambda^{k-2}(\mathsf{c})$, we have

$$\underline{\phi}^{-1}\left(\mathsf{t}(\overline{\theta}_\lambda^{k-1} - \theta_\lambda^{k-1})\right) \leq -\frac{1}{\mathsf{c}}\log(1-\mathsf{c}) \overset{\text{def.}}{=} \kappa_\mathsf{c}$$

thanks to Lemma 3. Thus, Eq. (59) becomes

$$\epsilon_k \leq \bar{\epsilon}_k + \kappa_\mathsf{c}\epsilon_{k-1}. \tag{60}$$

This being valid for all $i \leq k$ and since $\overline{\nu}_\lambda^0\left(\overline{\theta}_\lambda^1\right) = \nu_\lambda^0\left(\overline{\theta}_\lambda^1\right)$ we obtain that

$$\epsilon_k \leq \sum_{i=1}^{k} \bar{\epsilon}_i \kappa_\mathsf{c}^{k-i}. \tag{61}$$

Now plug the assumption of the proposition, namely that each problem is solved with precision

$$\bar{\epsilon}_k = \epsilon\frac{\kappa_\mathsf{c}^{k-t}}{t}$$

and use Eq. (61) at step t to obtain

$$\epsilon_t \leq \sum_{i=1}^{t} \kappa_\mathsf{c}^{i-t}\kappa_\mathsf{c}^{t-i}\frac{\epsilon}{t} = \epsilon. \tag{62}$$

Replacing $\overline{\theta}_\lambda^k \in \mathsf{D}_\lambda^{k-1}(\mathsf{c})$ with $\epsilon \leq \mathsf{c}\sqrt{\lambda}/R$. Let $k \geq 1$. Then, having

$$\overline{\theta}_\lambda^k \in \mathsf{D}_\lambda^{k-1}(\mathsf{c})$$

amounts by definition to have

$$\widetilde{\nu}_\lambda^{k-1}\left(\overline{\theta}_\lambda^k\right) \leq \mathsf{c},$$

which is also equivalent to

$$\nu_\lambda^{k-1}\left(\overline{\theta}_\lambda^k\right) \leq \mathsf{c}\,\underset{\lambda}{\mathsf{r}}(\overline{\theta}_\lambda^k).$$

We can use the crude lower bound $\mathsf{r}_\lambda(\cdot) \geq \sqrt{\lambda}/R$. Thus, the following implication holds:

$$\epsilon_k \overset{\text{def.}}{=} \nu_\lambda^{k-1}\left(\overline{\theta}_\lambda^k\right) \leq \mathsf{c}\frac{\sqrt{\lambda}}{R} \implies \overline{\theta}_\lambda^k \in \mathsf{D}_\lambda^{k-1}(\mathsf{c}). \tag{63}$$

Now, assume $\epsilon \leq \mathsf{c}\sqrt{\lambda}/R$. Then, we have that

$$\epsilon_1 = \bar{\epsilon}_1 = \epsilon \frac{\kappa_{\mathsf{c}}^{1-t}}{t} \leq \mathsf{c}\sqrt{\lambda}/R \frac{\kappa_{\mathsf{c}}^{1-t}}{t},$$

which gives

$$\epsilon_1 \leq \mathsf{c}\sqrt{\lambda}/R,$$

which implies $\overline{\theta}_\lambda^1 \in \mathsf{D}_\lambda^0(\mathsf{c})$ following Eq. (63). Then Eq. (60) holds with $k = 2$:

$$\epsilon_2 \leq \bar{\epsilon}_2 + \kappa_{\mathsf{c}} \epsilon_1.$$

For bigger k, proceed by induction. Let $k < t$ and assume for any $i < k$ that

$$\epsilon_{i+1} \leq \bar{\epsilon}_{i+1} + \kappa_{\mathsf{c}} \epsilon_i.$$

Then, we have that

$$\epsilon_k \leq \sum_{i=1}^{k} \bar{\epsilon}_i \kappa_{\mathsf{c}}^{k-i}$$

which gives the following bound, thanks to the assumption on ϵ and the $\bar{\epsilon}_i$:

$$\epsilon_k \leq \sum_{i=1}^{k} \epsilon \frac{\kappa_{\mathsf{c}}^{i-t}}{t} \kappa_c^{k-i} \leq \mathsf{c}\sqrt{\lambda}/R.$$

This implies $\overline{\theta}_\lambda^k \in \mathsf{D}_\lambda^{k-1}(\mathsf{c})$ following Eq. (63), and Eq. (60) holds at step $k+1$. Thus the induction hypothesis holds for all k and the conclusion of Eq. (62) holds.

Proof of Proposition 1. This result is a direct application of the previous one, where we set $\mathsf{c} = 1/2$. □

We see that the requirement $\epsilon \leq \mathsf{c}\sqrt{\lambda}/R$ is simply to ensures that a bound on the Newton decrement $\nu_\lambda^{k-1}\left(\overline{\theta}_\lambda^k\right)$ translates to a bound on the *normalized* Newton decrement $\widetilde{\nu}_\lambda^{k-1}\left(\overline{\theta}_\lambda^k\right)$ *via* the crude bound on the Dikin radius $\mathsf{r}_\lambda(\overline{\theta}_\lambda^k) \geq R/\sqrt{\lambda}$. Thus, the requirement on ϵ can be dropped if we assume $\overline{\theta}_\lambda^k \in \mathsf{D}_\lambda^{k-1}(\mathsf{c})$. Such condition is enforced in solver such as the one developed in [14].

Finally, we put in application this result with next proposition, which gives a bound on the excess risk with inexact solver.

Proposition 8 (Bound on the excess risk with inexact solver). *Assume that:*

- *the requirement of Proposition 7 hold;*
- *the requirement of Theorem 4 hold, namely* H_{1-5}*;*

The first is an hypothesis on the optimization procedure*, while the second in an hypothesis on the* statistics *of the learning task. Then, denoting* $\overline{\theta}_\lambda^t$ *the approximation of* $\widehat{\theta}_\lambda^t$ *as defined in Proposition 7, we have the following bound on the excess risk:*

$$L(\overline{\theta}_\lambda^t) - L(\theta^\star) \leq \mathsf{C}_{\text{bias}} \lambda^{2s} + \mathsf{C}_{\text{var}} \frac{\mathsf{df}_\lambda}{n} + \mathsf{E}_{\mathsf{c}}\, \epsilon, \quad s = (r + 1/2) \wedge t,$$

with:

$$\mathsf{E}_{\mathsf{c}} \stackrel{\text{def.}}{=} 4\Psi(4 - \log(1 - \mathsf{c})) \frac{e^4}{1 - \mathsf{c}} \kappa_{\mathsf{c}}^2, \quad \textit{e.g.} \quad \mathsf{E}_{1/2} \leq 4.3 \cdot 10^3.$$

Proof. The proof boils down to combining the statistical results held in Theorem 4 with the optimization result of Proposition 7. Begin by writing

$$\begin{aligned} L(\widehat{\theta}_\lambda^t) - L(\theta^\star) &\leq \Psi\left(\mathsf{t}(\overline{\theta}_\lambda^t - \theta^\star)\right) \left\|\overline{\theta}_\lambda^t - \theta^\star\right\|_{\mathbf{H}(\theta^\star)}^2 \\ &\leq 2\Psi\left(\mathsf{t}(\widehat{\theta}_\lambda^t - \overline{\theta}_\lambda^t) + \mathsf{t}(\widehat{\theta}_\lambda^t - \theta^\star)\right) \left[\left\|\widehat{\theta}_\lambda^t - \overline{\theta}_\lambda^t\right\|_{\mathbf{H}(\theta^\star)}^2 + \left\|\widehat{\theta}_\lambda^t - \theta^\star\right\|_{\mathbf{H}(\theta^\star)}^2\right]. \end{aligned}$$

We know how to handle the statistical term $\left\|\widehat{\theta}_\lambda^t - \theta^\star\right\|^2_{\mathbf{H}(\theta^\star)}$.

Bound on $\mathsf{t}(\widehat{\theta}_\lambda^t - \overline{\theta}_\lambda^t)$. As in the beginning of the proof of Proposition 4, we write:

$$\begin{aligned}
\mathsf{t}(\widehat{\theta}_\lambda^t - \overline{\theta}_\lambda^t) &\leq \frac{1}{\mathsf{r}_\lambda(\overline{\theta}_\lambda^t)} \left\|\overline{\theta}_\lambda^t - \widehat{\theta}_\lambda^t\right\|_{\mathbf{H}_\lambda(\overline{\theta}_\lambda^t)} \\
&\leq \frac{1}{\mathsf{r}_\lambda(\overline{\theta}_\lambda^t)} \underline{\phi}^{-1}\left(\mathsf{t}(\overline{\theta}_\lambda^t - \widehat{\theta}_\lambda^t)\right) \left\|\nabla \widehat{L}_\lambda^{t-1}(\overline{\theta}_\lambda^t)\right\|_{\mathbf{H}_\lambda^{-1}(\overline{\theta}_\lambda^t)} \\
&= \underline{\phi}^{-1}\left(\mathsf{t}(\overline{\theta}_\lambda^t - \widehat{\theta}_\lambda^t)\right) \widetilde{\nu}_\lambda^{t-1}\left(\overline{\theta}_\lambda^t\right) \\
&\leq \underline{\phi}^{-1}\left(\mathsf{t}(\overline{\theta}_\lambda^t - \widehat{\theta}_\lambda^t)\right) \mathsf{c}
\end{aligned}$$

where we used the fact that $\overline{\theta}_\lambda^t \in \mathsf{D}_\lambda^{t-1}(\mathsf{c})$, an assumption of Proposition 7. With the same reasoning of Eq. (52), we conclude:

$$\mathsf{t}(\widehat{\theta}_\lambda^t - \overline{\theta}_\lambda^t) \leq -\log(1-\mathsf{c}).$$

Bound on $\left\|\widehat{\theta}_\lambda^t - \overline{\theta}_\lambda^t\right\|^2_{\mathbf{H}(\theta^\star)}$. Use a similar reasoning as we used for the variance. Under H_1, we have (Proof of Theorem 2, 1st point)

$$\left\|\widehat{\theta}_\lambda^t - \overline{\theta}_\lambda^t\right\|^2_{\mathbf{H}(\theta^\star)} \leq 2\left\|\widehat{\theta}_\lambda^t - \overline{\theta}_\lambda^t\right\|^2_{\widehat{\mathbf{H}}_\lambda(\theta^\star)}.$$

First write:

$$\left\|\widehat{\theta}_\lambda^t - \overline{\theta}_\lambda^t\right\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star)} \leq e^{\mathsf{t}(\widehat{\theta}_\lambda^t - \theta^\star)/2} e^{\mathsf{t}(\overline{\theta}_\lambda^t - \widehat{\theta}_\lambda^t)/2} \left\|\widehat{\theta}_\lambda^t - \overline{\theta}_\lambda^t\right\|_{\widehat{\mathbf{H}}_\lambda(\overline{\theta}_\lambda^t)},$$

then, for each term, use:

- $\mathsf{t}(\widehat{\theta}_\lambda^t - \theta^\star) \leq 4$ (end of Theorem 4) so that $e^{\mathsf{t}(\widehat{\theta}_\lambda^t - \theta^\star)/2} \leq e^2$;
- $\mathsf{t}(\overline{\theta}_\lambda^t - \widehat{\theta}_\lambda^t) \leq -\log(1-\mathsf{c})$ so that $e^{\mathsf{t}(\overline{\theta}_\lambda^t - \widehat{\theta}_\lambda^t)/2} \leq (1-\mathsf{c})^{-1/2}$;
- and finally:

$$\begin{aligned}
\left\|\widehat{\theta}_\lambda^t - \overline{\theta}_\lambda^t\right\|_{\widehat{\mathbf{H}}_\lambda(\overline{\theta}_\lambda^t)} &\leq \underline{\phi}^{-1}\left(\mathsf{t}(\overline{\theta}_\lambda^t - \widehat{\theta}_\lambda^t)\right) \left\|\nabla \widehat{L}_\lambda^{t-1}(\overline{\theta}_\lambda^t)\right\|_{\mathbf{H}_\lambda^{-1}(\overline{\theta}_\lambda^t)} \\
&\leq \underline{\phi}^{-1}\left(\mathsf{t}(\overline{\theta}_\lambda^t - \widehat{\theta}_\lambda^t)\right) \nu_\lambda^{t-1}\left(\overline{\theta}_\lambda^t\right) \\
&\leq \left[-\frac{1}{\mathsf{c}}\log(1-\mathsf{c})\right]\epsilon \overset{\text{def.}}{=} \kappa_\mathsf{c}\epsilon.
\end{aligned}$$

Putting it all together. Thus, using the upper bound on the excess risk, we have with probability greater than $1-2\delta$

$$L(\overline{\theta}_\lambda^t) - L(\theta^\star) \leq \mathsf{C}_{\text{bias}}\lambda^{2s} + \mathsf{C}_{\text{var}}\frac{\mathsf{df}_\lambda}{n} + 4\Psi(4-\log(1-\mathsf{c}))\frac{e^4}{1-\mathsf{c}}\kappa_\mathsf{c}^2\epsilon, \quad s = (r+1/2)\wedge t.$$

Taking $\mathsf{c} = 1/2$, we have $\Psi(4-\log(1-\mathsf{c})) \leq 5$ and $\kappa_\mathsf{c} \leq 1.4$, which allows bounding the quantity in front of ϵ. □

D Technical lemmas

D.1 Concentration of Hermitian operators

In this section, we import results from [1] and [17]. The former provides a bound on $\left\|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta)\mathbf{H}_\lambda^{1/2}(\theta)\right\|$. The latter provides a bound on $\left\|\widehat{\mathbf{H}}_\lambda^{-1}(\theta)\mathbf{H}_\lambda(\theta)\right\|$, which is more difficult to obtain. They use the fact that $\mathsf{df}_\lambda = \operatorname{Tr}\mathbf{H}_\lambda(\theta)\mathbf{H}(\theta)$ for least square, but we can't use this very convenient relation here. Thus, we only use their result in the case $1/2 < r < 1$, which makes optimal rate still possible.

We will only use

$$\operatorname{Tr}\mathbf{H}_\lambda^{-1}(\theta)\widehat{\mathbf{H}}(\theta) \leq \frac{\mathsf{B}_2(\theta)}{\lambda}.$$

Proposition 9 (Concentration bound). *Let $\delta \in (0,1]$ and $\lambda > 0$. The following holds:*

$$n \geq 24\frac{\mathsf{B}_2(\theta)}{\lambda}\log\frac{8\mathsf{B}_2(\theta)}{\lambda\delta} \quad \Longrightarrow \quad \left\|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta)\mathbf{H}_\lambda^{1/2}(\theta)\right\| \leq \sqrt{2}, \tag{64}$$

$$n \geq 8\frac{\mathsf{B}_2(\theta)^2}{\lambda^2}\log^2\frac{2}{\delta} \quad \Longrightarrow \quad \left\|\widehat{\mathbf{H}}_\lambda^{-1}(\theta)\mathbf{H}_\lambda(\theta)\right\| \leq 2, \tag{65}$$

$$n \geq 2\left(1 \vee \frac{4\mathsf{B}_2(\theta)^2}{\lambda^{2s}}\right)\log\frac{2}{\delta} \quad \Longrightarrow \quad \left\|\mathbf{H}(\theta) - \widehat{\mathbf{H}}(\theta)\right\|_{HS} \leq \lambda^s, \tag{66}$$

where each bound hold with confidence $1-\delta$.

Proof. The first equation is Lemma 6 of [1]. The second equation can be adapted from Proposition 5.4 of [17], except that we use

$$\operatorname{Tr}\mathbf{H}_\lambda(\theta)\mathbf{H}(\theta) \leq \frac{\mathsf{B}_2(\theta)}{\lambda}$$

instead of df_λ. For the last inequality, use Bernstein inequality for random vectors. With probability $1-\delta$:

$$\left\|\mathbf{H}(\theta) - \widehat{\mathbf{H}}(\theta)\right\|_{HS} \leq \frac{2\mathsf{B}_2(\theta)\log 2/\delta}{n} + \mathsf{B}_2(\theta)\sqrt{\frac{2\log 2/\delta}{n}}.$$

Assuming $n \geq 2\log 2/\delta$, this bound becomes

$$\left\|\mathbf{H}(\theta) - \widehat{\mathbf{H}}(\theta)\right\|_{HS} \leq 2\mathsf{B}_2(\theta)\sqrt{\frac{2\log 2/\delta}{n}}.$$

Let $s > 0$. Further requiring $n \geq 8\mathsf{B}_2(\theta)\lambda^{-2s}\log 2/\delta$ gives:

$$\left\|\mathbf{H}(\theta) - \widehat{\mathbf{H}}(\theta)\right\|_{HS} \leq \lambda^s$$

which completes the proof. □

D.2 Inequalities on Hermitian operators

The following results are given in [17]. We redo the proof to track down and upper bound the constants which are discarded in the original paper.

Lemma 4 (Hermitian operator inequalities). *Let A, B be two non-negative self-adjoint operators on $\mathcal{H}$. Assume $\|A\|, \|B\| \leq \kappa$, where $\|\cdot\|$ denotes the operator norm. Then:*

$$\forall r \leq 1, \quad \|A^r - B^r\| \leq \|A - B\|^r \tag{67}$$

$$\forall r > 1, \quad \|A^r - B^r\| \leq w(r)\,\|A - B\| \tag{68}$$

$$\forall r \leq 1, \quad \|A^r B^r\| \leq \|AB\|^r \tag{69}$$

with $r2^{\lfloor r\rfloor+1}\kappa^r$.

Proof. For the first point, refer to [29] Theorem X.1.I, Eq. (X.2). For the third point, refer to Theorem IX.2.1 of the same book. It is also known as Cordes inequality [30]. The proofs involve positive semidefinite matrices but are directly applicable to non-negative self-adjoint Hermitian operators.

For the second point, assume $\|A\|, \|B\| \leq 1$. Consider the function $f(x) = (1-x)^r$, defined for $|x| \leq 1$. Its Taylor expansion reads:

$$f(x) = \sum_{n\geq 0} a_n x^n, \quad a_n = \frac{(-1)^n}{n!} \prod_{k=1}^{n} (r-k+1)$$

We have:

$$\left| \frac{a_{n+1}x^{n+1}}{a_n x^n} \right| = \left| \frac{r-n}{n+1} \cdot x \right| \underset{n\to\infty}{\to} |x|$$

so applying d'Alembert's rule, we have that the radius of the serie is 1. Now, we have that:

$$\begin{aligned} A^r - B^r = f(\mathbf{I}-A) - f(\mathbf{I}-B) &= \sum_{n\geq 0} a_n \left[(\mathbf{I}-A)^n - (\mathbf{I}-B)^n \right] \\ \implies \|A^r - B^r\| &\leq \sum_{n\geq 0} |a_n| \, \|(\mathbf{I}-A)^n - (\mathbf{I}-B)^n\| \end{aligned}$$

Using that $(\mathbf{I}-A)^n - (\mathbf{I}-B)^n = (\mathbf{I}-A)(\mathbf{I}-A)^{n-1} - (\mathbf{I}-B)^{n-1} - (B-A)(\mathbf{I}-B)^{n-1}$, we obtain:

$$\begin{aligned} \|(\mathbf{I}-A)^n - (\mathbf{I}-B)^n\| &\leq \left\|(\mathbf{I}-A)(\mathbf{I}-A)^{n-1} - (\mathbf{I}-B)^{n-1}\right\| + \left\|(B-A)(\mathbf{I}-B)^{n-1}\right\| \\ &\leq \left\|(\mathbf{I}-A)^{n-1} - (\mathbf{I}-B)^{n-1}\right\| + 1 \\ &\leq n \, \|A-B\| \end{aligned}$$

Denoting $g(x) = (1-x)^{r-1} = \sum b_n x^n$, we have $f'(x) = -rg(x)$ which gives $n\,|a_n| = r\,|b_n|$. Then:

$$\begin{aligned} \|A^r - B^r\| &\leq \|A-B\| \sum_{n\geq 0} n\,|a_n| \\ &\leq r\,\|A-B\| \sum_{n\geq 0} |b_n| \end{aligned}$$

We can somewhat painfully upper bound this last term. Notice that for $n > r$, all the b_n have the same sign $s = (-1)^{\lfloor r \rfloor}$. Thus, for $N > r$:

$$\begin{aligned} \sum_{n=0}^{N} |b_n| &= \sum_{n=0}^{\lfloor r \rfloor} |b_n| + s \sum_{n=\lfloor r \rfloor}^{N} b_n \\ &= \sum_{n=0}^{\lfloor r \rfloor} |b_n| + s \lim_{x\to 1} \sum_{n=\lfloor r \rfloor}^{N} b_n x^n \\ &\leq 2 \sum_{n=0}^{\lfloor r \rfloor} |b_n| + \lim_{x\to 1} g(x) \\ &\leq 2 \sum_{n=0}^{\lfloor r \rfloor} \frac{1}{n!} \prod_{k=1}^{n} (r-k+1) \\ &\leq 2 \sum_{n=0}^{\lfloor r \rfloor} \binom{\lfloor r \rfloor}{n} = 2^{\lfloor r \rfloor + 1} \end{aligned}$$

Finally, apply these properties to $A/\kappa, B/\kappa$ to obtain in general:

$$\|A^r - B^r\| \leq r 2^{\lfloor r \rfloor + 1} \kappa^r \, \|A-B\|$$

□

D.3 Basic calculus

This is a few line of computation, but useful in multiple places.

Lemma 5 (Bound on residual of IT's spectral function). *Let $r, t > 0$. Consider the following function defined on $[0, \kappa]$:*

$$h(\sigma) = \left(\frac{\lambda}{\lambda + \sigma}\right)^t \sigma^r.$$

Then:

$$\sup_{0 \leq \sigma \leq \kappa} h(\sigma) \leq \begin{cases} \left(r \cdot \frac{\lambda}{t}\right)^r & \text{if } \ r < t \\ \left(\frac{\lambda}{\kappa + \lambda}\right)^t \kappa^r & \text{otherwise } . \end{cases}$$

Proof. h is differentiable and

$$h'(\sigma) = \frac{\lambda^t \sigma^{r-1}}{(\sigma + \lambda)^{t+1}} \left[\sigma(r - t) + r\lambda\right].$$

If $t \leq r$, the regularization saturates and the maximum is in $\hat{\sigma} = \kappa$, which gives

$$\sup_{\sigma} h(\sigma) \leq \left(\frac{\lambda}{\kappa + \lambda}\right)^t \kappa^r \underset{\lambda \to 0}{\sim} \lambda^t \kappa^{r-t}.$$

Otherwise, if $t > r$, the maximum is in $\hat{\sigma} = \frac{r\lambda}{t-r}$ and it reads

$$\begin{aligned} \sup_{\sigma} h(\sigma) &\leq \left(\frac{t-r}{t}\right)^t \left(\frac{r\lambda}{t-r}\right)^r \\ &= \left(\frac{t-r}{t}\right)^{t-r} r^r \left(\frac{\lambda}{t}\right)^r. \end{aligned} \tag{70}$$

We can rewrite the prefactor in front of $(\lambda/t)^r$. First,

$$\left(\frac{t-r}{t}\right)^{t-r} r^r = \left(\frac{t-r}{t}\right)^t \left(\frac{rt}{t-r}\right)^r.$$

Then, use

$$\left(\frac{t-r}{t}\right)^t \leq e^{-r} \quad \text{when} \quad r < t. \tag{71}$$

Also,

$$\left(\frac{rt}{e(t-r)}\right)^r = \left(e\left(\frac{1}{r} - \frac{1}{t}\right)\right)^{-r} \leq (e/r)^{-r} \leq r^r. \tag{72}$$

Use Eq. (71) and Eq. (72) on the upper bound of Eq. (70), and the result is obtained. □

E Experiments

E.1 Technical details

Splines. The spline kernel of order q is defined on $[0, 1]^2$ as

$$\Lambda_q(x, z) = \sum_{k \in \mathbb{Z}} \frac{e^{2i\pi k(x-z)}}{|k|^q}.$$

A closed form expression is available when q is an even integer:

$$\Lambda_q(x, z) = 1 + \frac{(-1)^{q/2-1}}{q!} B_q(|x - z|).$$

B_q are Bernoulli polynomial of order q. They can be implemented easily. We also have the relation

$$\langle \Lambda_q(x, \cdot), \Lambda_{q'}(x', \cdot) \rangle_{L_2(\mathcal{X}, \rho_{\mathcal{X}})} = \Lambda_{q+q'}(x, x')$$

Our choice of r, α reflects the constraints on α and $(r + 1/2)\alpha + 1/2$ to be even integers.

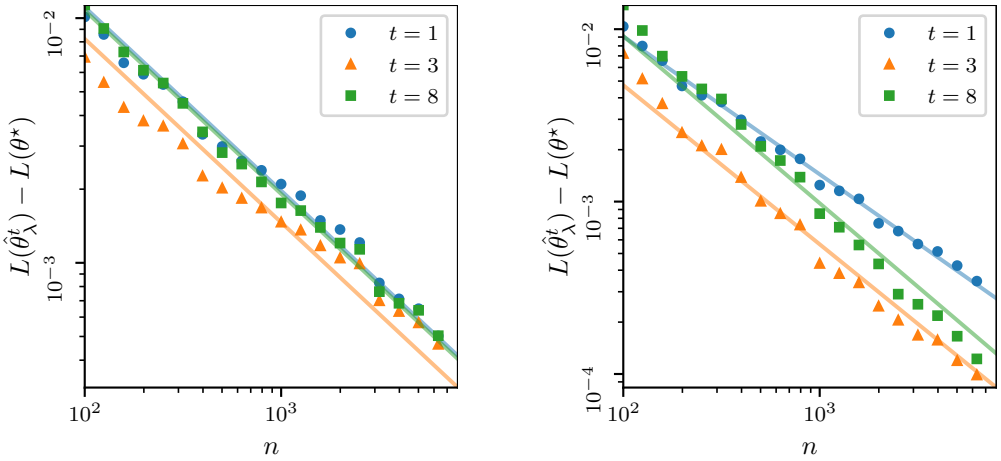

Figure 2: Excess risk with **least square** for various Iterated Tikhonov estimator, function of n. **Colors**: $t = 1$ (Tikhonov) estimator is shown in orange; $t = 2, 3$ in green, red. **Left**: from a difficult problem, $r = 1/4, \alpha = 2$. **Right**: easy problem, $r = 41/4, \alpha = 2$. Plain lines are predicted by theory, with slope $-\alpha(1+2s)/1+\alpha(1+2s)$, $s = \min\{r, t - 1/2\}$ (see main text). All plots are averaged over 100 different initialization.

Regularization. For both least square and logistic regression, the regularization λ is chosen among 50 log spaced values between 10^{-4} and 1.

Resources. Computation was carried by a `Intel(R) Xeon(R) CPU E5-1620 v2 @ 3.70GHz`, with 32GB of RAM.

E.2 Simulations with least square

Estimating $\widehat{\theta}^t_\lambda$. We leverage the very convenient filter interpretation with least square. We diagonalize the kernel matrix $K = UDU^\top$ once, then evaluate the estimator with

$$\widehat{\theta}^t_\lambda = \sum_{i=1}^n \alpha_i \phi(x_i),$$
$$\alpha = \frac{1}{n} U g^t_\lambda(D/n) D^\top y,$$

where g^t_λ is IT's filter, defined in (8).

Simulations. The simulations are reported in Figs. 2 and 3. The same broad conclusion as for the classification task with the logistic loss apply. Surprisingly, IT(8) seems to suffer from higher constant than its counterpart with low t.

E.3 Synthetic binary task

Derivation of the noise. We have $\theta^\star(x) = \Lambda_{(r+1/2)\alpha+\epsilon}(x, 0)$ a function of smoothness $r + 1/2$ in $L_2(\mathcal{X}, \rho_{\mathbf{x}})$. We want to use logistic regression. Thus, we need to choose the noise $\rho(y \mid x)$ so that

$$\theta^\star(x) = \arg\min_z \int_{\mathcal{Y}} \ell(y, z) \mathrm{d}\rho_{y|x}(y).$$

To keep things simple, we restrict the output space to $\mathcal{Y} = \{-1, 1\}$. Denote $a(x) = \mathbb{P}(y = 1 \mid x)$. We will have $\mathbb{P}(y = -1 \mid x) = 1 - a(x)$. Now we need to choose a s.t

$$a \in [0, 1] \quad \text{and} \quad \theta^\star(x) = \arg\min_z h(z) \stackrel{\text{def.}}{=} \log(1 + e^z)(1 - a) + \log(1 + e^{-z})a.$$

Having $a > 0, 1 - a > 0$ implies that h has a unique minimizer z^*. Then

$$h'(z) = \frac{1}{1 + e^z}\left((1 - a)e^z - a\right) \implies a = \frac{e^{z^*}}{1 + e^{z^*}}.$$

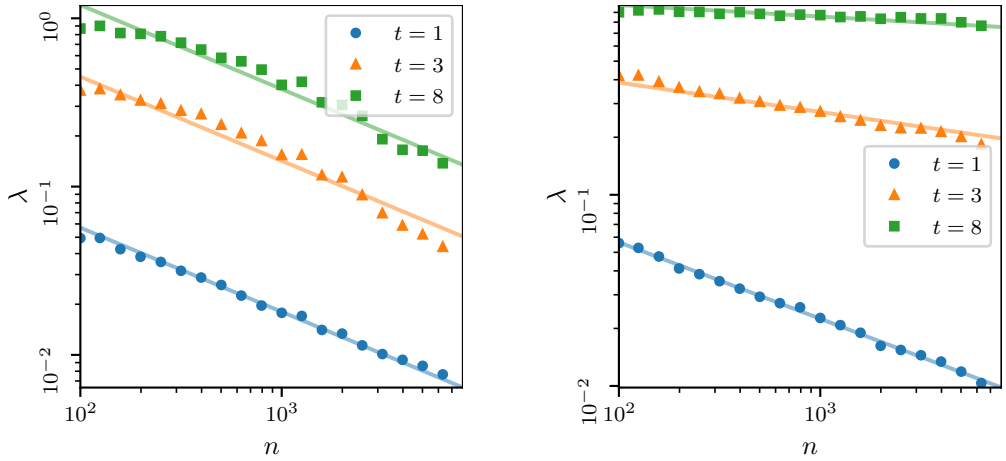

Figure 3: Chosen regularization λ with **least square** for various Iterated Tikhonov estimator, function of n. **Colors**: $t = 1$ (Tikhonov) estimator is shown in orange; $t = 3, 8$ in green, red. **Left**: from a difficult problem, $r = 1/4, \alpha = 2$. **Right**: easy problem, $r = 41/4, \alpha = 2$. Plain lines are predicted by theory, with slope $-\alpha/1+\alpha(1+2s)$, $s = \min\{r, t - 1/2\}$ (see main text). All plots are averaged over 100 different initialization.

Having required that $\theta^\star(x) = \arg\min_z h(z) \overset{\text{def.}}{=} z^*$, we can use the following output distribution:

$$\mathcal{Y} = \{-1, 1\}$$
$$\mathbb{P}(y = 1 \mid x) = \frac{1}{1 + e^{-\theta^\star(x)}}$$
$$\mathbb{P}(y = -1 \mid x) = \frac{1}{1 + e^{+\theta^\star(x)}}$$

which, in turn, ensures that $a(x) \in [0, 1]$.

Newton or first-order methods. In practice, the proximal operator is evaluated with a Newton method, or we use the toolbox Cyanure for big n [31]. Both are used with tolerance 10^{-10}, that is machine precision for single precision. Generally speaking, first-order methods are considered more performant than Newton methods. However, both practical and theoretical considerations motivate the use of second-order scheme in our statement of Proposition 1. Firstly, preconditionated iterative solver such as the one used in [14] provide very efficient results for ill-conditioned problems. Secondly, the analysis of GSC loss functions is well-suited to second-order scheme, as the Newton decrement is a natural quantity to keep track of the optimization error. Measuring the error differently would require additional assumption on the loss function.

Estimating the excess risk. The excess risk is estimated with Monte Carlo sampling, with 10^4 points:

$$ER(\theta) - ER(\theta^\star) \approx \frac{1}{n_{MC}} \sum_{i=1}^{n_{MC}} \frac{1}{1 + e^{-\theta^\star(x_i)}} \log\left(\frac{1 + e^{-\theta(x_i)}}{1 + e^{-\theta^\star(x_i)}}\right) + \frac{1}{1 + e^{\theta^\star(x_i)}} \log\left(\frac{1 + e^{\theta(x_i)}}{1 + e^{\theta^\star(x_i)}}\right)$$

Additional results. We report here the regularization λ chosen function of n and t for various IT regularized estimators. We confirm that the penalty used for IT is larger than of Tikhonov, to compensate for the fitting induced by the additional proximal steps. We also compare the excess risk achieved by IT with the excess risk of Tikhonov, and observe consistent improvement for easy task with a sufficiently high number of samples.

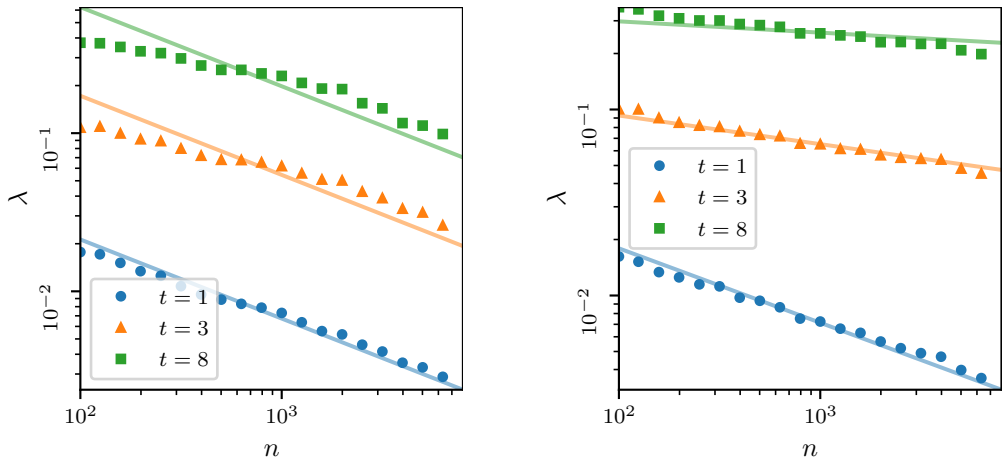

Figure 4: Chosen regularization λ for various Iterated Tikhonov estimator, function of n. **Colors**: $t = 1$ (Tikhonov) estimator is shown in orange; $t = 3, 8$ in green, red. **Left**: from a difficult problem, $r = 1/4, \alpha = 2$. **Right**: easy problem, $r = 41/4, \alpha = 2$. Plain lines are predicted by theory, with slope $-\alpha/1+\alpha(1+2s)$, $s = \min\{r, t - 1/2\}$ (see main text). All plots are averaged over 100 different initialization.

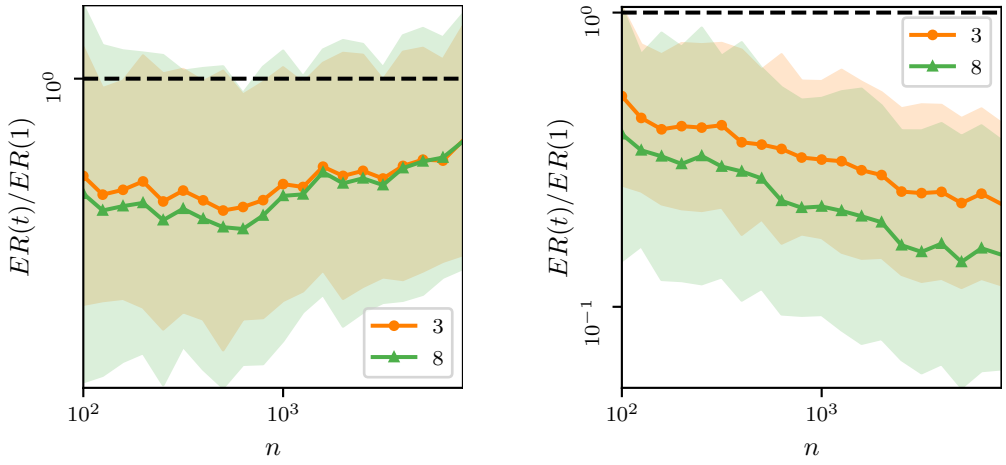

Figure 5: Ratio of IT's excess risk over Tikhonov's excess risk, function of n. **Left**: from a difficult problem, $r = 1/4, \alpha = 2$. **Right**: easy problem, $r = 41/4, \alpha = 2$. Whereas we expect the ratio to be consistently lower than 1, IT performs worse than Tikhonov in isolated cases, probably due to the optimization process and the chosen regularization path. Yet, it provides lower excess risk than Tikhonov overall, with up to an order of magnitude of improvement with as few as 1000 samples. All plots are averaged over 100 different initialization.