Self-concordant analysis for logistic regression

Francis Bach

INRIA - WILLOW Project-Team
Laboratoire d'Informatique de l'Ecole Normale Supérieure
(CNRS/ENS/INRIA UMR 8548)
23, avenue d'Italie, 75214 Paris, France
francis.bach@inria.fr

October 26, 2009

Abstract

Most of the non-asymptotic theoretical work in regression is carried out for the square loss, where estimators can be obtained through closed-form expressions. In this paper, we use and extend tools from the convex optimization literature, namely self-concordant functions, to provide simple extensions of theoretical results for the square loss to the logistic loss. We apply the extension techniques to logistic regression with regularization by the ℓ_2-norm and regularization by the ℓ_1-norm, showing that new results for binary classification through logistic regression can be easily derived from corresponding results for least-squares regression.

1 Introduction

The theoretical analysis of statistical methods is usually greatly simplified when the estimators have closed-form expressions. For methods based on the minimization of a certain functional, such as M-estimation methods [1], this is true when the function to minimize is quadratic, i.e., in the context of regression, for the square loss.

When such loss is used, asymptotic and non-asymptotic results may be derived with classical tools from probability theory (see, e.g., [2]). When the function which is minimized in M-estimation is not amenable to closed-form solutions, local approximations are then needed for obtaining and analyzing a solution of the optimization problem. In the asymptotic regime, this has led to interesting developments and extensions of results from the quadratic case, e.g., consistency or asymptotic normality (see, e.g., [1]). However, the situation is different when one wishes to derive non-asymptotic results, i.e., results where all constants of the problem are explicit. Indeed, in order to prove results as sharp as for the square loss, much notation and many assumptions have to be introduced regarding second and third derivatives; this makes the derived results much more complicated than the ones for closed-form estimators [3, 4, 5].

A similar situation occurs in convex optimization, for the study of Newton's method for obtaining solutions of unconstrained optimization problems. It is known to be locally quadratically convergent for convex problems. However, its classical analysis requires cumbersome notations and assumptions regarding second and third-order derivatives (see, e.g., [6, 7]). This situation was greatly enhanced with the introduction of the notion of *self-concordant functions*, i.e., functions whose third derivatives are controlled by their second derivatives. With this tool, the analysis is much more transparent [7, 8]. While Newton's method is a commonly used algorithm for logistic regression (see, e.g., [9, 10]), leading to iterative least-squares algorithms, we don't focus in the paper on the resolution of the optimization problems, but on the statistical analysis of the associated global minimizers.

In this paper, we aim to borrow tools from convex optimization and self-concordance to analyze the statistical properties of logistic regression. Since the logistic loss is not itself a self-concordant function, we introduce in Section 2 a new type of functions with a different control of the third derivatives. For these functions, we prove two types of results: first, we provide lower and upper Taylor expansions, i.e., Taylor expansions which are globally upper-bounding or lower-bounding a given function. Second, we prove results on the behavior of Newton's method which are similar to the ones for self-concordant functions. We then apply them in Sections 3, 4 and 5 to the one-step Newton iterate from the population solution of the corresponding problem (i.e., ℓ_2 or ℓ_1-regularized logistic regression). This essentially shows that the analysis of logistic regression can be done *non-asymptotically* using the local quadratic approximation of the logistic loss, *without* complex additional assumptions. Since this approximation corresponds to a weighted least-squares problem, results from least-squares regression can thus be naturally extended.

In order to consider such extensions and make sure that the new results closely match the corresponding ones for least-squares regression, we derive in Appendix G new Bernstein-like concentration inequalities for quadratic forms of bounded random variables, obtained from general results on U-statistics [11].

We first apply in Section 4 the extension technique to regularization by the ℓ_2-norm, where we consider two settings, a situation with no assumptions regarding the conditional distribution of the observations, and another one where the model is assumed well-specified and we derive asymptotic expansions of the generalization performance with explicit bounds on remainder terms. In Section 5, we consider regularization by the ℓ_1-norm and extend two known recent results for the square loss, one on model consistency [12, 13, 14, 15] and one on prediction efficiency [16]. The main contribution of this paper is to make these extensions as simple as possible, by allowing the use of non-asymptotic second-order Taylor expansions.

Notation. For $x \in \mathbb{R}^p$ and $q \geqslant 1$, we denote by $\|x\|_q$ the ℓ_q-norm of x, defined as $\|x\|_q^q = \sum_{i=1}^p |x_i|^q$. We also denote by $\|x\|_\infty = \max_{i \in \{1,\dots,p\}} |x_i|$ its ℓ_∞-norm. We denote by $\lambda_{\max}(Q)$ and $\lambda_{\min}(Q)$ the largest and smallest eigenvalue of a symmetric matrix Q. We use the notation $Q_1 \preccurlyeq Q_2$ (resp. $Q_1 \succcurlyeq Q_2$) for the positive semi-definiteness of the matrix

$Q_2 - Q_1$ (resp. $Q_1 - Q_2$).

For $a \in \mathbb{R}$, $\operatorname{sign}(a)$ denotes the sign of a, defined as $\operatorname{sign}(a) = 1$ if $a > 0$, -1 if $a < 0$, and 0 if $a = 0$. For a vector $v \in \mathbb{R}^p$, $\operatorname{sign}(v) \in \{-1, 0, 1\}^p$ denotes the vector of signs of elements of v.

Moreover, given a vector $v \in \mathbb{R}^p$ and a subset I of $\{1, \dots, p\}$, $|I|$ denotes the cardinal of the set I, v_I denotes the vector in $\mathbb{R}^{|I|}$ of elements of v indexed by I. Similarly, for a matrix $A \in \mathbb{R}^{p \times p}$, A_{IJ} denotes the submatrix of A composed of elements of A whose rows are in I and columns are in J. Finally, we let denote $\mathbb{P}$ and $\mathbb{E}$ general probability measures and expectations.

2 Taylor expansions and Newton's method

In this section, we consider a generic function $F : \mathbb{R}^p \to \mathbb{R}$, which is convex and three times differentiable. We denote by $F'(w) \in \mathbb{R}^p$ its gradient at $w \in \mathbb{R}^p$, by $F''(w) \in \mathbb{R}^{p \times p}$ its Hessian at $w \in \mathbb{R}^p$. We denote by $\lambda(w) \geqslant 0$ the smallest eigenvalue of the Hessian $F''(w)$ at $w \in \mathbb{R}^p$.

If $\lambda(w) > 0$, i.e., the Hessian is invertible at w, we can define the *Newton step* as $\Delta^N(w) = -F''(w)^{-1} F'(w)$, and the *Newton decrement* $\nu(F, w)$ at w, defined through:

$$\nu(F, w)^2 = F'(w)^\top F''(w)^{-1} F'(w) = \Delta^N(w)^\top F''(w) \Delta^N(w).$$

The *one-step Newton iterate* $w + \Delta^N(w)$ is the minimizer of the second-order Taylor expansion of F at w, i.e., of the function $v \mapsto F(w) + F'(w)(v - w) + \frac{1}{2}(v - w)^\top F''(w)(v - w)$. Newton's method consists in successively applying the same iteration until convergence. For more background and details about Newton's method, see, e.g., [7, 6, 17].

2.1 Self-concordant functions

We now review some important properties of self-concordant functions [7, 8], i.e., three times differentiable convex functions such that for all $u, v \in \mathbb{R}^p$, the function $g : t \mapsto F(u + tv)$ satisfies for all $t \in \mathbb{R}$, $|g'''(t)| \leqslant 2 g''(t)^{3/2}$.

The local behavior of self-concordant functions is well-studied and lower and upper Taylor expansions can be derived (similar to the ones we derive in Proposition 1). Moreover, bounds are available for the behavior of Newton's method; given a self-concordant function F, if $w \in \mathbb{R}^p$ is such that $\nu(F, w) \leqslant 1/4$, then F attains its unique global minimum at some $w^* \in \mathbb{R}^p$, and we have the following bound on the error $w - w^*$ (see, e.g., [8]):

$$(w - w^*)^\top F''(w)(w - w^*) \leqslant 4 \nu(F, w)^2. \tag{1}$$

Moreover, the newton decrement at the one-step Newton iterate from $w \in \mathbb{R}^p$ can be upper-bounded as follows:

$$\nu(F, w + \Delta^N(w)) \leqslant \nu(F, w)^2, \tag{2}$$

which allows to prove an upper-bound of the error of the one-step iterate, by application of Eq. (1) to $w + \Delta^N(w)$. Note that these bounds are not the sharpest, but are sufficient in our context. These are commonly used to show the global convergence of the damped Newton's method [8] or of Newton's method with backtracking line search [7], as well as a precise upper bound on the number of iterations to reach a given precision.

Note that in the context of machine learning and statistics, self-concordant functions have been used for bandit optimization and online learning [18], but for barrier functions related to constrained optimization problems, and not directly for M-estimation.

2.2 Modifications of self-concordant functions

The logistic function $u \mapsto \log(1+e^{-u})$ is not self-concordant as the third derivative is bounded by a constant times the second derivative (without the power $3/2$). However, similar bounds can be derived with a different control of the third derivatives. Proposition 1 provides lower and upper Taylor expansions while Proposition 2 considers the behavior of Newton's method. Proofs may be found in Appendix A and follow closely the ones for regular self-concordant functions found in [8].

Proposition 1 (Taylor expansions) *Let $F : \mathbb{R}^p \mapsto \mathbb{R}$ be a convex three times differentiable function such that for all $w, v \in \mathbb{R}^p$, the function $g(t) = F(w+tv)$ satisfies for all $t \in \mathbb{R}$, $|g'''(t)| \leqslant R\|v\|_2 \times g''(t)$, for some $R \geqslant 0$. We then have for all $w, v, z \in \mathbb{R}^p$:*

$$F(w+v) \;\geqslant\; F(w) + v^\top F'(w) + \frac{v^\top F''(w)v}{R^2\|v\|_2^2}\big(e^{-R\|v\|_2} + R\|v\|_2 - 1\big), \tag{3}$$

$$F(w+v) \;\leqslant\; F(w) + v^\top F'(w) + \frac{v^\top F''(w)v}{R^2\|v\|_2^2}\big(e^{R\|v\|_2} - R\|v\|_2 - 1\big), \tag{4}$$

$$\frac{z^\top[F'(w+v) - F'(w) - F''(w)v]}{[z^\top F''(w)z]^{1/2}} \leqslant [v^\top F''(w)v]^{1/2}\frac{e^{R\|v\|_2} - 1 - R\|v\|_2}{R\|v\|_2}, \tag{5}$$

$$e^{-R\|v\|_2}F''(w) \preccurlyeq F''(w+v) \preccurlyeq e^{R\|v\|_2}F''(w). \tag{6}$$

Inequalities in Eq. (3) and Eq. (4) provide upper and lower second-order Taylor expansions of F, while Eq. (5) provides a first-order Taylor expansion of F' and Eq. (6) can be considered as an upper and lower zero-order Taylor expansion of F''. Note the difference here between Eqs. (3-4) and regular third-order Taylor expansions of F: the remainder term in the Taylor expansion, i.e., $F(w+v) - F(w) - v^\top F'(w) - \frac{1}{2}v^\top F''(w)v$ is upper-bounded by $\frac{v^\top F''(w)v}{R^2\|v\|_2^2}(e^{R\|v\|_2} - \frac{1}{2}R^2\|v\|_2^2 - R\|v\|_2 - 1)$; for $\|v\|_2$ small, we obtain a term proportional to $\|v\|_2^3$ (like a regular local Taylor expansion), but the bound remains valid for all v and does not grow as fast as a third-order polynomial. Moreover, a regular Taylor expansion with a uniformly bounded third-order derivative would lead to a bound proportional to $\|v\|_2^3$, which does not take into account the local curvature of F at w. Taking into account this local curvature is key to obtaining sharp and simple bounds on the behavior of Newton's method (see proof in Appendix A):

Proposition 2 (Behavior of Newton's method) *Let $F : \mathbb{R}^p \mapsto \mathbb{R}$ be a convex three times differentiable function such that for all $w, v \in \mathbb{R}^p$, the function $g(t) = F(w + tv)$ satisfies for all $t \in \mathbb{R}$, $|g'''(t)| \leqslant R\|v\|_2 \times g''(t)$, for some $R \geqslant 0$. Let $\lambda(w) > 0$ be the lowest eigenvalue of $F''(w)$ for some $w \in \mathbb{R}^p$. If $\nu(F, w) \leqslant \frac{\lambda(w)^{1/2}}{2R}$, then F has a unique global minimizer $w^* \in \mathbb{R}^p$ and we have:*

$$\begin{aligned} \big(w - w^*\big)^\top F''(w)\big(w - w^*\big) &\leqslant 16\nu(F, w)^2, &(7)\\ \frac{R\nu(F, w + \Delta^N(w))}{\lambda(w + \Delta^N(w))^{1/2}} &\leqslant \left(\frac{R\nu(F, w)}{\lambda(w)^{1/2}}\right)^2, &(8)\\ \big(w + \Delta^N(w) - w^*\big)^\top F''(w)\big(w + \Delta^N(w) - w^*\big) &\leqslant \frac{16R^2}{\lambda(w)}\nu(F, w)^4. &(9)\end{aligned}$$

Eq. (7) extends Eq. (1) while Eq. (8) extends Eq. (2). Note that the notion and the results are not invariant by affine transform (contrary to self-concordant functions) and that we still need a (non-uniformly) lower-bounded Hessian. The last two propositions constitute the main technical contribution of this paper. We now apply these to logistic regression and its regularized versions.

3 Application to logistic regression

We consider n pairs of observations (x_i, y_i) in $\mathbb{R}^p \times \{-1, 1\}$ and the following objective function for logistic regression:

$$\hat{J}_0(w) = \frac{1}{n}\sum_{i=1}^n \log\Big(1 + \exp(-y_i w^\top x_i)\Big) = \frac{1}{n}\sum_{i=1}^n \Big\{\ell(w^\top x_i) - \frac{y_i}{2} w^\top x_i\Big\}, \qquad (10)$$

where $\ell : u \mapsto \log(e^{-u/2} + e^{u/2})$ is an even convex function. A short calculation leads to $\ell'(u) = -1/2 + \sigma(u)$, $\ell''(u) = \sigma(u)[1 - \sigma(u)]$, $\ell'''(u) = \sigma(u)[1 - \sigma(u)][1 - 2\sigma(u)]$, where $\sigma(u) = (1 + e^{-u})^{-1}$ is the sigmoid function. Note that we have for all $u \in \mathbb{R}$, $|\ell'''(u)| \leqslant \ell''(u)$. The cost function $\hat{J}_0$ defined in Eq. (10) is proportional to the negative conditional log-likelihood of the data under the conditional model $\mathbb{P}(y_i = \varepsilon_i|x_i) = \sigma(\varepsilon_i w^\top x_i)$.

If $R = \max_{i\in\{1,\dots,n\}} \|x_i\|_2$ denotes the maximum ℓ_2-norm of all input data points, then the cost function $\hat{J}_0$ defined in Eq. (10) satisfies the assumptions of Proposition 2. Indeed, we have, with the notations of Proposition 2,

$$\begin{aligned} |g'''(t)| &= \left|\frac{1}{n}\sum_{i=1}^n \ell'''[(w + tv)^\top x_i](x_i^\top v)^3\right| \\ &\leqslant \frac{1}{n}\sum_{i=1}^n \ell''[(w + tv)^\top x_i](x_i^\top v)^2\|v\|_2\|x_i\|_2 \leqslant R\|v\|_2 \times g''(t). \end{aligned}$$

Throughout this paper, we will consider a certain vector $w \in \mathbb{R}^p$ (usually defined through the population functionals) and consider the one-step Newton iterate from this w. Results from Section 2.2 will allow to show that this approximates the global minimum of $\hat{J}_0$ or a regularized version thereof.

Throughout this paper, we consider a *fixed design* setting (i.e., $x_1, \dots, x_n$ are consider deterministic) and we make the following assumptions:

(A1) *Independent outputs*: The outputs $y_i \in \{-1, 1\}$, $i = 1, \dots, n$ are independent (but not identically distributed).

(A2) *Bounded inputs*: $\max_{i \in \{1,\dots,n\}} \|x_i\|_2 \leqslant R$.

We define the model as *well-specified* if there exists $w_0 \in \mathbb{R}^p$ such that for all $i = 1, \dots, n$, $\mathbb{P}(y_i = \varepsilon_i) = \sigma(\varepsilon_i w_0^\top x_i)$, which is equivalent to $\mathbb{E}(y_i/2) = \ell'(w_0^\top x_i)$, and implies $\mathrm{var}(y_i/2) = \ell''(w_0^\top x_i)$. However, we do not always make such assumptions in the paper.

We use the matrix notation $X = [x_1, \dots, x_n]^\top \in \mathbb{R}^{n \times p}$ for the design matrix and $\varepsilon_i = y_i/2 - \mathbb{E}(y_i/2)$, for $i = 1, \dots, n$, which formally corresponds to the additive noise in least-squares regression. We also use the notation $Q = \frac{1}{n} X^\top \mathrm{Diag}(\mathrm{var}(y_i/2)) X \in \mathbb{R}^{p \times p}$ and $q = \frac{1}{n} X^\top \varepsilon \in \mathbb{R}^p$. By assumption, we have $\mathbb{E}(qq^\top) = \frac{1}{n} Q$.

We denote by J_0 the expectation of $\hat{J}_0$, i.e.:

$$J_0(w) = \mathbb{E}\big[\hat{J}_0(w)\big] = \frac{1}{n} \sum_{i=1}^n \Big\{ \ell(w^\top x_i) - \mathbb{E}(y_i/2) w^\top x_i \Big\}.$$

Note that with our notation, $\hat{J}_0(w) = J_0(w) - q^\top w$. In this paper we consider $J_0(\hat{w})$ as the generalization performance of a certain estimator $\hat{w}$. This corresponds to the average Kullback-Leibler divergence to the best model when the model is well-specified, and is common for the study of logistic regression and more generally generalized linear models [19, 20]. Measuring the classification performance through the 0–1 loss [21] is out of the scope of this paper.

The function J_0 is bounded from below, therefore it has a bounded infimum $\inf_{w \in \mathbb{R}^p} J_0(w) \geqslant 0$. This infimum might or might not be attained at a finite $w_0 \in \mathbb{R}^p$; when the model is well-specified, it is always attained (but this is not a necessary condition), and, unless the design matrix X has rank p, is not unique.

The difference between the analysis through self-concordance and the classical asymptotic analysis is best seen when the model is well-specified, and exactly mimics the difference between self-concordant analysis of Newton's method and its classical analysis. The usual analysis of logistic regression requires that the logistic function $u \mapsto \log(1 + e^{-u})$ is strongly convex (i.e., with a strictly positive lower-bound on the second derivative), which is true only on a compact subset of $\mathbb{R}$. Thus, non-asymptotic results such as the ones from [5, 3] requires an upper bound M on $|w_0^\top x_i|$, where w_0 is the generating loading vector; then, the second derivative of the logistic loss is lower bounded by $(1 + e^M)^{-1}$, and this lower bound may be very small when M gets large. Our analysis does not require such a bound because of the fine control of the third derivative.

4 Regularization by the ℓ_2-norm

We denote by $\hat{J}_\lambda(w) = \hat{J}_0(w) + \frac{\lambda}{2}\|w\|_2^2$ the empirical ℓ_2-regularized functional. For $\lambda > 0$, the function $\hat{J}_\lambda$ is strongly convex and we denote by $\hat{w}_\lambda$ the unique global minimizer of $\hat{J}_\lambda$. In this section, our goal is to find upper and lower bounds on the generalization performance $J_0(\hat{w}_\lambda)$, under minimal assumptions (Section 4.2) or when the model is well-specified (Section 4.3).

4.1 Reproducing kernel Hilbert spaces and splines

In this paper we focus explicitly on *linear* logistic regression, i.e., on a generalized linear model that allows linear dependency between x_i and the distribution of y_i. Although apparently limiting, in the context of regularization by the ℓ_2-norm, this setting contains *non-parametric* and *non-linear* methods based on splines or reproducing kernel Hilbert spaces (RKHS) [22]. Indeed, because of the representer theorem [23], minimizing the cost function

$$\frac{1}{n}\sum_{i=1}^{n}\Big\{\ell[f(x_i)] - \frac{y_i}{2}f(x_i)\Big\} + \frac{\lambda}{2}\|f\|_{\mathcal{F}}^2,$$

with respect to the function f in the RKHS $\mathcal{F}$ (with norm $\|\cdot\|_{\mathcal{F}}$ and kernel k), is equivalent to minimizing the cost function

$$\frac{1}{n}\sum_{i=1}^{n}\Big\{\ell[(T\beta)_i] - \frac{y_i}{2}(T\beta)_i\Big\} + \frac{\lambda}{2}\|\beta\|_2^2, \tag{11}$$

with respect to $\beta \in \mathbb{R}^p$, where $T \in \mathbb{R}^{n\times p}$ is a square root of the kernel matrix $K \in \mathbb{R}^{n\times n}$ defined as $K_{ij} = k(x_i, x_j)$, i.e., such that $K = TT^\top$. The unique solution of the original problem f is then obtained as $f(x) = \sum_{i=1}^n \alpha_i k(x, x_i)$, where α is any vector satisfying $TT^\top\alpha = T\beta$ (which can be obtained by matrix pseudo-inversion [24]). Similar developments can be carried out for smoothing splines (see, e.g., [22, 25]). By identifying the matrix T with the data matrix X, the optimization problem in Eq. (11) is identical to minimizing $\hat{J}_0(w) + \frac{\lambda}{2}\|w\|_2^2$, and thus our results apply to estimation in RKHSs.

4.2 Minimal assumptions (misspecified model)

In this section, we do not assume that the model is well-specified. We obtain the following theorem (see proof in Appendix B), which only assumes boundedness of the covariates and independence of the outputs:

Theorem 1 (Misspecified model) *Assume **(A1)**, **(A2)** and* $\lambda = 19R^2\sqrt{\frac{\log(8/\delta)}{n}}$, *with* $\delta \in (0,1)$. *Then, with probability at least* $1-\delta$, *for all* $w_0 \in \mathbb{R}^p$,

$$J_0(\hat{w}_\lambda) \leqslant J_0(w_0) + \big(10 + 100R^2\|w_0\|_2^2\big)\sqrt{\frac{\log(8/\delta)}{n}}. \tag{12}$$

In particular, if the global minimum of J_0 is attained at w_0 (which is not an assumption of Theorem 1), we obtain an oracle inequality as $J_0(w_0) = \inf_{w \in \mathbb{R}^p} J_0(w)$. The lack of additional assumptions unsurprisingly gives rise to a slow rate of $n^{-1/2}$.

This is to be compared with [26], which uses different proof techniques but obtains similar results for all convex Lipschitz-continuous losses (and not only for the logistic loss). However, the techniques presented in this paper allow the derivation of much more precise statements in terms of bias and variance (and with better rates), that involves some knowledge of the problem. We do not pursue detailed results here, but focus in the next section on well-specified models, where results have a simpler form.

This highlights two opposite strategies for the theoretical analysis of regularized problems: the first one, followed by [26, 27], is mostly loss-independent and relies on advanced tools from empirical process theory, namely uniform concentration inequalities. Results are widely applicable and make very few assumptions. However, they tend to give performance guarantees which are far below the observed performances of such methods in applications. The second strategy, which we follow in this paper, is to restrict the loss class (to linear or logistic) and derive the limiting convergence rate, which does depend on unknown constants (typically the best linear classifier itself). Once the limit is obtained, we believe it gives a better interpretation of the performance of these methods, and if one really wishes to make no assumption, taking upper bounds on these quantities, we may get back results obtained with the generic strategy, which is exactly what Theorem 1 is achieving.

Thus, a detailed analysis of the convergence rate, as done in Theorem 2 in the next section, serves two purposes: first, it gives a sharp result that depends on unknown constants; second the constants can be maximized out and more general results may be obtained, with fewer assumptions but worse convergence rates.

4.3 Well-specified models

We now assume that the model is well-specified, i.e., that the probability that $y_i = 1$ is a sigmoid function of a linear function of x_i, which is equivalent to:

(A3) *Well-specified model*: There exists $w_0 \in \mathbb{R}^p$ such that $\mathbb{E}(y_i/2) = \ell'(w_0^\top x_i)$.

Theorem 2 will give upper and lower bounds on the expected risk of the ℓ_2-regularized estimator $\hat{w}_\lambda$, i.e., $J_0(\hat{w}_\lambda)$. We use the following definitions for the two degrees of freedom and biases, which are usual in the context of ridge regression and spline smoothing (see, e.g., [22, 25, 28]):

$$\begin{array}{lrcl} \text{degrees of freedom (1) :} & d_1 & = & \operatorname{tr} Q(Q+\lambda I)^{-1}, \\ \text{degrees of freedom (2) :} & d_2 & = & \operatorname{tr} Q^2(Q+\lambda I)^{-2}, \\ \text{bias (1) :} & b_1 & = & \lambda^2 w_0^\top (Q+\lambda I)^{-1} w_0, \\ \text{bias (2) :} & b_2 & = & \lambda^2 w_0^\top Q(Q+\lambda I)^{-2} w_0. \end{array}$$

Note that we always have the inequalities $d_2 \leqslant d_1 \leqslant \min\{R^2/\lambda, n\}$ and $b_2 \leqslant b_1 \leqslant \min\{\lambda \|w_0\|_2^2, \lambda^2 w_0^\top Q^{-1} w_0\}$, and that these quantities depend on λ. In the context of RKHSs outlined in Section 4.1, we have $d_1 = \operatorname{tr} K(K + n\lambda \operatorname{Diag}(\sigma_i^2))^{-1}$, a quantity which is also usually referred to as the *degrees of freedom* [29]. In the context of the analysis of ℓ_2-regularized methods, the two degrees of freedom are necessary, as outlined in Theorems 2 and 3, and in [28].

Moreover, we denote by $\kappa > 0$ the following quantity

$$\kappa = \frac{R}{\lambda^{1/2}} \left(\frac{d_1}{n} + b_1 \right) \left(\frac{d_2}{n} + b_2 \right)^{-1/2}. \tag{13}$$

Such quantity is an extension of the one used by [30] in the context of kernel Fisher discriminant analysis used as a test for homogeneity. In order to obtain asymptotic equivalents, we require κ to be small, which, as shown later in this section, occurs in many interesting cases when n is large enough.

In this section, we will apply results from Section 2 to the functions $\hat{J}_\lambda$ and J_0. Essentially, we will consider local quadratic approximations of these functions around the generating loading vector w_0, leading to replacing the true estimator $\hat{w}_\lambda$ by the one-step Newton iterate from w_0. This is only possible if the Newton decrement $\nu(\hat{J}_\lambda, w_0)$ is small enough, which leads to additional constraints (in particular the upper-bound on κ).

Theorem 2 (Asymptotic generalization performance) *Assume **(A1)**, **(A2)** and **(A3)**. Assume moreover $\kappa \leqslant 1/16$, where κ is defined in Eq. (13). If $v \in [0, 1/4]$ satisfies $v^3(d_2 + nb_2)^{1/2} \leqslant 12$, then, with probability at least $1 - \exp(-v^2(d_2 + nb_2))$:*

$$\left| J_0(\hat{w}_\lambda) - J_0(w_0) - \frac{1}{2} \left(b_2 + \frac{d_2}{n} \right) \right| \leqslant \left(b_2 + \frac{d_2}{n} \right) (69v + 2560\kappa). \tag{14}$$

Relationship to previous work. When the dimension p of w_0 is bounded, then under the regular asymptotic regime (n tends to $+\infty$), $J_0(\hat{w}_\lambda)$ has the following expansion $J_0(w_0) + \frac{1}{2}\left(b_2 + \frac{d_2}{n}\right)$, a result which has been obtained by several authors in several settings [31, 32]. In this asymptotic regime, the optimal λ is known to be of order $O(n^{-1})$ [33]. The main contribution of our analysis is to allow a non asymptotic analysis with explicit constants. Moreover, note that for the square loss, the bound in Eq. (14) holds with $\kappa = 0$, which can be linked to the fact that our self-concordant analysis from Propositions 1 and 2 is applicable with $R = 0$ for the square loss. Note that the constants in the previous theorem could probably be improved.

Conditions for asymptotic equivalence. In order to have the remainder term in Eq. (14) negligible with high probability compared to the lowest order term in the expansion of $J_0(\hat{w}_\lambda)$, we need to have $d_2 + nb_2$ large and κ small (so that v can be taken taking small while $v^2(d_2 + nb_2)$ is large, and hence we have a result with high-probability). The assumption that $d_2 + nb_2$ grows unbounded when n tends to infinity is a classical assumption in the study of smoothing splines and RKHSs [34, 35], and simply states that the convergence rate of the excess risk $J_0(\hat{w}_\lambda) - J_0(w_0)$, i.e., $b_2 + d_2/n$, is slower than for parametric estimation, i.e., slower than n^{-1}.

Study of parameter κ. First, we always have $\kappa \geqslant \frac{R}{\lambda^{1/2}}\big(\frac{d_1}{n}+b_1\big)^{1/2}$; thus an upper bound on κ implies an upperbound on $\frac{d_1}{n}+b_1$ which is needed in the proof of Theorem 2 to show that the Newton decrement is small enough. Moreover, κ is bounded by the sum of $\kappa_{\text{bias}} = \frac{R}{\lambda^{1/2}} b_1 b_2^{-1/2}$ and $\kappa_{\text{var}} = \frac{R}{\lambda^{1/2}}\big(\frac{d_1}{n}\big)\big(\frac{d_2}{n}\big)^{-1/2}$. Under simple assumptions on the eigenvalues of Q or equivalently of $\text{Diag}(\sigma_i) K \,\text{Diag}(\sigma_i)$, one can show that κ_{var} is small. For example, if d of these eigenvalues are equal to one and the remaining ones are zero, then, $\kappa_{\text{var}} = \frac{R d^{1/2}}{\lambda^{1/2} n^{1/2}}$. And thus we simply need λ asymptotically greater than $R^2 d/n$. For additional conditions for κ_{var}, see [28, 30]. A simple condition for κ_{bias} can be obtained if $w_0^\top Q^{-1} w_0$ is assumed bounded (in the context of RKHSs this is a stricter condition that the generating function is inside the RKHS, and is used by [36] in the context of sparsity-inducing norms). In this case, the bias terms are negligible compared to the variance term as soon as λ is asymptotically greater than $n^{-1/2}$.

Variance term. Note that the diagonal matrix $\text{Diag}(\sigma_i^2)$ is upperbounded by $\frac{1}{4}I$, i.e., $\text{Diag}(\sigma_i^2) \preccurlyeq \frac{1}{4}I$, so that the degrees of freedom for logistic regression are always less than the corresponding ones for least-squares regression (for λ multiplied by 4). Indeed, the pairs (x_i, y_i) for which the conditional distribution is close to deterministic are such that σ_i^2 is close to zero. And thus it should reduce the variance of the estimator, as little noise is associated with these points, and the effect of this reduction is exactly measured by the reduction in the degrees of freedom.

Moreover, the rate of convergence d_2/n of the variance term has been studied by many authors (see, e.g., [22, 25, 30]) and depends on the decay of the eigenvalues of Q (the faster the decay, the smaller d_2). The degrees of freedom usually grows with n, but in many cases is slower than $n^{1/2}$, leading to faster rates in Eq. (14).

4.4 Smoothing parameter selection

In this section, we obtain a criterion similar to Mallow's C_L [37] to estimate the generalization error and select in a data-driven way the regularization parameter λ (referred to as the smoothing parameter when dealing with splines or RKHSs). The following theorem shows that with a data-dependent criterion, we may obtain a good estimate of the generalization performance, up to a constant term $q^\top w_0$ independent of λ (see proof in Appendix D):

Theorem 3 (Data-driven estimation of generalization performance) *Assume* ***(A1)***, ***(A2)*** *and* ***(A3)***. *Let $\hat{Q}_\lambda = \frac{1}{n}\sum_{i=1}^n \ell''(\hat{w}_\lambda^\top x_i) x_i x_i^\top$ and $q = \frac{1}{n}\sum_{i=1}^n (y_i/2 - \mathbb{E}(y_i/2)) x_i$. Assume moreover $\kappa \leqslant 1/16$, where κ is defined in Eq. (13). If $v \in [0, 1/4]$ satisfies $v^3(d_2 + nb_2)^{1/2} \leqslant 12$, then, with probability at least $1 - \exp(-v^2(d_2 + nb_2))$:*

$$\left| J_0(\hat{w}_\lambda) - \hat{J}_0(\hat{w}_\lambda) - \frac{1}{n} \operatorname{tr} \hat{Q}_\lambda (\hat{Q}_\lambda + \lambda I)^{-1} - q^\top w_0 \right| \leqslant \left(b_2 + \frac{d_2}{n} \right) (69v + 2560\kappa).$$

The previous theorem, which is essentially a non-asymptotic version of results in [31, 32] can be further extended to obtain oracle inequalities when minimizing the data-driven criterion $\hat{J}_0(\hat{w}_\lambda) + \frac{1}{n} \operatorname{tr} \hat{Q}_\lambda (\hat{Q}_\lambda + \lambda I)^{-1}$, similar to results obtained in [35, 28] for the square loss. Note that contrary to least-squares regression with Gaussian noise, there is no need to estimate the unknown noise variance (of course only when the logistic model is actually well-specified); however, the matrix Q used to define the degrees of freedom does depend on w_0 and thus requires that $\hat{Q}_\lambda$ is used as an estimate. Finally, criteria based on generalized cross-validation [38, 4] could be studied with similar tools.

5 Regularization by the ℓ_1-norm

In this section, we consider an estimator $\hat{w}_\lambda$ obtained as a minimizer of the ℓ_1-regularized empirical risk, i.e., $\hat{J}_0(w) + \lambda \|w\|_1$. It is well-known that the estimator has some zero components [39]. In this section, we extend some of the recent results [12, 13, 14, 15, 16, 40] for the square loss (i.e., the Lasso) to the logistic loss. We assume throughout this section that the model is well-specified, that is, that the observations y_i, $i = 1, \dots, n$, are generated according to the logistic model $\mathbb{P}(y_i = \varepsilon_i) = \sigma(\varepsilon_i w_0^\top x_i)$.

We denote by $K = \{j \in \{1, \dots, p\}, (w_0)_j \neq 0\}$ the set of non-zero components of w_0 and $s = \operatorname{sign}(w_0) \in \{-1, 0, 1\}^p$ the vector of signs of w_0. On top of Assumptions **(A1)**, **(A2)** and **(A3)**, we will make the following assumption regarding normalization for each covariate (which can always be imposed by renormalization), i.e.,

(A4) *Normalized covariates*: for all $j = 1, \dots, p$, $\frac{1}{n} \sum_{i=1}^n [(x_i)_j]^2 \leqslant 1$.

In this section, we consider two different results, one on model consistency (Section 5.1) and one on efficiency (Section 5.2). As for the square loss, they will both depend on additional assumptions regarding the square $p \times p$ matrix $Q = \frac{1}{n} \sum_{i=1}^n \ell''(w_0^\top x_i) x_i x_i^\top$. This matrix is a weighted Gram matrix, which corresponds to the unweighted one for the square loss. As already shown in [5, 3], usual assumptions for the Gram matrix for the square loss are extended, for the logistic loss setting using the weighted Gram matrix Q. In this paper, we consider two types of results based on specific assumptions on Q, but other ones could be considered as well (such as [41]). The main contribution of using self-concordant analysis is to allow simple extensions from the square loss with short proofs and sharper bounds, in particular by avoiding an exponential constant in the maximal value of $|w_0^\top x_i|$, $i = 1, \dots, n$.

5.1 Model consistency condition

The following theorem provides a sufficient condition for model consistency. It is based on the *consistency condition* $\|Q_{K^cK} Q_{KK}^{-1} s_K\|_\infty < 1$, which is exactly the same as the one for the square loss [15, 12, 14] (see proof in Appendix E):

Theorem 4 (Model consistency for ℓ_1-regularization) *Assume* ***(A1)***, ***(A2)***, ***(A3)*** *and* ***(A4)***. *Assume that there exists $\eta, \rho, \mu > 0$ such that*

$$\|Q_{K^cK} Q_{KK}^{-1} s_K\|_\infty \leqslant 1 - \eta, \tag{15}$$

$\lambda_{\min}(Q_{KK}) \geqslant \rho$ and $\min_{j\in K} |(w_0)_j| \geqslant \mu$. Assume $\lambda \leqslant \min\left\{\frac{\rho\mu}{4|K|^{1/2}}, \frac{\eta\rho^{3/2}}{64R|K|}\right\}$. Then the probability that the vector of signs of $\hat{w}_\lambda$ is different from $s = \text{sign}(w_0)$ is upperbounded by

$$2p\exp\left(-\frac{n\lambda^2\eta^2}{16}\right) + 2|K|\exp\left(-\frac{n\rho^2\mu^2}{16|K|}\right) + 2|K|\exp\left(-\frac{\lambda n\rho^{3/2}\eta}{64R|K|}\right). \qquad (16)$$

Comparison with square loss. For the square loss, the previous theorem simplifies [15, 12]: with our notations, the constraint $\lambda \leqslant \frac{\eta\rho^{3/2}}{64R|K|}$ and the last term in Eq. (16), which are the only ones depending on R, can be removed (indeed, the square loss allows the application of our adapted self-concordant analysis with the constant $R = 0$). On the one hand, the favorable scaling between p and n, i.e., $\log p = O(n)$ for a certain well-chosen λ, is preserved (since the logarithm of the added term is proportional to $-\lambda n$). However, on the other hand, the terms in R may be large as R is the radius of the entire data (i.e., with all p covariates). Bounds with the radius of the data on only the relevant features in K could be derived as well (see details in the proof in Appendix E).

Necessary condition. In the case of the square loss, a weak form of Eq. (15), i.e., $\|Q_{K^cK}Q_{KK}^{-1}s_K\|_\infty \leqslant 1$ turns out to be necessary and sufficient for asymptotic correct model selection [14]. While the weak form is clearly necessary for model consistency, and the strict form sufficient (as proved in Theorem 4), we are currently investigating whether the weak condition is also sufficient for the logistic loss.

5.2 Efficiency

Another type of result has been derived, based on different proof techniques [16] and aimed at efficiency (i.e., predictive performance). Here again, we can extend the result in a very simple way. We assume, given K the set of non-zero components of w_0:

(A5) *Restricted eigenvalue condition*:

$$\rho = \min_{\|\Delta_{K^c}\|_1 \leqslant 3\|\Delta_K\|_1} \frac{(\Delta^\top Q\Delta)^{1/2}}{\|\Delta_K\|_2} > 0.$$

Note that the assumption made in [16] is slightly stronger but only depends on the cardinality of K (by minimizing with respect to all sets of indices with cardinality equal to the one of K). The following theorem provides an estimate of the estimation error as well as an oracle inequality for the generalization performance (see proof in Appendix F):

Theorem 5 (Efficiency for ℓ_1-regularization) *Assume **(A1)**, **(A2)**, **(A3)**, **(A4)**, and **(A5)**. For all $\lambda \leqslant \frac{\rho^2}{48R|K|}$, with probability at least $1 - 2pe^{-\lambda n^2/5}$, we have:*

$$\begin{aligned} \|\hat{w}_\lambda - w_0\|_1 &\leqslant 12\lambda|K|\rho^{-2}, \\ J_0(\hat{w}_\lambda) - J_0(w_0) &\leqslant 12\lambda^2|K|\rho^{-2}. \end{aligned}$$

We obtain a result which directly mimics the one obtained in [16] for the square loss with the exception of the added bound on λ. In particular, if we take $\lambda = \sqrt{\frac{10\log(p)}{n}}$, we get with probability at least $1-2/p$, an upper bound on the generalization performance $J_0(\hat{w}_\lambda) \leqslant J_0(w_0) + 120\frac{\log p}{n}|K|\rho^{-2}$. Again, the proof of this result is a direct extension of the corresponding one for the square loss, with few additional assumptions owing to the proper self-concordant analysis.

6 Conclusion

We have provided an extension of self-concordant functions that allows the simple extensions of theoretical results for the square loss to the logistic loss. We have applied the extension techniques to regularization by the ℓ_2-norm and regularization by the ℓ_1-norm, showing that new results for logistic regression can be easily derived from corresponding results for least-squares regression, without added complex assumptions.

The present work could be extended in several interesting ways to different settings. First, for logistic regression, other extensions of theoretical results from least-squares regression could be carried out: for example, the analysis of sequential experimental design for logistic regression leads to many assumptions that could be relaxed (see, e.g., [42]). Also, other regularization frameworks based on sparsity-inducing norms could be applied to logistic regression with similar guarantees than for least-squares regression, such as group Lasso for grouped variables [43] or non-parametric problems [36], or resampling-based procedures [44, 45] that allow to get rid of sufficient consistency conditions.

Second, the techniques developed in this paper could be extended to other M-estimation problems: indeed, other generalized linear models beyond logistic regression could be considered where higher-order derivatives can be expressed through cumulants [19]. Moreover, similar developments could be made for density estimation for the exponential family, which would in particular lead to interesting developments for Gaussian models in high dimensions, where ℓ_1-regularization has proved useful [46, 47]. Finally, other losses for binary or multi-class classification are of clear interest [21], potentially with different controls of the third derivatives.

A Proofs of optimization results

We follow the proof techniques of [8], by simply changing the control of the third order derivative. We denote by $F'''(w)$ the third-order derivative of F, which is itself a function from $\mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^p$ to $\mathbb{R}$. The assumptions made in Propositions 1 and 2 are in fact equivalent to (see similar proof in [8]):

$$\forall u, v, w \in \mathbb{R}^p,\ |F'''[u,v,t]| \leqslant R\|u\|_2[v^\top F''(w)v]^{1/2}[t^\top F''(w)t]^{1/2}. \qquad (17)$$

A.1 Univariate functions

We first consider univariate functions and prove the following lemma that gives upper and lower Taylor expansions:

Lemma 1 *Let g be a convex three times differentiable function $g : \mathbb{R} \mapsto \mathbb{R}$ such that for all $t \in \mathbb{R}$, $|g'''(t)| \leqslant Sg''(t)$, for some $S \geqslant 0$. Then, for all $t \geqslant 0$:*

$$\frac{g''(0)}{S^2}(e^{-St} + St - 1) \leqslant g(t) - g(0) - g'(0)t \leqslant \frac{g''(0)}{S^2}(e^{St} - St - 1). \tag{18}$$

Proof Let us first assume that $g''(t)$ is strictly positive for all $t \in \mathbb{R}$. We have, for all $t \geqslant 0$: $-S \leqslant \frac{d \log g''(t)}{dt} \leqslant S$. Then, by integrating once between 0 and t, taking exponentials, and then integrating twice:

$$-St \leqslant \log g''(t) - \log g''(0) \leqslant St,$$

$$g''(0)e^{-St} \leqslant g''(t) \leqslant g''(0)e^{St}, \tag{19}$$

$$g''(0)S^{-1}(1 - e^{-St}) \leqslant g'(t) - g'(0) \leqslant g''(0)S^{-1}(e^{St} - 1),$$

$$g(t) \geqslant g(0) + g'(0)t + g''(0)S^{-2}(e^{-St} + St - 1), \tag{20}$$

$$g(t) \leqslant g(0) + g'(0)t + g''(0)S^{-2}(e^{St} - St - 1), \tag{21}$$

which leads to Eq. (18).

Let us now assume only that $g''(0) > 0$. If we denote by A the connected component that contains 0 of the open set $\{t \in \mathbb{R},\ g''(t) > 0\}$, then the preceding developments are valid on A; thus, Eq. (19) implies that A is not upper-bounded. The same reasoning on $-g$ ensures that $A = \mathbb{R}$ and hence $g''(t)$ is strictly positive for all $t \in \mathbb{R}$. Since the problem is invariant by translation, we have shown that if there exists $t_0 \in \mathbb{R}$ such that $g''(t_0) > 0$, then for all $t \in \mathbb{R}$, $g''(t) > 0$.

Thus, we need to prove Eq. (18) for g'' always strictly positive (which is done above) and for g'' identically equal to zero, which implies that g is linear, which is then equivalent to Eq. (18). ■

Note the difference with a classical uniform bound on the third derivative, which leads to a third-order polynomial lower bound, which tends to $-\infty$ more quickly than Eq. (20). Moreover, Eq. (21) may be interpreted as an upperbound on the remainder in the Taylor expansion of g around 0:

$$g(t) - g(0) - g'(0)t - \frac{g''(0)}{2}t^2 \leqslant g''(0)S^{-2}(e^{St} - \frac{1}{2}S^2t^2 - St - 1).$$

The right hand-side is equivalent to $\frac{St^3}{6}g''(0)$ for t close to zero (which should be expected from a three-times differentiable function such that $g'''(0) \leqslant Sg''(0)$), but still provides a good bound for t away from zero (which cannot be obtained from a regular Taylor expansion).

Throughout the proofs, we will use the fact that the functions $u \mapsto \frac{e^u - 1}{u}$ and $u \mapsto \frac{e^u - 1 - u}{u^2}$ can be extended to continuous functions on $\mathbb{R}$, which are thus bounded on any compact. The bound will depend on the compact and can be obtained easily.

A.2 Proof of Proposition 1

By applying Lemma 1 (Eq. (20) and Eq. (21)) to $g(t) = F(w+tv)$ (with constant $S = R\|v\|_2$) and taking $t = 1$, we get the desired first two inequalities in Eq. (3) and Eq. (4). By considering the function $g(t) = u^\top F''(w+tv)u$, we have $g'(t) = F'''(w+tv)[u,u,v]$, which is such that $|g'(t)| \leqslant \|v\|_2 R g(t)$, leading to $g(0)e^{-\|v\|_2 Rt} \leqslant g(t) \leqslant g(0)e^{\|v\|_2 Rt}$, and thus to Eq. (6) for $t = 1$ (when considered for all $u \in \mathbb{R}^p$).

In order to prove Eq. (5), we consider $h(t) = z^\top(F'(w+tv) - F'(w) - F''(w)vt)$. We have $h(0) = 0$, $h'(0) = 0$ and $h''(t) = F'''(w+tv)[v,v,z] \leqslant R\|v\|_2 e^{tR\|v\|_2}[z^\top F''(w)z]^{1/2}[v^\top F''(w)v]^{1/2}$ using Eq. (6) and Eq. (17). Thus, by integrating between 0 and t,

$$h'(t) \leqslant [z^\top F''(w)z]^{1/2}[v^\top F''(w)v]^{1/2}(e^{tR\|v\|_2} - 1),$$

which implies $h(1) \leqslant [z^\top F''(w)z]^{1/2}[v^\top F''(w)v]^{1/2}\int_0^1 (e^{tR\|v\|_2} - 1)dt$, which in turn leads to Eq. (5).

Using similar techniques, i.e., by considering the function $t \mapsto= z^\top[F''(w+tv) - F''(w)]u$, we can prove that for all $z, u, v, w \in \mathbb{R}^p$, we have:

$$z^\top[F''(w+v) - F''(w)]u \leqslant \frac{e^{R\|v\|_2} - 1}{\|v\|_2}[v^\top F''(w)v]^{1/2}[z^\top F''(w)z]^{1/2}\|u\|_2. \qquad (22)$$

A.3 Proof of Proposition 2

Since we have assumed that $\lambda(w) > 0$, then by Eq. (6), the Hessian of F is everywhere invertible, and hence the function F is strictly convex. Therefore, if the minimum is attained, it is unique.

Let $v \in \mathbb{R}^p$ be such that $v^\top F''(w)v = 1$. Without loss of generality, we may assume that $F'(w)^\top v$ is negative. This implies that for all $t \leqslant 0$, $F(w+tv) \geqslant F(w)$. Moreover, let us denote $\kappa = -v^\top F'(w)R\|v\|_2$, which is nonnegative and such that $\kappa \leqslant \frac{R|v^\top F'(w)|}{\lambda(w)^{1/2}} \leqslant \frac{R\nu(F,w)}{\lambda(w)^{1/2}} \leqslant 1/2$. From Eq. (3), for all $t \geqslant 0$, we have:

$$\begin{aligned} F(w+tv) &\geqslant F(w) + v^\top F'(w)t + \frac{1}{R^2\|v\|_2^2}(e^{-R\|v\|_2 t} + R\|v\|_2 t - 1) \\ &\geqslant F(w) + \frac{1}{R^2\|v\|_2^2}\left[e^{-R\|v\|_2 t} + (1-\kappa)R\|v\|_2 t - 1\right]. \end{aligned}$$

Moreover, a short calculation shows that for all $\kappa \in (0,1]$:

$$e^{-2\kappa(1-\kappa)^{-1}} + (1-\kappa)2\kappa(1-\kappa)^{-1} - 1 \geqslant 0. \qquad (23)$$

This implies that for $t_0 = 2(R\|v\|_2)^{-1}\kappa(1-\kappa)^{-1}$, $F(w+t_0 v) \geqslant F(w)$. Since $t_0 \leqslant \frac{2}{1-\kappa}|v^\top F'(w)| \leqslant 2\nu(F,w)\left(1 - \frac{\nu(F,w)R}{\lambda(w)^{1/2}}\right)^{-1} \leqslant 4\nu(F,w)$, we have $F(w+tv) \geqslant F(w)$ for $t = 4\nu(F,w)$.

Since this is true for all v such that $v^\top F''(w)v = 1$, this shows that the value of the function F on the entire ellipsoid (since $F''(w)$ is positive definite) $v^\top F''(w)v = 16\nu(F,w)^2$ is greater or equal to the value at w; thus, by convexity, there must be a minimizer w^*—which is unique because of Eq. (6)—of F such that

$$(w-w^*)^\top F''(w)(w-w^*) \leqslant 16\nu(F,w)^2,$$

leading to Eq. (7).

In order to prove Eq. (9), we will simply apply Eq. (7) at $w+v$, which requires to upper-bound $\nu(F,w+v)$. If we denote by $v = -F''(w)^{-1}F'(w)$ the Newton step, we have:

$$\begin{aligned}
& \|F''(w)^{-1/2}F'(w+v)\|_2 \\
= \;& \big\|F''(w)^{-1/2}[F'(w+v) - F'(w) - F''(w)v]\big\|_2 \\
= \;& \Big\| \int_0^1 F''(w)^{-1/2}[F''(w+tv) - F''(w)]v dt \Big\|_2 \\
\leqslant \;& \int_0^1 \Big\| F''(w)^{-1/2}[F''(w+tv) - F''(w)]F''(w)^{-1/2}F''(w)^{1/2}v \Big\|_2 dt \\
\leqslant \;& \int_0^1 \Big\| \Big[F''(w)^{-1/2}F''(w+tv)F''(w)^{-1/2} - I\Big] F''(w)^{1/2}v \Big\|_2 dt.
\end{aligned}$$

Moreover, we have from Eq. (6):

$$(e^{-tR\|v\|_2} - 1)I \preccurlyeq F''(w)^{-1/2}F''(w+tv)F''(w)^{-1/2} - I \preccurlyeq (e^{tR\|v\|_2} - 1)I.$$

Thus,

$$\begin{aligned}
\|F''(w)^{-1/2}F'(w+v)\|_2 &\leqslant \int_0^1 \max\{e^{tR\|v\|_2} - 1, 1 - e^{-tR\|v\|_2}\}\|F''(w)^{1/2}v\|_2 dt \\
&= \nu(F,w)\int_0^1 (e^{tR\|v\|_2} - 1)dt = \nu(F,w)\frac{e^{R\|v\|_2} - 1 - R\|v\|_2}{R\|v\|_2}.
\end{aligned}$$

Therefore, using Eq. (6) again, we obtain:

$$\nu(F,w+v) = \|F''(w+v)^{-1/2}F'(w+v)\|_2 \leqslant \nu(F,w)e^{R\|v\|_2/2}\frac{e^{R\|v\|_2} - 1 - R\|v\|_2}{R\|v\|_2}.$$

We have $R\|v\|_2 \leqslant R\lambda^{-1/2}\nu(F,w) \leqslant 1/2$, and thus, we have

$$e^{R\|v\|_2/2}\frac{e^{R\|v\|_2} - 1 - R\|v\|_2}{R\|v\|_2} \leqslant R\|v\|_2 \leqslant R\nu(F,w)\lambda(w)^{-1/2},$$

leading to:

$$\nu(F,w+v) \leqslant \frac{R}{\lambda(w)^{1/2}}\nu(F,w)^2. \tag{24}$$

Moreover, we have:

$$\begin{aligned}
\frac{R\nu(F,w+v)}{\lambda(w+v)^{1/2}} &\leqslant \frac{Re^{R\|v\|_2/2}}{\lambda(w)^{1/2}}\nu(F,w+v) \leqslant \frac{R}{\lambda(w)^{1/2}}\nu(F,w)e^{R\|v\|_2}\frac{e^{R\|v\|_2}-1-R\|v\|_2}{R\|v\|_2}, \\
&\leqslant \frac{R}{\lambda(w)^{1/2}}\nu(F,w) \times R\|v\|_2 \leqslant \left(\frac{R}{\lambda(w)^{1/2}}\nu(F,w)\right)^2 \leqslant 1/4,
\end{aligned}$$

which leads to Eq. (8). Moreover, it shows that we can apply Eq. (7) at $w+v$ and get:

$$\begin{aligned}
& [(w^*-w-v)^\top F''(w)(w^*-w-v)]^{1/2} \\
\leqslant\ & e^{R\|v\|_2/2}[(w^*-w-v)^\top F''(w+v)(w^*-w-v)]^{1/2} \\
\leqslant\ & 4e^{R\|v\|_2/2}\nu(F,w+v) \leqslant 4R\|v\|_2\nu(F,w),
\end{aligned}$$

which leads to the desired result, i.e., Eq. (9).

B Proof of Theorem 1

Following [26, 27], we denote by w_λ the unique global minimizer of the expected regularized risk $J_\lambda(w) = J_0(w) + \frac{\lambda}{2}\|w\|_2^2$. We simply apply Eq. (7) from Proposition 2 to $\hat{J}_\lambda$ and w_λ, to obtain, if the Newton decrement (see Section 2 for its definition) $\nu(\hat{J}_\lambda, w_\lambda)^2$ is less than $\lambda/4R^2$, that $\hat{w}_\lambda$ and its population counterpart w_λ are close, i.e.:

$$(\hat{w}_\lambda - w_\lambda)^\top J''_\lambda(w_\lambda)(\hat{w}_\lambda - w_\lambda) \leqslant 16\nu(\hat{J}_\lambda, w_\lambda)^2.$$

We can then apply the upper Taylor expansion in Eq. (4) from Proposition 1 to J_λ and w_λ, to obtain, with $v = \hat{w}_\lambda - w_\lambda$ (which is such that $R\|v\|_2 \leqslant 4\frac{R\nu(\hat{J}_\lambda,w_\lambda)}{\lambda^{1/2}} \leqslant 2$):

$$J_\lambda(\hat{w}_\lambda) - J_\lambda(w_\lambda) \leqslant \frac{v^\top J''_\lambda(w_\lambda)v}{R^2\|v\|_2^2}(e^{R\|v\|_2} - R\|v\|_2 - 1) \leqslant 20\nu(\hat{J}_\lambda, w_\lambda)^2.$$

Therefore, for any $w_0 \in \mathbb{R}^p$, since w_λ is the minimizer of $J_\lambda(w) = J_0(w) + \frac{\lambda}{2}\|w\|_2^2$:

$$J_0(\hat{w}_\lambda) \leqslant J_0(w_0) + \frac{\lambda}{2}\|w_0\|_2^2 + 20\nu(\hat{J}_\lambda, w_\lambda)^2. \tag{25}$$

We can now apply the concentration inequality from Proposition 4 in Appendix G, i.e., Eq. (42), with $u = \log(8/\delta)$. We use $\lambda = 19R^2\sqrt{\frac{\log(8/\delta)}{n}}$. In order to actually have $\nu(\hat{J}_\lambda, w_\lambda) \leqslant \lambda^{1/2}/2R$ (so that we can apply our self-concordant analysis), it is sufficient that:

$$41R^2u/\lambda n \leqslant \lambda/8R^2,\ 63(u/n)^{3/2}R^2/\lambda \leqslant \lambda/16R^2,\ 8(u/n)^2R^2/\lambda \leqslant \lambda/16R^2,$$

leading to the constraints $u \leqslant n/125$. We then get with probability at least $1-\delta = 1-8e^{-u}$ (for $u \leqslant n/125$):

$$J_0(\hat{w}_\lambda) \leqslant J_0(w_0) + \frac{\lambda}{2}\|w_0\|_2^2 + 20\frac{\lambda}{4R^2} \leqslant J_0(w_0) + \frac{(10+100R^2\|w_0\|_2^2)\sqrt{\log(8/\delta)}}{\sqrt{n}}.$$

For $u > n/125$, the bound in Eq. (12) is always satisfied. Indeed, this implies with our choice of λ that $\lambda \geqslant R^2$. Moreover, since $\|\hat{w}_\lambda\|_2^2$ is bounded from above by $\log(2)\lambda^{-1} \leqslant R^{-2}$,

$$J_0(\hat{w}_\lambda) \leqslant J_0(w_0) + \frac{R^2}{2}\|\hat{w}_\lambda - w_0\|_F^2 \leqslant J_0(w_0) + 1 + R^2\|w_0\|_2^2,$$

which is smaller than the right hand-side of Eq. (12).

C Proof of Theorem 2

We denote by J_0^T the second-order Taylor expansion of J_0 around w_0, equal to $J_0^T(w) = J_0(w_0) + \frac{1}{2}(w-w_0)^\top Q(w-w_0)$, with $Q = J_0''(w_0)$, and $\hat{J}_0^T$ the expansion of $\hat{J}_0$ around w_0, equal to $J_0^T(w) - q^\top w$. We denote by $\hat{w}_\lambda^N$ the one-step Newton iterate from w_0 for the function $\hat{J}_0$, defined as the global minimizer of $\hat{J}_0^T$ and equal to $\hat{w}_\lambda^N = w_0 + (Q+\lambda I)^{-1}(q-\lambda w_0)$.

What the following proposition shows is that we can replace $\hat{J}_0$ by $\hat{J}_0^T$ for obtaining the estimator and that we can replace J_0 by J_0^T for measuring its performance, i.e., we may do as if we had a weighted least-squares cost, as long as the Newton decrement is small enough:

Proposition 3 (Quadratic approximation of risks) *Assume* $\nu(\hat{J}_\lambda, w_0)^2 = (q-\lambda w_0)^\top (Q+\lambda I)^{-1}(q-\lambda w_0) \leqslant \frac{\lambda}{4R^2}$. *We have:*

$$|J_0(\hat{w}_\lambda) - J_0^T(\hat{w}_\lambda^N)| \leqslant \frac{15R\nu(\hat{J}_\lambda, w_0)^2}{\lambda^{1/2}}\|Q^{1/2}(\hat{w}_\lambda^N - w_0)\|_2 + \frac{40R^2}{\lambda}\nu(\hat{J}_\lambda, w_0)^4. \quad (26)$$

Proof We show that (1) $\hat{w}_\lambda^N$ is close to $\hat{w}_\lambda$ using Proposition 2 on the behavior of Newton's method, (2) that $\hat{w}_\lambda^N$ is close to w_0 by using its closed form $\hat{w}_\lambda^N = w_0 + (Q+\lambda I)^{-1}(q-\lambda w_0)$, and (3) that J_0 and J_0^T are close using Proposition 1 on upper and lower Taylor expansions.

We first apply Eq. (9) from Proposition 2 to get

$$(\hat{w}_\lambda - \hat{w}_\lambda^N)^\top \hat{J}_\lambda''(w_0)(\hat{w}_\lambda - \hat{w}_\lambda^N) \leqslant \frac{16R^2}{\lambda}\nu(\hat{J}_\lambda, w_0)^4. \quad (27)$$

This implies that $\hat{w}_\lambda$ and $\hat{w}_\lambda^N$ are close, i.e.,

$$\begin{aligned}\|\hat{w}_\lambda - \hat{w}_\lambda^N\|^2 &\leqslant \lambda^{-1}(\hat{w}_\lambda - \hat{w}_\lambda^N)^\top \hat{J}_\lambda''(w_0)(\hat{w}_\lambda - \hat{w}_\lambda^N) \\ &\leqslant \frac{16R^2}{\lambda^2}\nu(\hat{J}_\lambda, w_0)^4 \leqslant \frac{4}{\lambda}\nu(\hat{J}_\lambda, w_0)^2 \leqslant \frac{1}{R^2}.\end{aligned}$$

Thus, using the closed form expression for $\hat{w}_\lambda^N = w_0 + (Q+\lambda I)^{-1}(q - \lambda w_0)$, we obtain

$$\begin{aligned} \|\hat{w}_\lambda - w_0\| &\leqslant \|\hat{w}_\lambda - \hat{w}_\lambda^N\| + \|w_0 - \hat{w}_\lambda^N\| \\ &\leqslant 2\frac{\nu(\hat{J}_\lambda, w_0)}{\lambda^{1/2}} + \frac{\nu(\hat{J}_\lambda, w_0)}{\lambda^{1/2}} \leqslant \frac{3\nu(\hat{J}_\lambda, w_0)}{\lambda^{1/2}} \leqslant \frac{3}{2R}. \end{aligned}$$

We can now apply Eq. (3) from Proposition 2 to get for all v such that $R\|v\|_2 \leqslant 3/2$,

$$|J_0(w_0+v) - J_0^T(w_0+v)| \leqslant (v^\top Q v) R \|v\|_2/4. \tag{28}$$

Thus, using Eq. (28) for $v = \hat{w}_\lambda - w_0$ and $v = \hat{w}_\lambda^N - w_0$:

$$\begin{aligned} & |J_0(\hat{w}_\lambda) - J_0^T(\hat{w}_\lambda^N)| \\ \leqslant\ & |J_0(\hat{w}_\lambda) - J_0^T(\hat{w}_\lambda)| + |J_0^T(\hat{w}_\lambda^N) - J_0^T(\hat{w}_\lambda)|, \\ \leqslant\ & \frac{R}{4}\|\hat{w}_\lambda - w_0\|_2 \, \|Q^{1/2}(\hat{w}_\lambda - w_0)\|_2^2 + \frac{1}{2}\left| \|Q^{1/2}(\hat{w}_\lambda - w_0)\|_2^2 - \|Q^{1/2}(\hat{w}_\lambda^N - w_0)\|_2^2 \right|, \\ \leqslant\ & \frac{3R\nu(\hat{J}_\lambda, w_0)}{4\lambda^{1/2}}\|Q^{1/2}(\hat{w}_\lambda - w_0)\|_2^2 + \frac{1}{2}\left| \|Q^{1/2}(\hat{w}_\lambda - w_0)\|_2^2 - \|Q^{1/2}(\hat{w}_\lambda^N - w_0)\|_2^2 \right|, \\ \leqslant\ & \frac{3R\nu(\hat{J}_\lambda, w_0)}{4\lambda^{1/2}}\|Q^{1/2}(\hat{w}_\lambda^N - w_0)\|_2^2 + \left(\tfrac{1}{2} + \tfrac{3}{4}\right)\left| \|Q^{1/2}(\hat{w}_\lambda - w_0)\|_2^2 - \|Q^{1/2}(\hat{w}_\lambda^N - w_0)\|_2^2 \right|, \\ \leqslant\ & \frac{3R\nu(\hat{J}_\lambda, w_0)}{4\lambda^{1/2}}\|Q^{1/2}(\hat{w}_\lambda^N - w_0)\|_2^2 \\ & + \frac{5}{4}\|Q^{1/2}(\hat{w}_\lambda - \hat{w}_\lambda^N)\|_2^2 + \frac{5}{2}\|Q^{1/2}(\hat{w}_\lambda - \hat{w}_\lambda^N)\|_2 \|Q^{1/2}(\hat{w}_\lambda^N - w_0)\|_2. \end{aligned}$$

From Eq. (27), we have $\|Q^{1/2}(\hat{w}_\lambda - \hat{w}_\lambda^N)\|_2^2 \leqslant \frac{16R^2}{\lambda}\nu(\hat{J}_\lambda, w_0)^4$. We thus obtain, using that $\|Q^{1/2}(\hat{w}_\lambda^N - w_0)\|_2 \leqslant \nu(\hat{J}_0, w_0)$:

$$|J_0(\hat{w}_\lambda) - J_0^T(\hat{w}_\lambda^N)| \leqslant \left(\frac{3}{4} + \frac{5}{2}\sqrt{32}\right) \frac{\nu(\hat{J}_\lambda, w_0)^2}{R^{-1}\lambda^{1/2}} \|Q^{1/2}(\hat{w}_\lambda^N - w_0)\|_2 + \frac{40R^2}{\lambda}\nu(\hat{J}_\lambda, w_0)^4,$$

which leads to the desired result. ■

We can now go on with the proof of Theorem 2. From Eq. (26) in Proposition 3 above, we have, if $\nu(\hat{J}_\lambda, w_0)^2 \leqslant \lambda/4R^2$,

$$\begin{aligned} J_0(\hat{w}_\lambda) &= J_0^T(\hat{w}_\lambda^N) + B \\ &= J_0(w_0) + \frac{1}{2}(q - \lambda w_0)^\top Q (Q + \lambda I)^{-2}(q - \lambda w_0) + B \\ &= J_0(w_0) + \frac{d_2}{2n} + \frac{b_2}{2} + B + C, \end{aligned}$$

$$\begin{aligned} \text{with } C &= \lambda w_0^\top (Q+\lambda I)^{-2} Q q + \frac{1}{2}\operatorname{tr}(Q+\lambda I)^{-2} Q \left(q q^\top - \frac{1}{n} Q \right), \\ |B| &\leqslant \frac{15 R \nu(\hat{J}_\lambda, w_0)^2}{\lambda^{1/2}} \|Q^{1/2}(\hat{w}_\lambda^N - w_0)\|_2 + \frac{40R^2}{\lambda}\nu(\hat{J}_\lambda, w_0)^4. \end{aligned}$$

We can now bound each term separately and check that we indeed have $\nu(\hat{J}_\lambda, w_0)^2 \leqslant \lambda/4R^2$ (which allows to apply Proposition 2). First, from Eq. (13), we can derive

$$b_2 + \frac{d_2}{n} \leqslant b_1 + \frac{d_1}{n} \leqslant \frac{\kappa \lambda^{1/2}}{R}\big(b_2 + \frac{d_2}{n}\big)^{1/2} \leqslant \frac{\kappa \lambda^{1/2}}{R}\big(b_1 + \frac{d_1}{n}\big)^{1/2},$$

which implies the following identities:

$$b_2 + \frac{d_2}{n} \leqslant b_1 + \frac{d_1}{n} \leqslant \frac{\kappa^2 \lambda}{R^2}. \tag{29}$$

We have moreover:

$$\begin{aligned} \nu(\hat{J}_\lambda, w_0)^2 &= (q - \lambda w_0)^\top (Q + \lambda I)^{-1} (q - \lambda w_0) \\ &\leqslant b_1 + \frac{d_1}{n} + \operatorname{tr}\big(Q + \lambda I\big)^{-1}\bigg(qq^\top - \frac{Q}{n}\bigg) + 2\lambda w_0^\top (Q + \lambda I)^{-1} q. \end{aligned}$$

We can now apply concentration inequalities from Appendix G, together with the following applications of Bernstein's inequality. Indeed, we have $\lambda w_0^\top (Q + \lambda I)^{-2} Q q = \sum_{i=1}^n Z_i$, with

$$\begin{aligned} |Z_i| &\leqslant \frac{\lambda}{n} |w_0^\top (Q + \lambda I)^{-2} Q x_i| \\ &\leqslant \frac{\lambda}{2n}\Big(w_0^\top (Q + \lambda I)^{-2} Q w_0\Big)^{1/2}\Big(x_i^\top (Q + \lambda I)^{-2} Q x_i\Big)^{1/2} \leqslant \frac{b_2^{1/2}}{2n} R \lambda^{-1/2}. \end{aligned}$$

Moreover, $\mathbb{E} Z_i^2 \leqslant \frac{\lambda^2}{n} w_0^\top (Q + \lambda I)^{-2} Q^3 (Q + \lambda I)^{-2} w_0 \leqslant \frac{1}{n} b_2$. We can now apply Bernstein inequality [2] to get with probability at least $1 - 2e^{-u}$ (and using Eq. (29)):

$$\lambda w_0^\top (Q + \lambda I)^{-2} Q q \leqslant \sqrt{\frac{2 b_2 u}{n}} + \frac{u}{6n} b_2^{1/2} R \lambda^{-1/2} \leqslant \sqrt{\frac{2 b_2 u}{n}} + \frac{u\kappa}{6n}.$$

Similarly, with probability at least $1 - 2e^{-u}$, we have:

$$\lambda w_0^\top (Q + \lambda I)^{-1} q \leqslant \sqrt{\frac{2 b_2 u}{n}} + \frac{u\kappa}{6n}.$$

We thus get, through the union bound, with probability at least $1 - 20e^{-u}$:

$$\begin{aligned} \nu(\hat{J}_\lambda, w_0)^2 &\leqslant \bigg(b_1 + \frac{d_1}{n}\bigg) + \bigg(\frac{32 d_2^{1/2} u^{1/2}}{n} + \frac{18u}{n} + \frac{53 R d_1^{1/2} u^{3/2}}{n^{3/2} \lambda^{1/2}} + 9\frac{R^2 u^2}{\lambda n^2}\bigg) \\ &\qquad + \bigg(2\sqrt{\frac{2 b_2 u}{n}} + \frac{\kappa u}{6n}\bigg), \\ &\leqslant b_1 + \frac{d_1}{n} + \frac{64 u^{1/2}}{n^{1/2}}\big(b_2 + \frac{d_2}{n}\big)^{1/2} + \frac{u}{n}\big(18 + \frac{\kappa}{6}\big) + \frac{R^2}{\lambda}\frac{9u^2}{n^2} \\ &\qquad + \frac{53 n^{1/2} \kappa u^{3/2}}{n^{3/2}}, \\ &\leqslant \frac{\lambda \kappa^2}{R^2} + E, \end{aligned}$$

together with $C \leqslant E$. We now take $u = (nb_2 + d_2)v^2$ and assume $v \leqslant 1/4$, $\kappa \leqslant 1/16$, and $v^3(nb_2+d_2)^{1/2} \leqslant 12$, so that, we have

$$\begin{aligned}
E &\leqslant 64v\big(b_2 + \frac{d_2}{n}\big) + v^2\big(b_2 + \frac{d_2}{n}\big)\big(18 + \frac{\kappa}{6}\big) + \frac{9R^2}{\lambda}v^4\big(b_2 + \frac{d_2}{n}\big)^2 \\
&\qquad\qquad\qquad\qquad\qquad\qquad + 53n^{1/2}\kappa v^3\big(b_2 + \frac{d_2}{n}\big)^{3/2}, \\
&\leqslant \big(b_2 + \frac{d_2}{n}\big)\bigg(64v + \big(18 + \frac{\kappa}{6}\big)v^2 + \frac{9R^2}{\lambda}v^4\frac{\lambda\kappa^2}{R^2} + 53\kappa v^3(nb_2+d_2)^{1/2}\bigg), \\
&\leqslant \big(b_2 + \frac{d_2}{n}\big)\bigg(64v + 18v^2 + \frac{\kappa}{6}v^2 + 9\kappa^2 v^4 + 53\kappa v^3(nb_2+d_2)^{1/2}\bigg), \\
&\leqslant \big(b_2 + \frac{d_2}{n}\big)\bigg(68.5v + \frac{\kappa}{6\times 16} + 9\kappa/16\times 16\times 16 + 53\kappa \times \frac{12}{64}\bigg), \\
&\leqslant \big(b_2 + \frac{d_2}{n}\big)\big(69v + 10\kappa\big) \leqslant 20\big(b_2 + \frac{d_2}{n}\big).
\end{aligned}$$

This implies that $\nu(\hat{J}_\lambda, w_0)^2 \leqslant \frac{\lambda}{R^2}\frac{20}{256} \leqslant \frac{\lambda}{4R^2}$, so that we can apply Proposition 2. Thus, by denoting $e_2 = b_2 + \frac{d_2}{n}$, $e_1 = b_1 + \frac{d_1}{n}$, and $\alpha = 69v + 10\kappa \leqslant 20$, we get a global upper bound:

$$B + |C| \leqslant e_2\alpha + \frac{40R^2}{\lambda}(e_1 + e_2\alpha)^2 + \frac{15Re_2^{1/2}}{\lambda^{1/2}}(e_1 + e_2\alpha)(1+\alpha)^{1/2}.$$

With $e_1 + e_2\alpha \leqslant e_2^{1/2}(\kappa\lambda^{1/2}/R)(1+\alpha)$, we get

$$\begin{aligned}
B + |C| &\leqslant e_2\alpha + 40\kappa^2 e_2(1+\alpha)^2 + 15\kappa e_2(1+\alpha)^{3/2} \\
&\leqslant e_2\alpha + e_2\kappa(40\times 21\times 21/16 + 15(21)^{3/2}) \leqslant e_2(69v + 2560\kappa),
\end{aligned}$$

which leads to the desired result, i.e., Eq. (14).

D Proof of Theorem 3

We follow the same proof technique than for Theorem 2 in Appendix C. We have:

$$\begin{aligned}
J_0(\hat{w}_\lambda) &= \hat{J}_0(\hat{w}_\lambda) + q^\top(\hat{w}_\lambda - w_0) + q^\top w_0 \\
&= \hat{J}_0(\hat{w}_\lambda) + q^\top(\hat{w}_\lambda - \hat{w}_\lambda^{NN}) + q^\top(\hat{w}_\lambda^N - w_0) - q^\top \hat{J}_\lambda''(\hat{w}_\lambda^N)^{-1}\hat{J}_\lambda'(\hat{w}_\lambda^N) + q^\top w_0,
\end{aligned}$$

where $\hat{w}_\lambda^{NN}$ is the two-step Newton iterate from w_0. We have, from Eq. (24), $\nu(\hat{J}_\lambda, \hat{w}_\lambda^N) \leqslant \frac{2R}{\lambda^{1/2}}\nu(\hat{J}_\lambda, w_0)^2$, which then implies (with Eq. (9)):

$$(\hat{w}_\lambda - \hat{w}_\lambda^{NN})^\top(Q + \lambda I)(\hat{w}_\lambda - \hat{w}_\lambda^{NN}) \leqslant \frac{16R^2}{\lambda}\bigg(\frac{2R}{\lambda^{1/2}}\nu(\hat{J}_\lambda, w_0)^2\bigg)^4 \leqslant \frac{512R^6\nu(\hat{J}_\lambda, w_0)^8}{\lambda^3},$$

which in turn implies

$$\begin{aligned}
|q^\top(\hat{w}_\lambda - \hat{w}_\lambda^{NN})| &\leqslant [q(Q+\lambda I)^{-1}q]^{1/2}\frac{32R^3\nu(\hat{J}_\lambda,w_0)^4}{\lambda^{3/2}} \\
&\leqslant \frac{R[q(Q+\lambda I)^{-1}q]^{1/2}}{\lambda^{1/2}}\frac{32R^2\nu(\hat{J}_\lambda,w_0)^4}{\lambda}. \qquad (30)
\end{aligned}$$

Moreover, we have from the closed-form expression of $\hat{w}_\lambda^N$:

$$\Big|q^\top(\hat{w}_\lambda^N - w_0) - \frac{d_1}{n}\Big| \leqslant \Big|\operatorname{tr}(Q+\lambda I)^{-1}(qq^\top - Q/n)\Big| + \lambda w_0^\top(Q+\lambda I)^{-1}q. \qquad (31)$$

Finally, we have, using Eq. (5) from Proposition 1:

$$\begin{aligned}
\big|q^\top \hat{J}_\lambda''(\hat{w}_\lambda^N)^{-1}\hat{J}_\lambda'(\hat{w}_\lambda^N)\big| &= \big|q^\top \hat{J}_\lambda''(\hat{w}_\lambda^N)^{-1}[\hat{J}_0'(\hat{w}_\lambda^N) - \hat{J}_0'(w_0) - Q(\hat{w}_\lambda^N - w_0)]\big| \\
&\leqslant \big[q^\top \hat{J}_\lambda''(\hat{w}_\lambda^N)^{-1}Q\hat{J}_\lambda''(\hat{w}_\lambda^N)^{-1}q\big]^{1/2}\big[\Delta^\top Q\Delta\big]^{1/2}R\|\Delta\|_2 \\
&\leqslant 2\big[q^\top Q(Q+\lambda I)^{-2}q\big]^{1/2}\|Q^{1/2}\Delta\|_2\frac{R\nu(\hat{J}_\lambda,w_0)}{\lambda^{1/2}}, \qquad (32)
\end{aligned}$$

where $\Delta = \hat{w}_\lambda^N - w_0$.

What also needs to be shown is that $\big|\operatorname{tr}\hat{Q}_\lambda(\hat{Q}_\lambda+\lambda I)^{-1} - \operatorname{tr}Q(Q+\lambda I)^{-1}\big|$ is small enough; by noting that $Q = J_0''(w_0)$, $\hat{Q}_\lambda = J_0''(w_0+v)$, and $v = \hat{w}_\lambda - w_0$, we have, using Eq. (22) from Appendix A.2:

$$\begin{aligned}
& \big|\operatorname{tr}\hat{Q}_\lambda(\hat{Q}_\lambda+\lambda I)^{-1} - \operatorname{tr}Q(Q+\lambda I)^{-1}\big| \\
= \; & \lambda\big|\operatorname{tr}\big[(\hat{Q}_\lambda+\lambda I)^{-1}(Q-\hat{Q}_\lambda)(Q+\lambda I)^{-1}\big]\big| \\
\leqslant \; & \lambda\sum_{i=1}^p\big|\delta_i^\top(\hat{Q}_\lambda+\lambda I)^{-1}(Q-\hat{Q}_\lambda)(Q+\lambda I)^{-1}\delta_i\big| \\
\leqslant \; & \lambda R\sum_{i=1}^p\|Q^{1/2}(Q+\lambda I)^{-1}\delta_i\|_2\|(\hat{Q}_\lambda+\lambda I)^{-1}\delta_i\|_2\|Q^{1/2}v\|_2 \\
\leqslant \; & \lambda^{-1/2}R\|Q^{1/2}v\|_2\sum_{i=1}^p\delta_i^\top Q(Q+\lambda I)^{-1}\delta_i = \lambda^{-1/2}R\|Q^{1/2}v\|_2 d_1. \qquad (33)
\end{aligned}$$

All the terms in Eqs. (30,31,32,33) that need to be added to obtain the required upperbound are essentially the same than the ones proof of Theorem 2 in Appendix C (with smaller constants). Thus the rest of the proof follows.

E Proof of Theorem 4

We follow the same proof technique than for the Lasso [15, 12, 14], i.e., we consider $\tilde{w}$ the minimizer of $\hat{J}_0(w) + \lambda s^\top w$ subject to $w_{K^c} = 0$ (which is unique because Q_{KK} is invertible), and (1) show that $\tilde{w}_K$ has the correct (non zero) signs and (2) that it is actually the unrestricted minimum of $\hat{J}_0(w) + \lambda \|w\|_1$ over $\mathbb{R}^p$, i.e., using optimality conditions for nonsmooth convex optimization problems [48], that $\|[\hat{J}_0'(\tilde{w})]_{K^c}\|_\infty \leqslant \lambda$. All this will be shown by replacing $\tilde{w}$ by the proper one-step Newton iterate from w_0.

Correct signs on K. We directly use Proposition 2 with the function $w_K \mapsto \hat{J}_0(w_K, 0) + \lambda s_K^\top w_K$—where $(w_K, 0)$ denotes the p-dimensional vector obtained by completing w_K by zeros—to obtain from Eq. (7):

$$(\tilde{w}_K - (w_0)_K)^\top Q_{KK} (\tilde{w}_K - (w_0)_K) \leqslant 16 (q_K - \lambda s_K)^\top Q_{KK}^{-1} (q_K - \lambda s_K) = 16 \nu^2,$$

as soon as $\nu^2 = (q_K - \lambda s_K)^\top Q_{KK}^{-1} (q_K - \lambda s_K) \leqslant \frac{\rho}{4R^2}$, and thus as soon as $q_K Q_{KK}^{-1} q_K \leqslant \frac{\rho}{8R^2}$ and $\lambda^2 s_K^\top Q_{KK}^{-1} s_K \leqslant \frac{\rho}{8R^2}$. We thus have:

$$\|\tilde{w} - w_0\|_\infty \leqslant \|\tilde{w}_K - (w_0)_K\|_2 \leqslant \rho^{-1/2} \|Q_{KK}^{1/2} (\tilde{w}_K - (w_0)_K)\|_2 \leqslant 4 \rho^{-1/2} \nu.$$

We therefore get the correct signs for the covariates indexed by K, as soon as $\|\tilde{w} - w_0\|_\infty^2 \leqslant \min_{j \in K} |(w_0)_j|^2 = \mu^2$, i.e., as soon as

$$\max \left\{ q_K Q_{KK}^{-1} q_K, \lambda^2 s_K^\top Q_{KK}^{-1} s_K \right\} \leqslant \min \left\{ \frac{\rho}{16} \mu^2, \frac{\rho}{8R^2} \right\}.$$

Note that $s_K^\top Q_{KK}^{-1} s_K \leqslant |K| \rho^{-1}$, thus it is implied by the following constraint:

$$\lambda \leqslant \frac{\rho}{4|K|^{1/2}} \min \left\{ \mu, R^{-1} \right\}, \quad (34)$$

$$q_K Q_{KK}^{-1} q_K \leqslant \frac{\rho}{16} \min \left\{ \mu^2, R^{-2} \right\}. \quad (35)$$

Gradient condition on K^c. We denote by $\tilde{w}^N$ the one-step Newton iterate from w_0 for the minimization of $\hat{J}_0(w) + \lambda s^\top w$ restricted to $w_{K^c} = 0$, equal to $\tilde{w}_K^N = (w_0)_K + Q_{KK}^{-1}(q_K - \lambda s_K)$. From Eq. (9), we get:

$$(\tilde{w}_K - \tilde{w}_K^N)^\top Q_{KK} (\tilde{w}_K - \tilde{w}_K^N) \leqslant \frac{16R^2}{\rho} \left[(q_K - \lambda s_K)^\top Q_{KK}^{-1} (q_K - \lambda s_K) \right]^2 = \frac{16 R^2 \nu^4}{\rho}.$$

We thus have

$$\begin{aligned} \|\tilde{w} - \tilde{w}^N\|_2 &\leqslant \rho^{-1/2} \frac{4R\nu^2}{\rho^{1/2}} = \frac{4R\nu^2}{\rho} \leqslant 1/R, \\ \|w_0 - \tilde{w}^N\|_2 &\leqslant \rho^{-1/2} \nu \leqslant 1/2R, \\ \|\tilde{w} - w_0\|_2 &\leqslant \|\tilde{w} - \tilde{w}^N\|_2 + \|w_0 - \tilde{w}^N\|_2 \leqslant 3\nu\rho^{-1/2} \leqslant 3R/2. \end{aligned}$$

Note that up to here, all bounds R may be replaced by the maximal ℓ_2-norm of all data points, reduced to variables in K.

In order to check the gradient condition, we compute the gradient of $\hat{J}_0$ along the directions in K^c, to obtain for all $z \in \mathbb{R}^p$, using Eq. (5) and with any v such that $R\|v\|_2 \leqslant 3/2$:

$$\frac{\left|z^\top[\hat{J}_0'(w_0+v)-\hat{T}_0'(w_0+v)]\right|}{(z^\top Qz)^{1/2}} \leqslant (v^\top Qv)^{1/2}\frac{e^{R\|v\|_2}-1-R\|v\|_2}{R\|v\|_2} \leqslant 2(v^\top Qv)^{1/2}R\|v\|_2,$$

where $\hat{T}_0'(w) = \hat{J}_0'(w_0) + \hat{J}_0''(w_0)(w - w_0)$ is the derivative of the Taylor expansion of $\hat{J}_0$ around w_0. This implies, since $\operatorname{diag}(Q) \leqslant 1/4$, the following ℓ_∞-bound on the difference $\hat{J}_0$ and its Taylor expansion:

$$\|[\hat{J}_0'(w_0+v) - \hat{T}_0'(w_0+v)]_{K^c}\|_\infty \leqslant (v^\top Qv)^{1/2}R\|v\|_2.$$

We now have,

$$\begin{aligned}
\|\hat{J}_0'(\tilde{w})_{K^c}\|_\infty \leqslant\ & \|\hat{T}_0'(\tilde{w}^N)_{K^c}\|_\infty \\
& +\|\hat{T}_0'(\tilde{w}^N)_{K^c} - \hat{T}_0'(\tilde{w})_{K^c}\|_\infty + \|\hat{T}_0'(\tilde{w})_{K^c} - \hat{J}_0'(\tilde{w})_{K^c}\|_\infty, \\
\leqslant\ & \|[\hat{J}_0'(w_0) + Q(\tilde{w}^N - w_0)]_{K^c}\|_\infty \\
& +\|[Q(\tilde{w} - \tilde{w}^N)]_{K^c}\|_\infty + R\|\tilde{w} - w_0\|_2\|Q^{1/2}(\tilde{w} - w_0)\|_2, \\
\leqslant\ & \| - q_{K^c} + Q_{K^cK}Q_{KK}^{-1}(q_K - \lambda s_K)\|_\infty \\
& +\|Q_{K^cK}Q_{KK}^{-1/2}Q_{KK}^{1/2}(\tilde{w}_K - \tilde{w}_K^N)\|_\infty + 3\nu R\rho^{-1/2}(4R\nu^2\rho^{-1/2} + \nu), \\
\leqslant\ & \|q_{K^c} - Q_{K^cK}Q_{KK}^{-1}(q_K - \lambda s_K)\|_\infty + \frac{1}{4}\|Q_{KK}^{1/2}(\tilde{w}_K - \tilde{w}_K^N)\|_2 + \frac{9R}{\rho^{1/2}}\nu^2, \\
\leqslant\ & \|q_{K^c} - Q_{K^cK}Q_{KK}^{-1}(q_K - \lambda s_K)\|_\infty + \frac{1}{4}\frac{16R}{\rho^{1/2}}\nu^2 + \frac{9R}{\rho^{1/2}}\nu^2, \\
\leqslant\ & \|q_{K^c} - Q_{K^cK}Q_{KK}^{-1}(q_K - \lambda s_K)\|_\infty + \frac{16R}{\rho^{1/2}}\nu^2.
\end{aligned}$$

Thus, in order to get $\|\hat{J}_0'(\tilde{w})_{K^c}\|_\infty \leqslant \lambda$, we need

$$\|q_{K^c} - Q_{K^cK}Q_{KK}^{-1}q_K\|_\infty \leqslant \eta\lambda/4, \tag{36}$$

and

$$\max\left\{q_K Q_{KK}^{-1}q_K, \lambda^2 s_K^\top Q_{KK}^{-1}s_K\right\} \leqslant \frac{\lambda\eta\rho^{1/2}}{64R}. \tag{37}$$

In terms of upper bound on λ we then get:

$$\lambda \leqslant \min\left\{\frac{\rho}{4|K|^{1/2}}\mu, \frac{\rho}{4|K|^{1/2}}R^{-1}, \frac{\eta\rho^{3/2}}{64R|K|}\right\},$$

which can be reduced $\lambda \leqslant \min\left\{\frac{\rho}{4|K|^{1/2}}\mu, \frac{\eta\rho^{3/2}}{64R|K|}\right\}$. In terms of upper bound on $q_K^\top Q_{KK}^{-1}q_K$ we get:

$$q_K^\top Q_{KK}^{-1}q_K \leqslant \min\left\{\frac{\rho}{16}\mu^2, \frac{\rho}{16}R^{-2}, \frac{\lambda\eta\rho^{1/2}}{64R}\right\},$$

which can be reduced to $q_K^\top Q_{KK}^{-1} q_K \leqslant \min\left\{\frac{\rho}{16}\mu^2, \frac{\lambda\eta\rho^{1/2}}{64R}\right\}$, using the constraint on λ.

We now derive and use concentration inequalities. We first use Bernstein's inequality (using for all k and i, $|(x_i)_k - Q_{kK}Q_{KK}^{-1}(x_i)_K||\varepsilon_i| \leqslant R/\rho^{1/2}$ and $Q_{kk} \leqslant 1/4$), and the union bound to get

$$\begin{aligned}\mathbb{P}(\|q_{K^c} - Q_{K^cK}Q_{KK}^{-1}q_K\|_\infty \geqslant \lambda\eta/4) &\leqslant 2p\exp\left(-\frac{n\lambda^2\eta^2/32}{1/4 + R\lambda\eta\rho^{-1/2}/12}\right)\\ &\leqslant 2p\exp\left(-\frac{n\lambda^2\eta^2}{16}\right),\end{aligned}$$

as soon as $R\lambda\eta\rho^{-1/2} \leqslant 3$, i.e., as soon as, $\lambda \leqslant 3\rho^{1/2}R^{-1}$, which is indeed satisfied because of our assumption on λ. We also use Bernstein's inequality to get

$$\mathbb{P}(q_K^\top Q_{KK}^{-1}q_K \geqslant t) \leqslant \mathbb{P}\left(\|q_K\|_\infty \geqslant \sqrt{\frac{\rho t}{|K|}}\right) \leqslant 2|K|\exp\left(-\frac{n\rho t}{|K|}\right).$$

The union bound then leads to the desired result.

F Proof of Theorem 5

We follow the proof technique of [16]. We have $\hat{J}_0(\hat{w}_\lambda) = J_0(\hat{w}_\lambda) - q^\top\hat{w}_\lambda$. Thus, because $\hat{w}_\lambda$ is a minimizer of $\hat{J}_0(w) + \lambda\|w\|_1$,

$$J_0(\hat{w}_\lambda) - q^\top\hat{w}_\lambda + \lambda\|\hat{w}_\lambda\|_1 \leqslant J_0(w_0) - q^\top w_0 + \lambda\|w_0\|_1, \tag{38}$$

which implies, since $J_0(\hat{w}_\lambda) \geqslant J_0(w_0)$:

$$\begin{aligned}\lambda\|\hat{w}_\lambda\|_1 &\leqslant \lambda\|w_0\|_1 + \|q\|_\infty\|\hat{w}_\lambda - w_0\|_1,\\ \lambda\|(\hat{w}_\lambda)_K\|_1 + \lambda\|(\hat{w}_\lambda)_{K^c}\|_1 &\leqslant \lambda\|(w_0)_K\|_1 + \|q\|_\infty\big(\|(\hat{w}_\lambda)_K - (w_0)_K\|_1 + \|(\hat{w}_\lambda)_{K^c}\|_1\big).\end{aligned}$$

If we denote by $\Delta = \hat{w}_\lambda - w_0$ the estimation error, we deduce:

$$(\lambda - \|q\|_\infty)\|\Delta_{K^c}\|_1 \leqslant (\lambda + \|q\|_\infty)\|\Delta_K\|_1.$$

If we assume $\|q\|_\infty \leqslant \lambda/2$, then, we have $\|\Delta_{K^c}\|_1 \leqslant 3\|\Delta_K\|_1$, and thus using **(A5)**, we get $\Delta^\top Q\Delta \geqslant \rho^2\|\Delta_K\|_2^2$. From Eq. (38), we thus get:

$$\begin{aligned}J_0(\hat{w}_\lambda) - J_0(w_0) &\leqslant q^\top(\hat{w}_\lambda - w_0) - \lambda\|\hat{w}_\lambda\|_1 + \lambda\|w_0\|_1,\\ J_0(w_0 + \Delta) - J_0(w_0) &\leqslant (\|q\|_\infty + \lambda)\|\Delta\|_1 \leqslant \frac{3\lambda}{2}\|\Delta\|_1.\end{aligned} \tag{39}$$

Using Eq. (3) in Proposition 1 with J_0, we obtain:

$$J_0(w_0 + \Delta) - J_0(w_0) \geqslant \frac{\Delta^\top Q\Delta}{R^2\|\Delta\|_2^2}\big(e^{-R\|\Delta\|_2} + R\|\Delta\|_2 - 1\big),$$

which implies, using $\Delta^\top Q \Delta \geqslant \rho^2 \|\Delta_K\|_2^2$ and Eq. (39):

$$\frac{\rho^2 \|\Delta_K\|_2^2}{R^2 \|\Delta\|_2^2} \big(e^{-R\|\Delta\|_2} + R\|\Delta\|_2 - 1 \big) \leqslant \frac{3\lambda}{2} \|\Delta\|_1. \tag{40}$$

We can now use, with $s = |K|$, $\|\Delta\|_2 \leqslant \|\Delta\|_1 \leqslant 4\|\Delta_K\|_1 \leqslant 4\sqrt{s}\|\Delta_K\|_2$ to get:

$$\rho^2 \big(e^{-R\|\Delta\|_2} + R\|\Delta\|_2 - 1 \big) \leqslant \frac{3\lambda}{2} \frac{(4\sqrt{s}\|\Delta_K\|_2)^2 R \|\Delta\|_2}{\|\Delta_K\|_2^2} \leqslant 24\lambda s R^2 \|\Delta\|_2.$$

This implies using Eq. (23), that $R\|\Delta\|_2 \leqslant \frac{48\lambda R s/\rho^2}{1 - 24\lambda s R/\rho^2} \leqslant 2$ a soon as $R\lambda s \rho^{-2} \leqslant 1/48$, which itself implies that $\frac{1}{(R\|\Delta\|_2)^2}\big(e^{-R\|\Delta\|_2} + R\|\Delta\|_2 - 1 \big) \geqslant 1/2$, and thus, from Eq. (40),

$$\|\Delta_K\|_2 \leqslant \frac{3\lambda}{2} \times 4\sqrt{s}\|\Delta_K\|_2.$$

The second result then follows from Eq. (39) (using Bernstein inequality for an upper bound on $\mathbb{P}(\|q\|_\infty \geqslant \lambda/2)$).

G Concentration inequalities

In this section, we derive concentration inequalities for quadratic forms of bounded random variables that extend the ones already known for Gaussian random variables [28]. The following proposition is a simple corollary of a general concentration result on U-statistics [11].

Proposition 4 *Let $y_1, \dots, y_n$ be n vectors in $\mathbb{R}^p$ such that $\|y_i\|_2 \leqslant b$ for all $i = 1, \dots, n$ and $Y = [y_1^\top, \dots, y_n^\top]^\top \in \mathbb{R}^{n \times p}$. Let $\varepsilon \in \mathbb{R}^n$ be a vector of zero-mean independent random variables almost surely bounded by 1 and with variances σ_i^2, $i = 1, \dots, n$. Let $S = \mathrm{Diag}(\sigma_i)^\top Y Y^\top \mathrm{Diag}(\sigma_i)$. Then, for all $u \geqslant 0$:*

$$\begin{aligned} \mathbb{P}\big[\, |\varepsilon^\top Y Y^\top \varepsilon - \operatorname{tr} S| \geqslant 32 \operatorname{tr}(S^2)^{1/2} u^{1/2} + 18 \lambda_{\max}(S) u \\ + 126 b (\operatorname{tr} S)^{1/2} u^{3/2} + 39 b^2 u^2 \big] \leqslant 8 e^{-u}. \end{aligned} \tag{41}$$

Proof We apply Theorem 3.4 from [11], with $T_i = \varepsilon_i$, $g_{i,j}(t_i, t_j) = y_i^\top y_j t_i t_j$ if $|t_i|, |t_j| \leqslant 1$ and zero otherwise. We then have (following notations from [11]):

$$\begin{aligned} A &= \max_{i,j} |y_i^\top y_j| \leqslant b^2, \\ B^2 &= \max_{i \in \{1,\dots,n\}} \sum_{j<i} (y_i^\top y_j)^2 \sigma_j^2 \leqslant \max_{i \in \{1,\dots,n\}} \sum_{j<i} y_i^\top y_i b^2 \sigma_j^2 \leqslant b^2 \operatorname{tr}(S), \\ C^2 &= \sum_{j<i} (y_i^\top y_j)^2 \sigma_j^2 \sigma_i^2 \leqslant \frac{1}{2} \operatorname{tr}(S^2), \\ D &\leqslant \frac{1}{2} \lambda_{\max}(S). \end{aligned}$$

Thus (using $\varepsilon = 4$ in [11]):

$$\mathbb{P}\left(\left| \sum_{j \neq i} y_i^\top y_j \varepsilon_i \varepsilon_j \right| \geqslant 44.8 C u^{1/2} + 35.36 D u + 124.56 B u^{3/2} + A 38.26 u^2 \right) \leqslant 5.542 e^{-u}.$$

Moreover, we have from Bernstein's inequality [2]:

$$\mathbb{P}\left(\left| \sum_{i=1}^n y_i^\top y_i (\varepsilon_i^2 - \sigma_i^2) \right| \geqslant u^{1/2} \sqrt{2 b^2 \operatorname{tr} S} + \frac{b^2 u}{3} \right) \leqslant 2 e^{-u},$$

leading to the desired result, noting that for $u \leqslant \log(8)$, the bound is trivial. ■

We can apply to our setting to get, with $y_i = \frac{1}{n}(P + \lambda I)^{-1/2} x_i$ (with $\|x_i\|_2 \leqslant R$), leading to $b = \frac{1}{2} R n^{-1} \lambda^{-1/2}$ and $S = \frac{1}{n} \operatorname{Diag}(\sigma) X (P + \lambda I)^{-1} X^\top \operatorname{Diag}(\sigma)$.

Misspecified models. If no assumptions are made, we simply have: $\lambda_{\max}(S) \leqslant (\operatorname{tr} S^2)^{1/2} \leqslant \operatorname{tr}(S) \leqslant R^2 / \lambda n$ and we get after bringing terms together:

$$\mathbb{P}\left[q^\top (P + \lambda I)^{-1} q \geqslant \frac{41 R^2 u}{\lambda n} + \frac{R^2}{\lambda} \left(8 \frac{u^2}{n^2} + 63 \frac{u^{3/2}}{n^{3/2}} \right) \right] \leqslant 8 e^{-u}. \qquad (42)$$

Well-specified models In this case, $P = Q$ and $\lambda_{\max}(S) \leqslant 1/n$, $\operatorname{tr} S = d_1/n$, $\operatorname{tr} S^2 = d_2/n^2$.

$$\mathbb{P}\left[\left| q^\top (P + \lambda I)^{-1} q - \frac{d_1}{n} \right| \geqslant \frac{32 d_2^{1/2} u^{1/2}}{n} + \frac{18 u}{n} + \frac{53 R d_1^{1/2} u^{3/2}}{n^{3/2} \lambda^{1/2}} + 9 \frac{R^2 u^2}{\lambda n^2} \right] \leqslant 8 e^{-u}. \quad (43)$$

Acknowledgements

I would like to thank Sylvain Arlot, Jean-Yves Audibert and Guillaume Obozinski for fruitful discussions related to this work. This work was supported by a French grant from the Agence Nationale de la Recherche (MGA Project ANR-07-BLAN-0311).

References

[1] A. W. Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[2] P. Massart. *Concentration Inequalities and Model Selection: Ecole d'été de Probabilités de Saint-Flour 23*. Springer, 2003.

[3] S.A. Van De Geer. High-dimensional generalized linear models and the Lasso. *Annals of Statistics*, 36(2):614, 2008.

[4] C. Gu. Adaptive spline smoothing in non-gaussion regression models. *Journal of the American Statistical Association*, pages 801–807, 1990.

[5] F. Bunea. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electronic Journal of Statistics*, 2:1153–1194, 2008.

[6] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, 1999.

[7] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2003.

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

[9] R. Christensen. *Log-linear models and logistic regression*. Springer, 1997.

[10] D.W. Hosmer and S. Lemeshow. *Applied logistic regression*. Wiley-Interscience, 2004.

[11] C. Houdré and P. Reynaud-Bouret. Exponential inequalities, with constants, for U-statistics of order two. In *Stochastic inequalities and applications, Progress in Probability, 56*, pages 55–69. Birkhäuser, 2003.

[12] P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

[13] M. Yuan and Y. Lin. On the non-negative garrotte estimator. *Journal of The Royal Statistical Society Series B*, 69(2):143–161, 2007.

[14] H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, December 2006.

[15] M. J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 2009. To appear.

[16] P. J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 2009. To appear.

[17] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizbal. *Numerical Optimization Theoretical and Practical Aspects*. Springer, 2003.

[18] J. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, pages 263–274, 2008.

[19] P. McCullagh and J.A. Nelder. *Generalized linear models*. Chapman & Hall/CRC, 1989.

[20] B. Efron. The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–633, 2004.

[21] P.L. Bartlett, M.I. Jordan, and J.D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[22] G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.

[23] G. S. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33:82–95, 1971.

[24] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996.

[25] C. Gu. *Smoothing spline ANOVA models*. Springer, 2002.

[26] K. Sridharan, N. Srebro, and S. Shalev-Shwartz. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems*, 2008.

[27] I. Steinwart, D. Hush, and C. Scovel. A new concentration result for regularized risk minimizers. *High Dimensional Probability: Proceedings of the Fourth International Conference*, 51:260–275, 2006.

[28] S. Arlot and F. Bach. Data-driven calibration of linear estimators with minimal penalties. In *Advances in Neural Information Processing Systems*, 2009.

[29] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman & Hall, 1990.

[30] Z. Harchaoui, F. R. Bach, and E. Moulines. Testing for homogeneity with kernel fisher discriminant analysis. Technical Report 00270806, HAL, 2008.

[31] R. Shibata. Statistical aspects of model selection. In *From Data to Model*, pages 215–240. Springer, 1989.

[32] H. Bozdogan. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

[33] P. Liang, F. Bach, G. Bouchard, and M. I. Jordan. An asymptotic analysis of smooth regularizers. In *Advances in Neural Information Processing Systems*, 2009.

[34] P. Craven and G. Wahba. Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31(4):377–403, 1978/79.

[35] K.-C. Li. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Annals of Statistics*, 15(3):958–975, 1987.

[36] F. Bach. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

[37] C. L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.

[38] F. O'Sullivan, B.S. Yandell, and W.J. Raynor Jr. Automatic smoothing of regression functions in generalized linear models. *Journal of the American Statistical Association*, pages 96–103, 1986.

[39] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of The Royal Statistical Society Series B*, 58(1):267–288, 1996.

[40] T. Zhang. Some sharp performance bounds for least squares regression with ℓ_1 regularization. *Annals of Statistics*, 2009. To appear.

[41] A. Juditsky and A. S. Nemirovski. On Verifiable Sufficient Conditions for Sparse Signal Recovery via ℓ_1 Minimization. Technical Report 0809.2650, arXiv, 2008.

[42] P. Chaudhuri and P.A. Mykland. Nonlinear experiments: Optimal design and inference based on likelihood. *Journal of the American Statistical Association*, pages 538–546, 1993.

[43] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B*, 68(1):49–67, 2006.

[44] F. Bach. Bolasso: model consistent Lasso estimation through the bootstrap. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

[45] N. Meinshausen and P. Bühlmann. Stability selection. Technical report, arXiv: 0809.2932, 2008.

[46] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of statistics*, 34(3):1436, 2006.

[47] O. Banerjee, L. El Ghaoui, and A. dAspremont. Model selection through sparse maximum likelihood estimation. *Journal of Machine Learning Research*, 9:485–516, 2008.

[48] J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization*. Number 3 in CMS Books in Mathematics. Springer, 2000.