Finite-sample analysis of M-estimators using self-concordance

Dmitrii Ostrovskii, Francis Bach

Finite-sample analysis of M-estimators using self-concordance*

Dmitrii M. Ostrovskii† Francis Bach‡

Abstract

The classical asymptotic theory for parametric M-estimators guarantees that, in the limit of infinite sample size, the excess risk has a chi-square type distribution, even in the misspecified case. We demonstrate how *self-concordance* of the loss allows to characterize the *critical sample size* sufficient to guarantee a chi-square type in-probability bound for the excess risk. Specifically, we consider two classes of losses: (i) self-concordant losses in the classical sense of Nesterov and Nemirovski, i.e., whose third derivative is uniformly bounded with the $3/2$ power of the second derivative; (ii) *pseudo* self-concordant losses, for which the power is removed. These classes contain losses corresponding to several generalized linear models, including the logistic loss and pseudo-Huber losses.

Our basic result under minimal assumptions bounds the critical sample size by $O(d \cdot d_{\text{eff}})$, where d the parameter dimension and d_{eff} the effective dimension that accounts for model misspecification. In contrast to the existing results, we only impose *local* assumptions that concern the population risk minimizer θ_*. Namely, we assume that the calibrated predictors, i.e., predictors scaled by the square root of the second derivative of the loss, is subgaussian at θ_*. Besides, for type-ii losses we require boundedness of certain measure of curvature of the population risk at θ_*.

Our improved result bounds the critical sample size from above as $O(\max\{d_{\text{eff}}, d\log d\})$ under slightly stronger assumptions. Namely, the local assumptions must hold in the neighborhood of θ_* given by the Dikin ellipsoid of the population risk. Interestingly, we find that, for logistic regression with Gaussian design, there is no actual restriction of conditions: the subgaussian parameter and curvature measure remain near-constant over the Dikin ellipsoid. Finally, we extend some of these results to ℓ_1-penalized estimators in high dimensions.

1 Introduction and problem formulation

Recall the standard statistical learning setup: given a set $\Theta \subseteq \mathbb{R}^d$ that parametrizes the space of possible hypotheses, and observing a random $Z \in \mathcal{Z}$ with unknown distribution $\mathcal{P}$, one would like to minimize the *population risk* $L(\theta) := \mathbb{E}[\ell_Z(\theta)]$. For each possible observation z of Z, the *loss* $\ell_z : \Theta \to \mathbb{R}$ specifies the cost of choosing θ under the outcome $\{Z = z\}$, and $\mathbb{E}[\cdot]$ is the expectation with respect to the distribution $\mathcal{P}$. This distribution is assumed unknown, so the population risk cannot be computed and minimized directly. Instead, one is granted access to the sample $(Z_1, ..., Z_n)$ of independent copies of Z, and uses it to construct an estimate $\widehat{\theta}$ of the population risk minimizer,

$$\theta_* \in \operatorname*{Argmin}_{\theta\in\Theta} L(\theta),$$

assuming that such a minimizer exists. As such, we can consider the empirical distribution $\mathcal{P}_n$ – uniform probability measure supported on the sample – and the empirical risk $L_n(\theta)$, defined as the observable counterpart of $L(\theta)$, namely,

$$L_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell_{Z_i}(\theta).$$

*To appear in *Electronic Journal of Statistics* (as of November 2020).

†University of Southern California, Viterbi School of Engineering. 3650 McKlintock Avenue, CA 90089, Los Angeles, USA. Email: dmitrii.ostrovskii@inria.fr.

‡INRIA Paris and École Normale Supérieure. 2 rue Simone Iff, 75012 Paris, France. Email: francis.bach@inria.fr.

Ideally, we would like to have an estimator with small *excess risk* $L(\widehat{\theta}) - L(\theta_*)$, in probability or in expectation over the sample. Since for each fixed value θ of the parameter, $L_n(\theta)$ is an unbiased estimate of $L(\theta)$ which converges to $L(\theta)$ almost surely by the law of large numbers, a natural candidate estimator of θ_* is the *empirical risk minimizer* (ERM), defined as

$$\widehat{\theta}_n \in \operatorname*{Argmin}_{\theta\in\Theta} L_n(\theta).$$

In this paper, we are concerned with establishing high-probability finite-sample bounds on the excess risk $L(\widehat{\theta}_n) - L(\theta_*)$ of this estimator. The classical Fisher theorem ([LC06]) implies the rescaled excess risk has a chi-square type limiting behavior, under weak conditions, when $n \to \infty$. When stated informally, our goal in this paper is to characterize the *critical sample size* sufficient to enter the this "asymptotic regime", i.e., to guarantee a chi-square type high-probability bound for the excess risk in finite sample. Elaborating on this goal in more detail and stating our results would be impossible without first giving a brief overview of the classical asymptotic theory. We give such overview in the next section.

1.1 Classical asymptotic theory

Our main focus in this paper is the setting where $L_n(\theta)$ is a negative log-likelihood, that is $\ell_z(\theta) = -\log p_\theta(z)$ where $p_\theta(\cdot)$ is some probability density supported on $\mathcal{Z}$. In this case, $\widehat{\theta}_n$ maximizes the likelihood of observing the i.i.d. sample $(Z_1, ..., Z_n)$ from $\mathcal{P}_\theta$ ranging over a *parametric family* $\mathscr{P} = \{\mathcal{P}_\theta,\ \theta \in \Theta\}$. In reality, $\mathscr{P}$ may or may not contain the actual data-generating distribution $\mathcal{P}$. When $\mathcal{P} \in \mathscr{P}$, we say that the parametric model corresponding to $\mathscr{P}$ is *well-specified*; in this case, ERM becomes the maximum-likelihood estimator (MLE). Otherwise, the model is called *misspecified*, and ERM can be regarded as MLE under model misspecification, or *quasi* maximum likelihood estimator [Whi82]. In this case, $\mathcal{P}_{\theta_*}$ corresponds to the "projection" of $\mathcal{P}$ onto the family $\mathscr{P}$ in the sense of the Kullback-Leibler divergence, and the quasi MLE approximates $\mathcal{P}_{\theta_*}$ by replacing $\mathcal{P}$ with the empirical distribution $\mathcal{P}_n$.

Our goal in this section is to give a brief overview of the main results of the asymptotic theory of M-estimation. Most of them, see monographs [LC06, IH13, vdV98, Bor98], rely on the *local regularity* assumptions about the loss, allowing for second-order Taylor expansion of $L(\theta)$ around θ_*. In particular, it is assumed that $L(\theta)$ is sufficiently smooth at θ_*, which is an interior point of Θ, so that the first-order optimality condition for θ_* reduces to $\nabla L(\theta_*) = 0$. Moreover, the Hessian

$$\mathbf{H} := \nabla^2 L(\theta_*)$$

is assumed to be non-degenerate, i.e., $\mathbf{H} \succ 0$. Finally, the empirical risk is assumed to be three times continuously differentiable at θ_*, see, e.g., [LC06]. When combined together, these assumptions allow to derive, as a starting point, the *local asymptotic normality* of quasi MLE: when $n \to \infty$ with fixed d,

$$\sqrt{n}\mathbf{H}^{1/2}(\widehat{\theta}_n - \theta_*) \rightsquigarrow \mathcal{N}(0, \mathbf{H}^{-1/2}\mathbf{G}\mathbf{H}^{-1/2}), \tag{1}$$

where $\rightsquigarrow$ denotes convergence in law, and $\mathbf{G}$ is the covariance matrix of the loss gradient at θ_* (also called Fisher's information matrix):

$$\mathbf{G} := \mathbb{E}[\nabla\ell_Z(\theta_*)\nabla\ell_Z(\theta_*)^\top].$$

Matrices $\mathbf{G}$ and $\mathbf{H}$ remain fixed as n grows. Hence, under mild regularity assumptions,[1] one also has that the variance of $\widehat{\theta}_n$ decreases as $O(1/n)$. Moreover, in the well-specified case $\mathbf{G} = \mathbf{H}$, see, e.g., [Bar53], which leads to *Fisher's theorem*:

$$\sqrt{n}\mathbf{H}^{1/2}(\widehat{\theta}_n - \theta_*) \rightsquigarrow \mathcal{N}(0, \mathbf{I}_d),$$

[1] It suffices for $\rho_n := \sqrt{n}\|\widehat{\theta}_n\|_2$ to be uniformly integrable, i.e., $\lim_{\varepsilon\to 0}\sup_n \mathbb{E}[\rho_n \mathbb{1}_{\rho_n\geq\varepsilon}] = 0$. This is a very weak condition; see [Kle13, Sec. 6.2] for stronger (but easier to verify) conditions.

where $\mathbf{I}_d$ is the identity matrix of size d. Thus, denoting $\|\cdot\|_{\mathbf{J}}$ the norm linked to positive semidefinite matrix $\mathbf{J}$ by $\|x\|_{\mathbf{J}} = \|\mathbf{J}^{1/2}x\|_2$, we have $n\|\widehat{\theta}_n - \theta_*\|^2_{\mathbf{H}} \rightsquigarrow \chi^2_d$, where χ^2_d is the chi-square law with d degrees of freedom. The second-order Taylor expansion of the average risk around θ_* then allows to derive the same asymptotic law for the scaled excess risk $2n[L(\widehat{\theta}_n) - L(\theta_*)]$ – this result is known as *Wilks' theorem*. In turn, this implies (under mild regularity conditions) that

$$\mathbb{E}_n[L(\widehat{\theta}_n)] - L(\theta_*) = \frac{d}{2n} + o(n^{-1}), \text{ as } n \to \infty, \tag{2}$$

where $\mathbb{E}_n$ is the expectation over the product distribution $\mathcal{P}^{\otimes n}$ of $(Z_1, ..., Z_n)$. More precisely, by the standard chi-square deviation bounds (see e.g., [LM00, Lemma 1]) one has that, with probability $\geq 1 - \delta$,

$$L(\widehat{\theta}_n) - L(\theta_*) = \frac{(\sqrt{d} + \sqrt{2\log(1/\delta)})^2}{2n} + o(n^{-1}). \tag{3}$$

Finally, these $O(d/n)$ asymptotic bounds can be extended to the general situation of misspecified models by introducing the *effective dimension*:

$$d_{\text{eff}} := \mathbb{E}[\|\nabla \ell_Z(\theta_*)\|^2_{\mathbf{H}^{-1}}] = \text{Tr}(\mathbf{H}^{-1/2}\mathbf{G}\mathbf{H}^{-1/2}).$$

Note that in a well-specified model, $d_{\text{eff}} = d$ since $\mathbf{G} = \mathbf{H}$; moreover, in the ill-specified case one can still have $d_{\text{eff}} = O(d)$ "in favorable circumstances" – we will consider one such situation, that of misspecified linear regression, later on.[2] The expected excess risk bound (2) then changes to

$$\mathbb{E}_n[L(\widehat{\theta}_n)] - L(\theta_*) = \frac{d_{\text{eff}}}{2n} + o(n^{-1}), \tag{4}$$

and the corresponding in-probability bound (see again [LM00, Lemma 1]) is

$$L(\widehat{\theta}_n) - L(\theta_*) = \frac{d_{\text{eff}}(1 + \sqrt{2\log(1/\delta)})^2}{2n} + o(n^{-1}). \tag{$\star$}$$

In fact, the main term in the right-hand side of (4) is the minimum possible asymptotic variance of any unbiased estimator; this result is known as the Cramér-Rao bound.

For what goes next, it is important to note that the asymptotic approach can be summarized as follows:

- First, the estimate is *localized*: $\|\widehat{\theta}_n - \theta_*\|^2_{\mathbf{H}}$ is upper-bounded with the squared "natural" norm of the score, $\|\nabla L_n(\theta_*)\|^2_{\mathbf{H}^{-1}}$, which can be controlled by the central limit theorem.
- Then, using the second-order Taylor expansion of $L(\theta)$ around θ_*, similar behavior is obtained for the excess risk $L(\widehat{\theta}_n) - L(\theta_*)$.

Paying tribute to the clarity and historical significance of the classical asymptotic theory, one should keep in mind that its operating regime $n \to \infty$ with fixed parameter dimension usually cannot be applied in the modern context. The recent works [DM16, BKM+18] extend the classical results to the asymptotic high-dimensional regime $d \to \infty$ with $d = O(n)$, analyzing M-estimator as the fixed point of the approximate message passing algorithm. However, existing analysis of approximate message passing in finite samples is scarce: the only work we are aware of is [RV18], which only considers fixed-design linear regression. Postponing a more detailed review of related work to Section 7, let us briefly overview the main approaches in finite-sample analysis.

[2] We can also have $d_{\text{eff}} < d$ if we get "extremely lucky". For example, consider the Gaussian shift model $y \sim \mathcal{N}(\theta, 1)$, and let in reality $y \sim \mathcal{N}(0, \sigma)$. Then $d_{\text{eff}} = \sigma^2$ while $d = 1$.

1.2 Finite-sample regime and empirical processes

This work has been motivated by the following question:

For what finite n the excess risk admits a chi-square type bound akin to ($\star$)?

One rather general approach towards answering this question, i.e., addressing the fully finite-sample regime, has been outlined in [Spo12], and can be described as follows. First, the parameter space Θ is divided into the local subset, given as the intersection of Θ and the (unit-radius) *Dikin ellipsoid* of θ_*,

$$\Theta_1(\theta_*) := \{\theta \in \mathbb{R}^d : \|\theta - \theta_*\|_{\mathbf{H}} \leq 1\}, \tag{5}$$

and the complement subset $\Theta \setminus \Theta_1(\theta_*)$. Then, the second step of the asymptotic approach is replaced with so-called *quadratic bracketing:* the excess risk is "sandwiched" on $\Theta_1(\theta_*)$ between two quadratic forms which correspond to the inflation and deflation of $\|\theta - \theta_*\|_{\mathbf{H}}^2$. On the other hand, the first step (localization of the estimate) is done via the control of the event $\{\widehat{\theta}_n \notin \Theta_1(\theta_*)\}$, by bounding the uniform deviations of the empirical risk $L_n(\theta) - L_n(\theta_*)$ via advanced tools from empirical process theory such as generic chaining [Tal06]. This approach is quite powerful, allowing to derive the counterparts of asymptotic results in the non-asymptotic regime $n \geq c_\delta \cdot d_{\text{eff}}$, where the constant c_δ only depends on the desired confidence level $1 - \delta$. However, it requires rather strong *global* assumptions on the pointwise deviations of the empirical risk process, which are necessary to control its uniform deviations, see [Spo12, Sections 2.2 and 4]. Close in spirit to [Spo12] are the techniques developed in [CCK17] to study Gaussian approximation of the maxima of the sums of i.i.d. random variables. The main highlight of [CCK17] is the ability to handle the regime of exponentially large dimensionality, with respect to the sample size, due to the special structure of the statistics under study. However, much like in [Spo12], the techniques of [CCK17] rely on the advanced machinery of empirical processes.

Meanwhile, in the special case of random-design least-squares, finite-sample analysis is way simpler, and heavy-weight machinery of empirical processes is not needed. In this case, the problem is reduced to the control of a *single* random matrix, the sample covariance matrix of the design vector, which encapsulates the second-order information about the risk. Our primal goal in this work is to extend these ideas to a wider class of models with non-quadratic losses of certain types, including the losses arising in conditional generalized linear models and robust regression. For these classes of losses, one may carefully exploit their regularity properties, which allows to avoid using the empirical processes machinery – and the associated *global* conditions – when localizing the empirical risk minimizer. Deferring further discussion of our contributions to Sec. 1.4 and related work to Sec. 7, let us overview the case of least-squares.

1.3 Simple case: least-squares

An original approach introduced in [HKZ12a] allows to obtain finite-sample excess risk bounds in the setting of unconstrained least-squares linear regression with random design. Here, $\Theta = \mathbb{R}^d$, and the observations take the form $Z = (X, Y)$ where $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$. The goal is to predict *response* Y as a linear combination of *design* X with parameter $\theta \in \mathbb{R}^d$, and one takes $\ell_Z(\theta)$ to be $\ell_Z(\theta) = \frac{1}{2\sigma^2}(Y - X^\top\theta)^2$. ERM then reduces to the ordinary least-squares estimator. Least-squares correspond to the implicit assumption that the residual $\varepsilon = Y - X^\top\theta_*$ has Gaussian distribution $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma > 0$, and is independent of X, which allows to factor out the distribution of X from the model. Note that the rate $O(d/n)$ translates to the well-known minimax rate $O(d\sigma^2/n)$ for the mean square error $\mathbb{E}[(Y - X^\top\theta)^2] - \sigma^2$. Moreover, sometimes the Gaussian assumption on ε can be relaxed, and the misspecified situation becomes essentially as favorable as the well-specified one, at least from the asymptotic point of view. Indeed, normalizing the noise to have unit variance, and using that

$$\nabla\ell_Z(\theta_*) = \varepsilon X \quad \text{and} \quad \mathbf{H} = \mathbb{E}[XX^\top],$$

we get $d_{\text{eff}} = \mathbb{E}[\varepsilon^2 \|\mathbf{H}^{-1/2} X\|^2]$. Hence, $d_{\text{eff}} = d$ for any distribution of ε with $\mathbb{E}[\varepsilon^2] = 1$, provided that ε and X are independent. Moreover, assuming that Y and all one-dimensional marginals of X have finite fourth moment, i.e.,

$$\sqrt{\mathbb{E}[Y^4 | X = x]} \le \kappa_\varepsilon \mathbb{E}[Y^2 | X = x], \quad \forall x \in \mathbb{R}^d,$$
$$\sqrt{\mathbb{E}[\langle u, X\rangle^4]} \le \kappa_X \mathbb{E}[\langle u, X\rangle^2], \quad \forall u \in \mathbb{R}^d,$$

we can bound d_{eff} as $d_{\text{eff}} \le \kappa_X \cdot \kappa_\varepsilon \cdot d$. In other words, d_{eff} and d are comparable.

Now, the approach of [HKZ12a] exploits the fact that $L(\theta)$ is a quadratic form,

$$L(\theta) - L(\theta_*) = \tfrac{1}{2}\|\theta - \theta_*\|^2_{\mathbf{H}}, \tag{6}$$

and the empirical risk is a quadratic form corresponding to $\mathbf{H}_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$:

$$L_n(\theta) - L_n(\theta_*) = \tfrac{1}{2}\|\theta - \theta_*\|^2_{\mathbf{H}_n} + \langle \nabla L_n(\theta_*), \theta - \theta_*\rangle.$$

As such, the *global* curvature information about $L(\theta)$ is encapsulated in a single matrix $\mathbf{H}$, and we have at our disposal an unbiased estimate $\mathbf{H}_n$ of this matrix. This observation allows to dramatically simplify the analysis: it suffices to control the deviations of $\mathbf{H}_n$ from its expectation, which can be done using the standard tools of random matrix theory. In particular, in [Ver12], see also Theorem A.2 in Appendix, it is shown that whenever X is K-subgaussian in all directions, and

$$n \gtrsim K^4(d + \log(1/\delta)), \tag{7}$$

where symbol $\gtrsim$ hides a constant factor, with probability at least $1 - \delta$ it holds

$$\tfrac{1}{2}\|\Delta\|^2_{\mathbf{H}} \le \|\Delta\|^2_{\mathbf{H}_n} \le 2\|\Delta\|^2_{\mathbf{H}}, \quad \forall \Delta \in \mathbb{R}^d. \tag{8}$$

In other words, the sample second-moment matrix $\mathbf{H}_n$ approximates $\mathbf{H}$, up to a constant factor, in the sense of the corresponding Mahalanobis distances (in particular, $\mathbf{H}_n$ is non-degenerate whenever $\mathbf{H}$ is). This result can then be exploited as follows: since $\nabla L_n(\widehat{\theta}_n) = 0$, and $\mathbf{H}_n \succ 0$,

$$\|\widehat{\theta}_n - \theta_*\|^2_{\mathbf{H}_n} = \|\nabla L_n(\theta_*)\|^2_{\mathbf{H}_n^{-1}}. \tag{9}$$

Using (8), this gives $\tfrac{1}{2}\|\widehat{\theta}_n - \theta_*\|^2_{\mathbf{H}} \le 2\|\nabla L_n(\theta_*)\|^2_{\mathbf{H}^{-1}}$, which, via (6), results in

$$L(\widehat{\theta}_n) - L(\theta_*) \le 2\|\nabla L_n(\theta_*)\|^2_{\mathbf{H}^{-1}}.$$

Finally, a non-asymptotic version of ($\star$) is obtained by controlling the squared norm $\|\nabla L_n(\theta_*)\|^2_{\mathbf{H}^{-1}}$ under light-tailed (say, subgaussian or subexponential) assumptions on $\nabla \ell_Z(\theta_*) = \varepsilon X$, through standard concentration inequalities. These light-tailed assumptions can further be relaxed to fourth-moment assumption, using the generalized median-of-means estimator (see [HS16]),On the other hand, it is much more challenging to get rid of the light-tailed assumption on X, as obtaining covariance estimators with guarantees of the type (8) under weak moment assumptions is by itself a non-trivial problem. Recently, this problem has been addressed in [OR19], whose authors then proposed an estimator for ridge and ridgeless regression with near-optimal high-probability guarantees under heavy-tailed assumptions on X (see [OR19, Theorem 6.1]).[3]

The remarkable feature of the outlined analysis is that, as soon as the curvature of $L(\theta)$, as given by $\mathbf{H}$, is reliably estimated, the localization step is *"automatic"* due to (9). The only requirement is for n

[3] Another possibility is to use a rejection sampling argument similar to the one employed in the proof of our Theorem 3.2. This, however, prohibits us from taking small values of the confidence parameter δ, namely, those decreasing polynomially fast with $\min(d, n)$, cf. (31).

to reach the lower bound (7), so that one could relate the norms $\|\cdot\|_{\mathbf{H}_n}$ and $\|\cdot\|_{\mathbf{H}}$. The crucial fact here is that for the quadratic loss, the curvature information is *global*, i.e., is encoded in a single matrix. However, for more general losses this is not the case, and there seems to be no direct way of extending the above argument. As discussed before, the known solution to the problem [Spo12] involved localization of the estimate, through the control of the *global* uniform deviations of $L_n(\theta)$, to a neighborhood of θ_* where a local quadratic approximation can be used; this approach requires global assumptions on the pointwise deviations of $L_n(\theta)$. Yet, we will show that in some other models beyond linear regression with quadratic loss, the *local* analysis suffices to provide localization of the estimate, and the complicated and opaque localization step using generic chaining, as in [Spo12], can be circumvented.

1.4 Contributions and outline

Our analysis applies to *linear prediction models*: observing a pair $Z = (X, Y)$ with $X \in \mathcal{X} \subseteq \mathbb{R}^d$ and $Y \in \mathcal{Y} \subseteq \mathbb{R}$, one predicts Y through linear combination $\eta = X^\top\theta$ with $\theta \in \Theta \subseteq \mathbb{R}^d$. Accordingly, we consider losses given by

$$\ell_Z(\theta) = \ell(Y, X^\top\theta)$$

for some function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ assumed to be sufficiently smooth in its second argument. This subsumes regression ($\mathcal{Y} = \mathbb{R}$) and classification ($\mathcal{Y} = \{0, 1\}$). Moreover, we assume the ability to bound the third derivative of $\ell(y, \eta)$ with respect to η via the second derivative in two alternative ways, as will be detailed in Section 2. Such *self-concordance* assumptions originate from [NN94], where they were used in the context of interior-point methods; later on, they were modified and used in the statistical analysis of logistic regression [Bac10, BM13]. We consider both variants of self-concordance in our analysis, and show that the canonical self-concordance assumption, introduced in [NN94], leads to somewhat better bounds on the critical sample size than its modification suggested in [Bac10] (see Sections 3–4). In addition to self-concordance of the loss, we make some assumptions on the *local* behavior of the gradient and Hessian of the empirical risk at the population risk minimizer θ_*, or its neighbordhood given by the unit Dikin ellipsoid (5) of the population risk at θ_*. To prove our main results (cf. Theorems 4.1–4.2), we carefully combine these assumptions through a non-standard covering argument, which allows us to control the uniform deviations of $\nabla^2 L_n(\theta)$ from $\nabla^2 L(\theta)$ over the Dikin ellipsoid, and implies localization of the estimator. We mention once again that *global* assumptions in the vein of [Spo12] about the deviations of the empirical risk, its gradient and Hessian can be avoided by using self-concordance.

Our framework includes random-design least-squares linear regression as a baseline. However, as we show in Section 2, it is in fact much more general. First, it encompasses some conditional *generalized linear models* with random design. Here we find that both versions of self-concordance are related to natural assumptions about the moments of Y, and discover several generalized linear models amenable to our analysis, including logistic regression. Second, we can address some common losses in *robust estimation*, which turn out to be pseudo self-concordant in the sense of [Bac10]. Moreover, we show how to slightly modify these losses to make them *canonically* self-concordant, while preserving their first- and second-order structure. According to our theory, this leads to the improved statistical performance of the M-estimator, as characterized by the sufficient sample size to reach the asymptotically optimal rate for the excess risk.

Our analysis carries out the following plan. First, the local assumptions allow to make sure that starting from the certain sample size, the sample Hessian

$$\mathbf{H}_n = \nabla^2 L_n(\theta_*)$$

approximates the true Hessian $\mathbf{H} = \nabla^2 L(\theta_*)$ up to a constant factor, completely analogous to the case of least squares. After that, self-concordance comes at play. First, using simple analytic arguments, we prove that with high probability, $\nabla^2 L_n(\theta)$ remains nearly constant in a Dikin ellipsoid of a smaller radius

of order $O(1/\sqrt{d})$, leading to a larger critical sample size than in the case of least-squares. We then use these initial results to prove that under slightly stronger – but still local – assumptions, $\nabla^2 L_n(\theta)$ in fact remains constant in a *constant-radius* Dikin ellipsoid, leading to the critical sample size comparable to that in least-squares (cf. Theorems 4.1–4.2). This is done via a simple but somewhat non-trivial covering argument, which might be of independent interest.

Let us now give a more detailed overview of the obtained results.

In Section 3, we show that for *pseudo* self-concordant losses [Bac10], the asymptotically optimal (up to a costant factor) bound on the excess risk is guaranteed when the sample size reaches $O(\rho \cdot d \cdot d_{\text{eff}})$ up to a logarithmic factor in $1/\delta$, where ρ is the *local curvature* parameter linking $\mathbf{H}$ and $\mathbf{\Sigma} := \mathbb{E}[XX^\top]$ by

$$\mathbf{\Sigma} \preccurlyeq \rho \mathbf{H}.$$

Moreover, for *canonically* self-concordant losses in the sense of [NN94], the dependency on ρ can be eliminated, and the critical sample size becomes $O(d \cdot d_{\text{eff}})$. We now give a simplified (and slightly vulgarized) formulation of these two results.

Theorem 1.1 (Simplified formulation of Theorems 3.1–3.2)**.** *Assume that $\ell(y, \cdot)$ is self-concordant, for any y, in the sense of Nesterov and Nemirovski [NN94], i.e.,*

$$|\ell'''_\eta(y, \eta)| \le 2\ell''_\eta(y, \eta)^{3/2}, \quad \forall \eta \in \mathbb{R}, \tag{10}$$

and that $\ell'_\eta(Y, X^\top \theta_*)X =: \nabla \ell_Z(\theta_*)$ and $\ell''_\eta(Y, X^\top \theta_*)^{1/2} X$ are subgaussian. Then

$$L(\widehat{\theta}_n) - L(\theta_*) \lesssim \|\widehat{\theta}_n - \theta_*\|^2_{\mathbf{H}} \lesssim \|\nabla L_n(\theta_*)\|^2_{\mathbf{H}^{-1}} \lesssim \frac{d_{\text{eff}} \log(e/\delta)}{n} \tag{11}$$

with probability $\ge 1 - \delta$, $\delta \in (0, 1)$, as long as

$$n \gtrsim d_{\text{eff}} \cdot d \cdot \log(ed/\delta), \tag{12}$$

where $\lesssim, \gtrsim$ hide constants. Moreover, if the loss satisfies the modified assumption

$$|\ell'''_\eta(y, \eta)| \le \ell''_\eta(y, \eta), \quad \forall \eta \in \mathbb{R} \tag{13}$$

instead of (10), X is as well subgaussian, and $\mathbf{\Sigma} \preccurlyeq \rho \mathbf{H}$, then (11) is valid once

$$n \gtrsim \rho \cdot d_{\text{eff}} \cdot d \cdot \log(ed/\delta). \tag{14}$$

While the only available generic upper bound on ρ is given by the inverse of the *global* strong convexity modulus of the loss, and can be very large or even infinite in the case of unbounded predictors, the *actual* value of ρ depends on the data distribution, and is moderate when this distribution is not chosen adversarially, as discussed in [BM13, Sections 3.1, 4.2] and in our Section 2.2. In this vein, we show in Appendix D that $\rho \lesssim 1 + \|\theta_*\|^3_{\mathbf{\Sigma}}$ in logistic regression with Gaussian design $X \sim \mathcal{N}(0, \mathbf{\Sigma})$. Motivated by this result, we propose canonically self-concordant losses for classification and robust regression in Section 2.1.

In Section 4, we obtain improved bounds for the critical sample size, scaling *near-linearly in the parameter dimension,* under slightly stronger assumption on the data distribution. Essentially, we now require that the *calibrated design* $\widetilde{X}(\theta) := [\ell''(Y, X^\top \theta)]^{1/2} X$, is subgaussian uniformly over θ in the set

$$\Theta_r(\theta_*) := \{\theta : \|\theta - \theta_*\|_{\mathbf{H}} \le r\} \tag{15}$$

– the r-radius *Dikin ellipsoid* of the population risk at θ_*. specifically, we require $r = 1$ for canonically self-concordant losses, and $r = 1/\sqrt{\rho}$ for pseudo self-concordant losses. This assumption is still local, and is not much more restrictive in some practical situations: in Appendix D we show, informally, that in the case of logistic regression with Gaussian design, the tails of $\widetilde{X}(\theta)$ over $\theta \in \Theta_{1/\sqrt{\rho}}(\theta_*)$ are not heavier than those of $\widetilde{X}(\theta_*)$ (see Proposition D.1). It allows to control the uniform deviations of the empirical Hessians from their means on $\Theta_r(\theta_*)$, leading to the reduced sample size as per the following result.

Theorem 1.2 (Simplified formulation of Theorems 4.1–4.2)**.** *In addition to the premise of Theorem 1.2, assume that the vectors* $\widetilde{X}(\theta) := [\ell''(Y, X^\top\theta)]^{1/2}X$ *are subgaussian for* $\theta \in \Theta_r(\theta_*)$*, cf.* (15)*, with* $r = 1$ *in the case of* (10) *and* $r = 1/\sqrt{\rho}$ *in the case of* (13)*. Then bounds* (11) *in Theorem 1.1 are valid once*

$$n \gtrsim \begin{cases} d_{\text{eff}} \vee d\log d & \text{under (10)}, \\ \rho \cdot d_{\text{eff}} \vee d\log d & \text{under (13)}. \end{cases} \tag{16}$$

The main technical challenge when proving this result is the fact that, while (pseudo) self-concordance of the *population* risk over $\Theta_r(\theta_*)$ with appropriate r follows from that of the loss function (by relating the directional derivatives of $L(\theta)$ to the corresponding moments of $\widetilde{X}(\theta)$), this fails to hold for the *empirical* risk. Hence, we cannot uniformly control its Hessians on $\Theta_r(\theta_*)$ by simply integrating the directional third derivatives of the empirical risk. Instead, such control is attained by observing that self-concordance of the losses suffices to control Hessians in a smaller Dikin ellipsoid with radius $O(1/d^\kappa)$ for some $\kappa \geq 1/2$, and combining this observation with a somewhat non-standard covering argument. We hypothesize that the bounds (16) are optimal up to the $\log(d)$ factor, i.e., ERM cannot provably achieve the nonasymptotic version of ($\star$) in the regime where n is sublinear in d_{eff} or d. This hypothesis is motivated by the observation that $n \gtrsim d$ is necessary to estimate the local norm $\|\cdot\|_{\mathbf{H}}$, whereas $n \gtrsim d_{\text{eff}}$ is necessary to have $\|\nabla L_n(\theta_*)\|_{\mathbf{H}} \leq c$, which, in turn, allows to localize $\widehat{\theta}_n$ near θ_*.

In Section 5, we extend some of the above results to the high-dimensional setup. Specifically, we obtain analogues of Theorem 1.1 for ℓ_1-regularized M-estimators, assuming that the optimal parameter θ_* is $\mathtt{s}$-sparse, the matrices $\mathbf{G}$ and $\mathbf{H}$ are bounded in the operator norm, and the design is uncorrelated (the last assumption can in principle be relaxed). In the case of pseudo self-concordant losses (Theorem 5.1), we replace $\max(d, d_{\text{eff}})$ with $O(\rho\mathtt{s}\log(d))$, both in the error rates and the minimal sample size requirements. Unfortunately, for canonically self-concordant losses, we do not get the expected improvement by ρ (see Theorem 5.2), and the bounds essentially remain the same as in the case of pseudo self-concordance. This, however, is not surprising, since sparsity and ℓ_1-regularization depend on the choice of the basis, and are not affine-invariant, which prevents us from fully exploiting self-concordance in the analysis by forcing to rely on the usual ℓ_1- and ℓ_2-norms instead of $\|\cdot\|_{\mathbf{H}}$. More detailed discussion of these results and their comparison with related work is deferred to Section 5.

1.5 Notation

We write $f \lesssim g$ or $f = O(g)$ to state that $f(\cdot) \leq Cg(\cdot)$ for any admissible arguments of $f(\cdot)$, $g(\cdot)$ and some constant $C > 0$; analogously for $f \gtrsim g$ or $f = \Omega(g)$. Notation $f \approx g$ means $f \lesssim g \lesssim f$. $[n]$ is the set of integers $\{1, 2, ..., n\}$. Throughout, θ_* is the unique minimizer of $L(\theta)$, Similarly, $\widehat{\theta}_n$ is the minimizer of $L_n(\theta)$, which will be (provably) unique with high probability in all cases. Random vectors are denoted with capital letters (such as Z), and matrices with bold capital letters (such as $\mathbf{A}$). $\mathbf{I}_d$ is the $d \times d$ identity matrix. $\mathbf{A}^\top$ is the transpose of $\mathbf{A}$. For two square matrices $\mathbf{A}_1, \mathbf{A}_2$ of the same size, we write $\mathbf{A}_1 \prec \mathbf{A}_2$ (resp., $\mathbf{A}_1 \preccurlyeq \mathbf{A}_2$) when $\mathbf{A}_2 - \mathbf{A}_1$ is positive (semi)definite. We denote with $\|\cdot\|_p$ the ℓ_p-norm on $\mathbb{R}^d$ and the Schatten ℓ_p-norm of a matrix; in particular, $\|\mathbf{A}\|_2$ is the Frobenius and $\|\mathbf{A}\|_\infty$ the operator norm. For $\mathbf{A} \succeq 0$, we define the seminorm $\|\theta\|_{\mathbf{A}} := \|\mathbf{A}^{1/2}\theta\|_2$.

2 Assumptions and examples

Before introducing the assumptions, we remind that the loss $\ell_Z : \Theta \to \mathbb{R}$ is modeled as $\ell_Z(\theta) = \ell(Y, X^\top\theta)$ for some function $\ell(y, \eta)$ on $\mathcal{Y} \times \mathbb{R}^{(+)}$, where $\mathcal{Y}$ is a subset of $\mathbb{R}$, and $\mathbb{R}^{(+)}$ is allowed to be either $\mathbb{R}$ or the ray $\mathbb{R}^+$ of strictly positive numbers, which allows to encompass the exponential response model (cf. Section 2.1). We refer to both $\ell_Z(\theta)$ and $\ell(y, \eta)$ as *the loss*; which of the two we mean is clear from context. The derivatives of $\ell(y, \eta)$ are with respect to η.

2.1 Self-concordance assumptions

Let us introduce the assumptions related purely to the loss, rather than to the data distribution. Our standing assumption, which we silently use later on, is that the loss $\ell_z(\cdot)$ is three times differentiable and convex on Θ for any $z \in \mathcal{Z}$.

We first present the assumption of *pseudo self-concordance*, introduced in [Bac10] for the analysis of logistic regression. The reader may refer to [STD18, TDKC15, BM13] for the uses of generalized self-concordance in the context of quasi-Newton algorithms.

Assumption SCa. *For any $y \in \mathcal{Y}$ and $\eta \in \mathbb{R}^{(+)}$, the loss satisfies*

$$|\ell'''(y, \eta)| \leq \ell''(y, \eta).$$

We also consider the *canonical* self-concordance assumption first introduced in [NN94] in the context of interior-point algorithms. The constant 2 is standard in the literature, but can be replaced with arbitrary constant by rescaling the loss.

Assumption SCb. *For any $y \in \mathcal{Y}$ and $\eta \in \mathbb{R}^{(+)}$, the loss satisfies*

$$|\ell'''(y, \eta)| \leq 2[\ell''(y, \eta)]^{3/2}.$$

We now present some examples in which either of these assumptions is satisfied.

2.1.1 Generalized linear models over canonical exponential family

In generalized linear models (GLM) with canonical link function ([MN89]), one has

$$\ell(y, \eta) = -y\eta + a(\eta) - b(y), \tag{17}$$

where the *cumulant* $a(\eta) : \mathbb{R}^{(+)} \to \mathbb{R}$ normalizes $-\ell(y, \eta)$ to be a log-likelihood:

$$a(\eta) = \log \int_{\mathcal{Y}} \exp(y\eta + b(y)) \, \mathrm{d}y.$$

With $\eta = X^\top \theta$, we have a conditional GLM for Y given $\eta = X^\top \theta$.

Note that the second and third derivatives of $\ell(y, \eta)$ with respect to η coincide with those of $a(\cdot)$, hence ℓ satisfies the basic smoothness/convexity assumption whenever $a(\cdot)$ is three times differentiable (as such, $a(\cdot)$ must be convex). In fact, the cumulant derivatives correspond to the central moments of Y:

$$a'(\eta) = \mathbb{E}_\eta[Y], \quad a''(\eta) = \mathbb{E}_\eta[(Y - \mathbb{E}_\eta[Y])^2], \quad a'''(\eta) = \mathbb{E}_\eta[(Y - \mathbb{E}_\eta[Y])^3],$$

where $\mathbb{E}_\eta[\cdot]$ is expectation with respect to the distribution with negative log-likelihood given by (17). Hence, Assumption SCb states precisely that the *skewness* of the model distribution is bounded by a constant uniformly over $\eta \in \mathbb{R}^{(+)}$. This is the case in the *exponential response* GLM where $Y \sim \mathrm{Exp}(\eta)$ and $a(\eta) = -\log(\eta)$ defined on $\mathbb{R}^{(+)} = \mathbb{R}^+$.

On the other hand, Assumption SCa is satisfied whenever the third absolute central moment of Y is uniformly bounded by the variance of Y, without the $3/2$ power. This is the case in *Poisson regression:* $Y \sim \mathrm{Poisson}(\lambda)$ with $\lambda = \exp(\eta)$; then $b(y) = -\log(y!)$ and $a(\eta) = \exp(\eta)$ so that $a'''(\eta) = a''(\eta)$. This model is appropriate for count data where the rate of arrival itself depends multiplicatively on the canonical parameter η; see, e.g., [Chr06]. Perhaps most importantly, Assumption SCa is automatically satisfied in *logistic regression* in which $\mathcal{Y} = \{0, 1\}$, and Y is modeled as a Bernoulli random variable with $\mathbb{P}_\eta\{Y = 1\} = \sigma(\eta)$ where $\sigma(\eta) = 1/(1 + e^{-\eta})$ is the sigmoid function. In this case, $a(\eta) = \log(1 + e^\eta)$, and one can verify that $a'''(\eta) = a''(\eta)(1 - 2\sigma(\eta))$, so Assumption SCa is satisfied since $|\sigma(\eta)| < 1$ for any $\eta \in \mathbb{R}$. Another way to see this is by looking at the cumulant and using that $\mathcal{Y} = \{0, 1\}$:

$$|a'''(\eta)| \leq |Y - \mathbb{E}_\eta[Y]| \cdot \mathbb{E}_\eta[(Y - \mathbb{E}_\eta[Y])^2] \leq \mathbb{E}_\eta[(Y - \mathbb{E}_\eta[Y])^2] = a''(\eta).$$

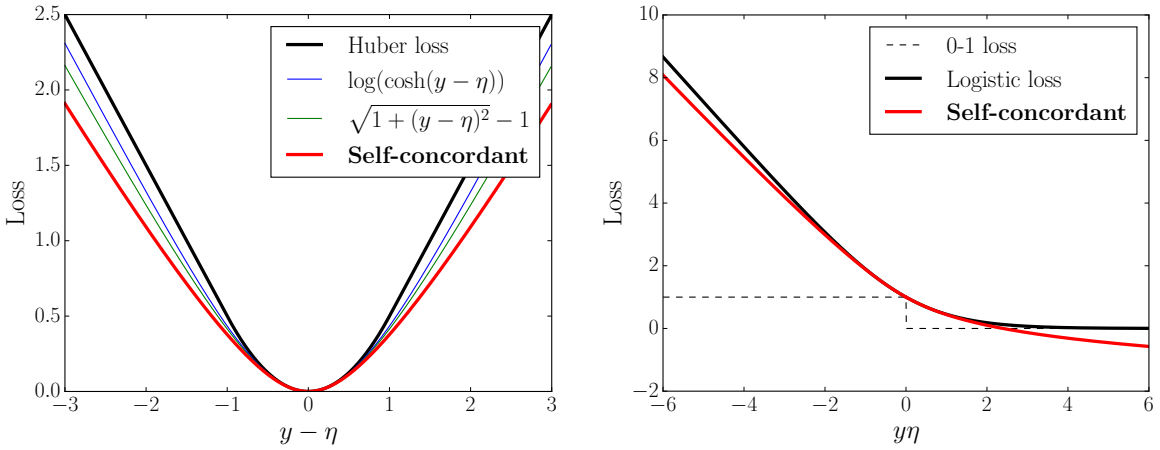

Figure 1: *Left:* self-concordant pseudo-Huber loss, cf. (21). *Right*: self-concordant analogue of the logistic loss suitable for classification, cf. (22). Although our classification loss does not upper-bound the 0-1 loss on $\mathbb{R}^+$, it can be lower-bounded as $\Omega(-\log(y\eta))$ whenever $y\eta > 0$.

2.1.2 Robust estimation

Here, $\mathcal{Y} = \mathbb{R}$, and $\ell(y, \eta) = \varphi(y - \eta)$ for some *contrast* $\varphi : \mathbb{R} \to \mathbb{R}$, a function minimized in the origin and usually even. Crucially, $\varphi(\cdot)$ must be globally Lipschitz-continuous, which guarantees robustness of the M-estimator, see [Hub11]. On the other hand, from the statistical perspective, one can motivate contrasts that are locally quadratic, i.e., such that $\varphi''(0)$ exists and is strictly positive, see, e.g., [Loh17].[4] These considerations, along with some minimax optimality results, lead to the Huber loss (see [Hub64]):

$$\varphi_\tau(t) = \begin{cases} t^2/2, & |t| \le \tau, \\ \tau t - \tau^2/2, & |t| > \tau. \end{cases} \tag{18}$$

The Huber loss is parametrized by $\tau > 0$, which allows to control the tradeoff between robustness and statistical performance. Indeed, on one hand, $|\varphi'_\tau(t)| \le \tau$ for any $t \in \mathbb{R}$, and we make estimation more robust by decreasing τ; on the other hand, the variance of the corresponding M-estimator usually decreases as τ. However, finite-sample statistical analysis of the Huber loss is complicated by the fact that $\varphi(t)$ is not C^3-smooth. This is also unfavorable from the algorithmic perspective, as it complicates the analysis of Newton-type algorithms for the computation of the M-estimator. These issues can be circumvented if one instead uses *pseudo-Huber losses*, which retain the favorable properties of the Huber loss, yet are C^3-smooth. E.g., such are contrasts of the form $\varphi_\tau(t) = \tau^2\varphi(t/\tau)$ with

$$\varphi(t) = \log\left(\cosh(t)\right), \quad \varphi(t) = \sqrt{1+t^2} - 1. \tag{19}$$

In both cases, the resulting $\phi''_\tau(\cdot)$ satisfies $\phi''_\tau(0) = 1$ for any $\tau > 0$, and $|\varphi'_\tau(t)| \le \tau$ for any $t \in \mathbb{R}$. Moreover, simple algebra shows that both functions in (19) satisfy Assumption SCa up to $c = 3$, whence $|\varphi'''_\tau(t)| \le \frac{3}{\tau}\phi''_\tau(t)$. As such, our theory is applicable to both these losses if they are properly renormalized.

2.1.3 Novel self-concordant losses

Here we construct a *canonically self-concordant* (up to a constant) pseudo-Huber loss, and similarly, a canonically self-concordant loss suitable for classification and similar to the logistic loss. This construction

[4] However, this condition is *not* necessary for the asymptotic normality of M-estimator. For example, the sample median ($\varphi(t) = |t|$) in the model $y = \theta + \varepsilon \in \mathbb{R}$ is asymptotically normal provided that the density of ε does not vanish at 0.

is motivated by the observation that our theory has a somewhat tighter guarantee on the critical sample size (after which the fast rates occur) under the canonical self-concordance assumption. (However, in practice the situation might be different as we explore in Sec. 6.) The key idea in this construction is that self-concordance is preserved under convex conjugation (see, e.g., [STD18, Prop. 6]), while at the same time one can control the range of the function through the domain of its convex conjugate (see [Roc70]). Namely, consider $\phi : (-1, 1) \to \mathbb{R}^+$:

$$\phi(u) = -\log(1 - u^2)/2, \tag{20}$$

that is, the negative log-barrier on $[-1, 1]$ normalized by $\phi''(0) = 1$. Its convex conjugate $\varphi(t)$ can be explicitly computed:

$$\varphi(t) = \frac{1}{2}\left[\sqrt{1 + 4t^2} - 1 + \log\left(\frac{\sqrt{1 + 4t^2} - 1}{2t^2}\right)\right]. \tag{21}$$

Note that $\phi(\cdot)$ is even, satisfies $\phi''(0) = 1$ and $|\phi'''(u)| \le 2\sqrt{2}[\phi''(u)]^{3/2}$, since both functions $\log(1 \pm u)$ satisfy Assumption SCb. By simple calculations detailed in Appendix C, $\varphi(t)$ defined in (21) has all the same properties. On the other hand, we have $|\varphi'(t)| < 1$ since $\phi(u)$ is a barrier on $[-1, 1]$. Thus, $\varphi(t)$ has all properties desired for a robust loss, and besides is canonically self-concordant (albeit with constant $2\sqrt{2}$ instead of 2). As illustrated in Figure 1, the quality of approximating the Huber loss for the new loss is essentially as good as for the commonly used pseudo-Huber losses (19). The new loss has a rescaled version $\varphi_\tau(t) = \tau^2\varphi(t/\tau)$, for which $\varphi_\tau''(0) = 1$, $|\varphi_\tau'(t)| \le \tau$, and $|\varphi_\tau'''(t)| \le (2/\tau)[\varphi_\tau''(t)]^{3/2}$.

Similarly, we can construct a self-concordant counterpart of the logistic loss suited for classification. In this case, we take $\phi(u) = -\log(u(1 + u))/2$, the normalized log-barrier of $[-1, 0]$, whose convex conjugate is

$$\phi^*(t) = \frac{1}{2}\left[-1 - t + \sqrt{1 + t^2} + \log\left(\frac{\sqrt{1 + t^2} - 1}{2t^2}\right)\right].$$

The derivative of $\phi^*(\cdot)$ must belong to $(-1, 0)$, and is canonically self-concordant (up to a constant) by the same reasoning as before. By rescaling and shifting it, we obtain the loss

$$\ell(y, \eta) = 2 + \frac{1}{2\log 2}\left[-1 - y\eta + \sqrt{1 + (y\eta)^2} + \log\left(\frac{\sqrt{1 + (y\eta)^2} - 1}{2(y\eta)^2}\right)\right] \tag{22}$$

which can be understood as a convex surrogate of the 0-1 loss similar to the logistic loss, see Figure 1. However, this loss is negative for $y\eta > 2.4$, and therefore does not globally upper-bound the 0-1 loss. Fortunately, its right branch can be lower-bounded with $\Omega(-\log(y\eta))$, so the resulting "leakage" is insignificant. On the other hand, this defect is unavoidable: one can show that a canonically self-concordant function on $\mathbb{R}^+$ cannot have a horizontal asymptote: this would imply $\varphi''(t) \to_{t\to+\infty} 0$, contradicting Assumption SCb reformulated as $|([\varphi''(t)]^{-1/2})'| \le 1$. Finally, let us remark that the "leakage" effect can also be quantified using the so-called calibration theory [BJM06].

2.2 Distribution assumptions

Preliminaries. We now introduce additional assumptions that ar e related to the distribution of the design scaled by the derivatives of the loss at the true optimum θ_*. All these assumptions are fully *local*, i.e., they only concern the true optimal point θ_*. We begin with the basic assumptions. First, we assume the existence of the matrices

$$\mathbf{\Sigma} := \mathbb{E}[XX^\top], \quad \mathbf{G} := \mathbb{E}[\nabla\ell_Z(\theta_*)\nabla\ell_Z(\theta_*)^\top], \quad \mathbf{H} := \mathbb{E}[\nabla^2\ell_Z(\theta_*)];$$

Generally, $\boldsymbol{\Sigma} \neq \mathbf{H}$ (unless for least-squares), and $\mathbf{G} \neq \mathbf{H}$ (unless in a well-specified model). Recall that $\mathbb{E}[\nabla \ell_Z(\theta_*)] = 0$; as such, $\mathbf{G}$ is the covariance matrix of $\nabla \ell_Z(\theta_*)$. For future reference we also note that, for any $\theta \in \Theta$, one has

$$\nabla \ell_Z(\theta) = \ell'(Y, X^\top \theta) X, \quad \nabla^2 \ell_Z(\theta) = \ell''(Y, X^\top \theta) X X^\top. \tag{23}$$

We assume that $X^\top \theta \in \mathbb{R}^{(+)}$ for any $\theta \in \Theta$ and $X \in \mathcal{X}$. This assumption is non-trivial only when $\mathbb{R}^{(+)} = \mathbb{R}^+$ which is of interest in the exponential response model. In this case, one can assume $\Theta \subseteq \mathbb{R}^d_+$ and $\mathcal{X} \subseteq \mathbb{R}^d_+$ where $\mathbb{R}^d_+$ is the positive orthant, or replace the pair $(\mathbb{R}^d_+, \mathbb{R}^d_+)$ with other pairs of mutually dual convex cones in $\mathbb{R}^d$.

Following [Ver12], we use the formalism of subgaussian, or ψ_2-norms. The ψ_2-norm $\|\xi\|_{\psi_2}$ of a random variable $\xi \in \mathbb{R}$ can be defined in a number of equivalent ways (see Appendix A), e.g., as $\|\xi\|_{\psi_2} := \{\sigma > 0 : \mathbb{E}[e^{\xi^2/\sigma^2}] \leq e\}$. This definition extends to random vectors $Z \in \mathbb{R}^d$ in a standard way:

$$\|Z\|_{\psi_2} := \sup\{\|\langle Z, \theta \rangle\|_{\psi_2} : \|\theta\|_2 \leq 1\}.$$

In other words, $\|Z\|_{\psi_2}$ is the maximal $\|\cdot\|_{\psi_2}$-norm for all one-dimensional marginals of Z. See Appendix A on more details on subgaussian random variables.

Assumption D0. *The decorrelated design is subgaussian: it holds*

$$\|\boldsymbol{\Sigma}^{-1/2} X\|_{\psi_2} \leq K_0.$$

Assumption D0 is often satisfied with a constant K_0 not depending on n or d. For example, this is the case for zero-mean Gaussian design $X \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, or design with independent Bernoulli components. Moreover, it can be shown that affine transformation of the design X that satisfies Assumption D0 also satisfies it, with at worst twice larger K_0 (see Lemma A.5 in Appendix).

Assumption D1. *The decorrelated loss gradient at θ_* is subgaussian:*

$$\|\mathbf{G}^{-1/2} \nabla \ell_Z(\theta_*)\|_{\psi_2} \leq K_1.$$

Note that Assumption D1 can be reformulated in terms of the design vector scaled by the loss derivative at θ_* since $\nabla \ell_Z(\theta_*) = \ell'(Y, X^\top \theta_*) X$. Similarly, we can consider the random vector

$$\widetilde{X} := [\ell''(Y, X^\top \theta_*)]^{1/2} X \tag{24}$$

which we call the *calibrated design*. Note that $\widetilde{X}$ is linked with $\mathbf{H}$ by $\mathbb{E}[\widetilde{X}\widetilde{X}^\top] = \mathbf{H}$, cf. (23). As stated next, we assume that the calibrated design is subgaussian. This allows to control the deviations of $\mathbf{H}_n$ using Theorem A.2 in Appendix.

Assumption D2. *The calibrated design $\widetilde{X} := [\ell''(Y, X^\top \theta_*)]^{1/2} X$ satisfies*

$$\|\mathbf{H}^{-1/2} \widetilde{X}\|_{\psi_2} \leq K_2.$$

Assumption D2 can be reformulated in terms of the loss Hessian $\nabla^2 \ell_Z(\theta_*)$ due to (23). However, this formulation does not give new ideas, and we omit it.

Remark 2.1. *The quantities K_0, K_1, K_2 are necessarily bounded with some absolute constant from below. This fact follows from the moment characterization of the ψ_2-norm (Item 2 of Lemma A.1 in Appendix), combined with the bound $(\mathbb{E}|\xi|^4)^{1/4} \geq (\mathbb{E}|\xi|^2)^{1/2}$ for any random variable $\xi \in \mathbb{R}$, and allows to simplify the formulation of the subsequent results.*

Remark 2.2. *Assumptions D1–D2 are quite restrictive, even under Assumption D0. In particular, in GLMs with canonical link function (cf. Section 2.1), the calibrated design at point θ_* is given by $\widetilde{X}(\theta) = [a''(X^\top\theta)]^{1/2}X$ where $a(\eta)$ is the cumulant function. The transform $[a''(X^\top\theta)]^{1/2}$ that scales X along a direction θ can be highly-non-linear, breaking subgaussianity for $\widetilde{X}(\theta)$. For example, Assumption D2 does not hold in Poisson regression. Another limitation of our approach is that the constants K_1, K_2 in Assumptions D1–D2 can depend on the magnitude of θ_*. In fact, for logistic regression with Gaussian design $X \sim \mathcal{N}(0, \mathbf{\Sigma})$, one has*

$$K_2 \lesssim \log(1 + \|\theta_*\|_{\mathbf{\Sigma}})\sqrt{1 + \|\theta_*\|_{\mathbf{\Sigma}}}.$$

This proof of this estimate (see Appendix D) is highly non-trivial, and relies on the Gaussianity of X. We also show that

$$K_1 \lesssim 1 + \|\theta_*\|_{\mathbf{\Sigma}}^{3/2}$$

if the logistic model for $Y|X$ is well-specified. This improves to $K_1 \lesssim 1 + \|\theta_*\|_{\mathbf{\Sigma}}^{1/2}$ if the subgaussian norm $\|\cdot\|_{\psi_2}$ is replaced with the subexponential norm $\|\cdot\|_{\psi_1}$ (see Appendix D and Section 3 for details). In other applications, one should carefully verify Assumptions D1–D2, bounding the constants K_1 and K_2. This can be a non-trivial task itself, especially when the distribution of X is unknown.

Finally, for pseudo self-concordant losses we need *compatibility* of $\mathbf{\Sigma}$ and $\mathbf{H}$.

Assumption C. *It holds $\mathbf{\Sigma} \preccurlyeq \rho\mathbf{H}$ for some $\rho < \infty$.*

Assumption C has already appeared in the statistical analysis of logistic regression in [BM13]. Note that the simplest *generic* upper bound for ρ is

$$\rho \le \left(\inf_{(y,\eta)\in\mathcal{Y}\times\mathbb{R}^{(+)}} \ell''(y,\eta)\right)^{-1}, \tag{25}$$

and unless $\ell''(y,\cdot)$ is strictly convex on $\mathbb{R}^{(+)}$ (which is usually not the case), this bound is vacuous. On the other hand, the infinum in (25) can be taken on the subset of $\mathbb{R}^{(+)}$ corresponding to possible values of $X^\top\theta_*$, but such bound can still be very conservative: for example, it only gives $\rho = O(e^{RD})$ in the case of logistic regression with $\|X\|_2 \le R$ *a.s.* and $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le D\}$. However, the *actual* value of ρ depends on the true distribution of the data, and is usually much smaller, see, e.g., dicsussion in [BM13, Sections 3.1, 4.2] for the case of logistic regression. For example, consider a "quasi well-specified" robust regression model: $\ell(Y, X^\top\theta) = \varphi(Y - X^\top\theta)$ with even contrast $\varphi(\cdot)$ and unconstrained parameter. Suppose that the true distribution of Y is given by $Y = X^\top\theta_* + \varepsilon$. with ε being independent from X, zero-mean, and symmetrically distributed. One can check that in this case, $L(\theta)$ is minimized at θ_*, and $\rho = 1/\mathbb{E}[\varphi''(\varepsilon)]$. On the other hand, the worst-case bounds on ρ can be enforced if the data distribution is chosen *adversarially*. In particular, for logistic regression [HKL14] construct an adversarial distribution that enforces $\rho = \Omega(e^{RD})$ as long as $n = O(e^{RD})$.

3 Results under minimal assumptions

In this section, we present extensions of the asymptotic deviation bound ($\star$) to the finite-sample regime *under minimal assumptions*. We then refine these results in Section 4, under a slightly strengthened version of Assumption D2, through a more subtle analysis. In the proofs, we use some probabilistic tools collected in Appendix A; in particular, we use deviation bounds for the quadratic forms (Theorem A.1) and for sample covariance matrices (Theorem A.2) of subgaussian vectors. We also use technical results on (pseudo) self-concordant functions collected in Appendix B. Some of them appear to be new, and are of independent interest. To improve readability, we defer the proofs to Appendix C.

Preliminaries. In the results which we are about to present, there is a technical difficulty arising due to the unboundedness of the vectors X and $\widetilde{X}$, cf. (24).[5] To this end, we observe that, due to Assumptions D0 and D2, these vectors admit $O(\sqrt{d})$ high-probability bound on their norms – more precisely, the events

$$\mathcal{E}_0 := \left\{ \|X\|_{\mathbf{\Sigma}^{-1}} \lesssim K_0 \sqrt{d \log(e/\delta)} \right\}, \quad \mathcal{E}_2 := \left\{ \|\widetilde{X}\|_{\mathbf{H}^{-1}} \lesssim K_2 \sqrt{d \log(e/\delta)} \right\}$$

hold with probability $\geq 1-\delta$, correspondingly, under Assumptions D0 and D2. To exploit this fact, we replace the population risk $L(\theta)$ with the *restricted risks*:

$$L_{\mathcal{E}_0}(\theta) := \mathbb{E}[\ell_Z(\theta) \mathbb{1}\{X \in \mathcal{E}_0\}]; \quad L_{\mathcal{E}_2}(\theta) := \mathbb{E}[\ell_Z(\theta) \mathbb{1}\{\widetilde{X} \in \mathcal{E}_2\}], \tag{26}$$

where we exclude from averaging the low-probability outcomes in which the norms of X and $\widetilde{X}$ are too large. Provided that δ is small enough, we can show that $\nabla L_{\mathcal{E}_0}(\theta_*) \approx \nabla L_{\mathcal{E}_2}(\theta_*) \approx 0$ and $\nabla^2 L_{\mathcal{E}_0}(\theta_*) \approx \nabla^2 L_{\mathcal{E}_2}(\theta_*) \approx \nabla^2 L(\theta_*)$, so that the second-order structure of the population risk is preserved; at the same time, we can now work with X and $\widetilde{X}$ as if they were almost surely bounded.

We now present our basic result for M-estimators with self-concordant losses.

Theorem 3.1. *Let Assumptions SCa, D0, D1, D2, and C hold. Whenever*

$$n \gtrsim \max\left\{ K_2^4 (d + \log(1/\delta)), \quad \rho K_0^2 K_1^2 d_{\text{eff}}\, d \log(ed/\delta) \right\}, \tag{27}$$

with probability at least $1-\delta$ it holds

$$\|\nabla L_n(\theta_*)\|^2_{\mathbf{H}^{-1}} \lesssim \frac{K_1^2 d_{\text{eff}} \log(e/\delta)}{n}, \tag{28}$$

$$\|\widehat{\theta}_n - \theta_*\|^2_{\mathbf{H}} \lesssim \|\nabla L_n(\theta_*)\|^2_{\mathbf{H}^{-1}}. \tag{29}$$

Moreover, one has

$$L_{\mathcal{E}_0}(\widehat{\theta}_n) - L_{\mathcal{E}_0}(\theta_*) \lesssim \frac{K_1^2 d_{\text{eff}} \log(e/\delta)}{n} \tag{30}$$

provided that

$$\delta \lesssim \min\left\{ \left(\frac{1}{\sqrt{n \log(e d_{\text{eff}})}} \right)^{1+1/\log(d_{\text{eff}})}, \left(\frac{1}{K_2^2 d \log(ed)} \right)^{1+1/\log(d)} \right\}. \tag{31}$$

The main message of Theorem 3.1 is that, under minimal assumptions, the "quadratic" behavior of the population risk, as given by (28)–(30), is guaranteed for sample sizes growing quadratically in parameter dimension – more precisely, for $n = \widetilde{\Omega}(\rho \cdot d \cdot d_{\text{eff}})$, cf. the second bound in (27), where $\widetilde{\Omega}$ hides subgaussian constants and the logarithmic factor in δ. Technically, the curvature parameter ρ appears in (27) because of the "incorrect" power of the second derivative in Assumption SCa as compared to power $3/2$ in Assumption SCb. Indeed, for canonically self-concordant losses, the factor ρK_0^2 in the bound for the critical sample size get replaced with K_2^2, and Assumptions C and D0 are not needed.

Theorem 3.2. *Let Assumptions SCb, D1, D2 hold, and assume that δ satisfies (31). Then, (28)–(30) are satisfied, with $L_{\mathcal{E}_2}(\cdot)$ instead of $L_{\mathcal{E}_0}(\cdot)$, whenever*

$$n \gtrsim \max\left\{ K_2^4 (d + \log(1/\delta)), \; K_1^2 K_2^2 d_{\text{eff}}\, d \log(ed/\delta) \right\}. \tag{32}$$

We also note that both of the above results include a technical condition (31) that does not minimal violation probability δ. This condition is mild, as the admissible δ depends polynomially on n and d. Moreover, this condition can be dropped if one reinforces Assumption D0 (resp., D2) by assuming that $\mathbf{\Sigma}^{-1/2} X$ (resp., $\mathbf{H}^{-1/2} \widetilde{X}$) is almost surely bounded. The corresponding modifications of Theorems 3.1–3.2 are given in the *arXiv* version of this paper [OB18, Thms 3.1–3.2].

As we previously discussed (cf. Remark 2.2), Assumptions D0–D2, although local, are quite restrictive, as they assume light-tailed behavior. Next we discuss how these assumptions can be relaxed.

[5] This issue arises due to working with individual losses; as a result, it does not appear in our refined results, presented in Section 4, in which we analyze the empirical risk "as a whole".

Extension to heavy-tailed distributions. To extend the results, we might use the confidence-boosting technique based on a version of the multi-dimensional sample median as proposed in [HS16]. This allows to completely get rid of Assumption D1, only assuming the existence of the covariance matrix $\mathbf{G}(\theta_*)$. To use the technique, one first divides the sample into $k = \log(e/\delta)$ non-overlapping subsamples, and computes the corresponding M-estimators $\widehat{\theta}^{(1)}, ..., \widehat{\theta}^{(k)}$ over each subsample. Then, one aggregates them through [HS16, Algorithm 3], by using

$$\mathrm{dist}^{(i)}(\theta) := \|\theta - \widehat{\theta}^{(i)}\|_{\widehat{\mathbf{H}}^{(i)}}, \quad \widehat{\mathbf{H}}^{(i)} := \nabla^2 L_n(\widehat{\theta}^{(i)})$$

as the random distance oracle related to $\widehat{\theta}^{(i)}$. The final estimator is $\widehat{\theta}^{(\widehat{i})}$ with

$$\widehat{i} \in \operatorname*{Argmin}_{i\in[k]} \left\{ \mathrm{Median} \left[\left(\mathrm{dist}^{(j)}(\widehat{\theta}^{(i)}) \right)_{j\in[k]} \right] \right\}.$$

By Chebyshev's inequality, each $\widehat{\theta}^{(i)}$ admits a fixed-probability version of (28), say, with $\delta = 2/3$. On the other hand, for each $i \in [k]$, one has

$$\tfrac{1}{2}\mathbf{H} \preccurlyeq \nabla^2 L_n(\widehat{\theta}^{(i)}) \preccurlyeq 2\mathbf{H}$$

with fixed probability. Indeed, $\frac{1}{2}L(\theta_*) \preccurlyeq L_n(\theta_*) \preccurlyeq 2L(\theta_*)$ by the analysis in Theorems 3.1–3.2. Then, our integration argument (cf. the proof of Lemmas B.1–B.3 in appendix) allows to relate $L_n(\theta_*)$ to $L_n(\widehat{\theta}^{(i)})$ and results in $\frac{1}{2}L_n(\theta_*) \preccurlyeq L_n(\widehat{\theta}^{(i)}) \preccurlyeq 2L_n(\theta_*)$. Finally, the estimators over different subsamples are mutually independent. Thus, we can apply Theorem 11 of [HS16], which finally yields (30).

A similar technique also allows to somewhat weaken Assumptions D0 and D2, replacing the subgaussian norm $\|\cdot\|_{\psi_2}$ with the subexponential norm $\|\cdot\|_{\psi_1}$ at the expense of an extra logarithmic factor. (By definition, $X \in \mathbb{R}^d$ satisfies $\|X\|_{\psi_1} \le K$ if for any u on the unit sphere one has $(\mathbb{E}[|\langle X, u\rangle|^p])^{1/p} \lesssim Kp$, compared to $K\sqrt{p}$ in the case of $\|\cdot\|_{\psi_2}$, cf. Lemma A.1 in Appendix.) This can be done by replacing Theorem A.2 (high-probability bound for subgaussian distributions) with [Ver12, Theorem 5.48] (fixed-probability bound for subexponential distributions), controlling $\mathbb{E}[\max_{i\in[n]} \|X_i\|^2_{\mathbf{H}}]$ and $\mathbb{E}[\max_{i\in[n]} \|\widetilde{X}_i\|^2_{\mathbf{H}}]$ via Bernstein's inequality (Theorem A.1 in Appendix). However, this technique is limited to subexponential distributions of X and $\widetilde{X}$ as required by [Ver12, Theorem 5.48].

On the other hand, replacing Assumptions D0 and D2 with finite-moment assumptions (ideally, finite kurtoses of vectors X and $\widetilde{X}$) is challenging. First of all, sample covariance estimators $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{H}}$ would have to be replaced by some estimators $\bar{\mathbf{\Sigma}}$ and $\bar{\mathbf{H}}$ that admit affine-invariant bounds of the form

$$\tfrac{1}{2}\mathbf{\Sigma} \preccurlyeq \bar{\mathbf{\Sigma}} \preccurlyeq 2\mathbf{\Sigma}, \quad \tfrac{1}{2}\mathbf{H} \preccurlyeq \bar{\mathbf{H}} \preccurlyeq 2\mathbf{H} \tag{33}$$

with high probability, under the existence of only finite moments (ideally, the fourth moment) of X and $\widetilde{X}$ in any direction. Such estimators were recently obtained in [OR19] based on the iterative appication of the truncated covariance estimator analyzed in [WM17]. Computing such an estimator on the hold-out sample would allow to get rid of Assumption D0 in Theorem 3.1. However, this technique by itself does not allow relax Assumption D2, note first that we do not know the true minimizer θ_*, and hence cannot directly compute the robust estimator $\bar{\mathbf{H}}$. A possible remedy, leading to the extension of Theorems 3.1–3.2, is to apply an approximation technique on top of the affine-invariant covariance estimator, similarly to the one used below to prove Theorems 4.1–4.2 with improved critical sample size. As we will discuss in the end of Section 4, this would allow to get rid of Assumptions D0 and D2 in Theorems 3.1–3.2 but not in Theorems 4.1–4.2.

4 Improved results: near-linear critical sample size

As we demonstrate next, the previously obtained bounds on the critical sample size can be improved: essentially, the product of d_{eff} and d can be replaced with their maximum. This requires to slightly strengthen Assumption D2 as follows.

Assumption D2*. *The calibrated design* $\widetilde{X}(\theta) := [\ell''(Y, X^\top\theta)]^{1/2}X$ *satisfies*

$$\|\mathbf{H}(\theta)^{-1/2}\widetilde{X}(\theta)\|_{\psi_2} \le \bar{K}_2(r),$$

where $\mathbf{H}(\theta) = \mathbb{E}[\widetilde{X}(\theta)\widetilde{X}(\theta)^\top]$*, for any* θ *in the Dikin ellipsoid* $\Theta_r(\theta_*)$ *given by*

$$\Theta_r(\theta_*) := \{\theta \in \mathbb{R}^d : \|\theta - \theta_*\|_{\mathbf{H}(\theta_*)} \le r\}.$$

Note that Assumption D2 corresponds to Assumption D2* with $r = 0$, the correspondence being given by $K_2 = \bar{K}_2(0)$. On the other hand, the strengthened assumption is still *local*, i.e., it only concerns the points r-close to θ_*, in the local Hessian metric, rather than in the whole domain Θ. With the new assumption at hand, we now state the improved result for canonically self-concordant losses.

Theorem 4.1. *Assume SCb, D1, and D2* with* $r \gtrsim 1$*. Then, (28), (29) and*

$$L(\widehat{\theta}_n) - L(\theta_*) \lesssim \frac{K_1^2 d_{\text{eff}} \log(e/\delta)}{n} \tag{34}$$

hold as long as

$$n \gtrsim \max\left\{\bar{K}_2^4(r) d \log(ed/\delta),\ K_1^2\bar{K}_2^6(r) d_{\text{eff}} \log(e/\delta)\right\}. \tag{35}$$

Let us briefly explain the key ideas behind this result. First of all, recall that the extra factor d in the bound of Theorem 3.2 appears because self-concordance of the *individual losses* only allows to obtain a second-order approximation of the empirical risk in a small Dikin ellipsoid with radius $O(1/\sqrt{d})$, due to the fact that $\|\widetilde{X}\|_{\mathbf{H}^{-1}} = \Omega(\sqrt{d})$ with high probability. This second-order approximation then allows to localize the estimate as soon as $\|\nabla L_n(\theta_*)\|_{\mathbf{H}^{-1}}$ becomes smaller than the radius of the ellipsoid in which such an approximation holds, cf. the proof of Proposition B.3. Hence, the extra factor d would be eliminated if we managed to provide a second-order Taylor approximation of $L_n(\theta)$ in the constant-radius Dikin ellipsoid $\Theta_c(\theta_*)$. The immediately arising difficulty is that unlike the individual losses, the empirical risk is *not* self-concordant, hence, the desired Taylor approximation cannot be obtained purely by integration. Instead, we conduct a somewhat non-standard argument (see Figure 2) which combines (i) self-concordance of the *population risk* following from Assumption D2*; (ii) self-concordance of the individual losses; (iii) a covering argument in which ellipsoid $\Theta_c(\theta_*)$ is covered with small ellipsoids with radius $O(1/d^\gamma)$ for some $\gamma \ge 1/2$. In particular, we choose $\gamma = 2$: this simplifies the calculations in the final step of the proof without breaking (35) since d^γ enters the analysis under logarithm, when bounding covering numbers.

Next we present a counterpart of Theorem 4.1 for pseudo self-concordant losses. As one might expect, the bound on the critical sample size degrades by ρ.

Theorem 4.2. *Assume SCa, D0, D1,C, and D2* with* $r \gtrsim 1/\sqrt{\rho}$*. Then, (28), (29) and (34) hold whenever*

$$n \gtrsim \max\left\{\bar{K}_2^4(r) d \log(ed/\delta),\ \rho K_0^2 K_1^2 \bar{K}_2^4(r) d_{\text{eff}} \log(e/\delta)\right\}. \tag{36}$$

The two results above merit some discussion.

First, note that, in the case of pseudo self-concordance, the radius of the Dikin ellipsoid in which Assumption D2* is required to hold is $\sqrt{\rho}$ times smaller than in the case of canonical self-concordance. As it will become clear from the proof of Theorem 4.2, this deflation is related to the fact that we cannot control the Hessians of $L(\theta)$ over Dikin ellipsoids with a larger radius, even when Assumption D2* holds on such an ellipsoid. On the other hand, decreasing the radius r of the Dikin ellipsoid allows to control $\bar{K}_2(r)$: in Appendix D we show that, in logistic regression with Gaussian design $X \sim \mathcal{N}(0, \mathbf{\Sigma})$,

$$\bar{K}_2^2(r) \lesssim \bar{K}_2^2(0) + r\sqrt{\rho}.$$

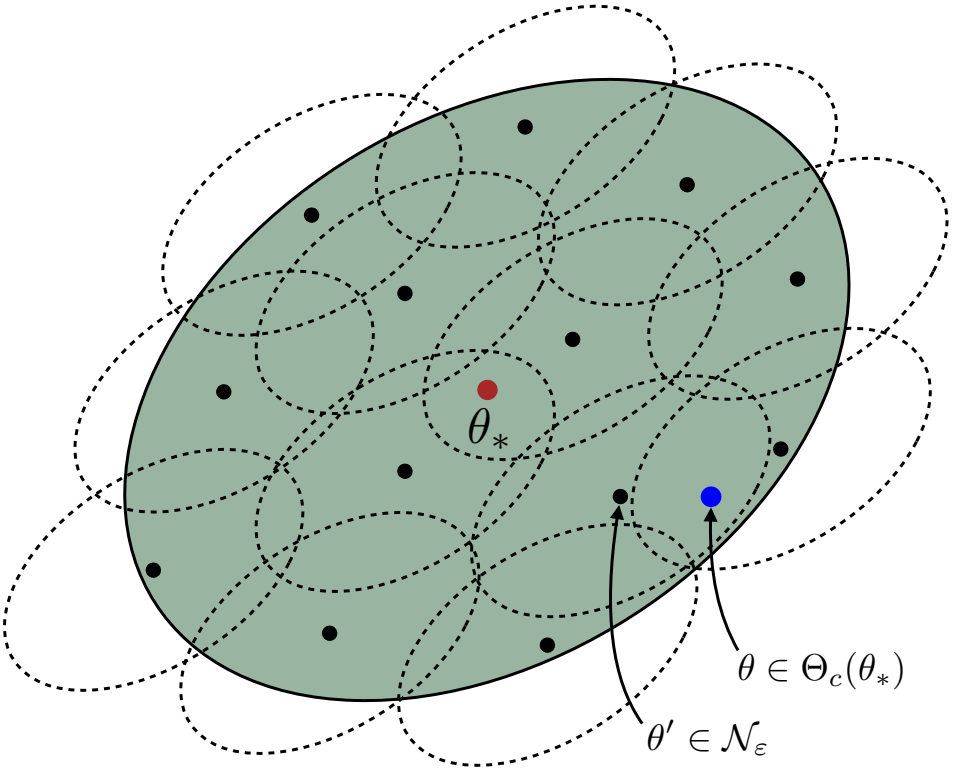

Figure 2: The crucial step in the proof of Theorem 4.1 is to ensure that $\frac{1}{2}\mathbf{H}(\theta_*) \preccurlyeq \mathbf{H}_n(\theta) \preccurlyeq 2\mathbf{H}(\theta_*)$ holds with high probability uniformly over the constant-radius Dikin ellipsoid $\Theta_c(\theta_*)$ (in green). Using Assumption D2*, we first prove that $\frac{1}{2}\mathbf{H}(\theta_*) \preccurlyeq \mathbf{H}(\theta) \preccurlyeq 2\mathbf{H}(\theta_*)$ for any $\theta \in \Theta_c(\theta_*)$. On the other hand, self-concordance of individual losses provides a constant-order approximation of $\mathbf{H}_n(\cdot)$ within a smaller ellipsoid with radius $O(1/d^\gamma)$, for some $\gamma \geq 1/2$, around θ. As such, the problem is reduced to the control of the uniform deviations of $\mathbf{H}_n(\theta)$ from $\mathbf{H}(\theta)$ for $\theta \in \mathcal{N}_\varepsilon$, where $\mathcal{N}_\varepsilon$ is the epsilon-net of $\Theta_1(\theta_*)$ with respect to the norm $\|\cdot\|_{\mathbf{H}(\theta_*)}$ with $\varepsilon = O(1/d^\gamma)$. This is done by using Theorem A.2.

Thus Assumption D2* with $r = 1/\sqrt{\rho}$ is essentially *equivalent* to Assumption D2.

Second, note that the second threshold in (35) has an extra $\bar{K}_2^4(r)$ factor compared to that in (32) if we do not distinguish between $\bar{K}_2(r)$ and $K_2 = \bar{K}_2(0)$, and similarly when comparing (36) and (27). This can be a substantial difference since K_2 and $\bar{K}_2(r)$ can both depend on the norm of θ_*. In fact, in Appendix D (Proposition D.1) we show, by a technical calculation, that in logistic regression with $X \sim \mathcal{N}(0, \mathbf{\Sigma})$ one has

$$\rho \lesssim (1 + \|\theta_*\|_{\mathbf{\Sigma}})^3,$$

this bound being tight, while the bound on $\bar{K}_2(1/\sqrt{\rho})$ is

$$\bar{K}_2(1/\sqrt{\rho}) \lesssim \sqrt{1 + \|\theta_*\|_{\mathbf{\Sigma}}}$$

up to a logarithmic factor. Thus, $\bar{K}_2^4(1/\sqrt{\rho})$ can potentially be as large as $\rho^{2/3}$. On the other hand, when the distribution of $\widetilde{X}(\theta)$ is *log-concave* and centrally symmetric at any $\theta \in \Theta_r(\theta_*)$, the factor $\bar{K}_2^4(r)$ can be eliminated. This amounts to using the improved relation between the third and second moments of the marginals of $\mathbf{H}(\theta)^{-1/2}\widetilde{X}(\theta)$ in step $\mathbf{1^o}$ of the analysis in Theorems 4.1–4.2:

$$\mathbb{E}[|\langle \mathbf{H}(\theta)^{-1/2}\widetilde{X}(\theta), u\rangle|^3| \leq 7(\mathbb{E}[\langle u, \mathbf{H}(\theta)^{-1/2}\widetilde{X}(\theta), u\rangle^2])^{3/2},$$

as follows from [BE15, Lem. 2] by simple algebra, using log-concavity of $\mathbf{H}(\theta)^{-\frac{1}{2}}\widetilde{X}(\theta)$.

Extending Theorems 4.1–4.2 to heavy-tailed distributions. One fact playing the key role in the proofs of the last two theorems is that in the bound

$$K^2(d+\log(e/\delta)) \tag{37}$$

for the sample complexity of estimating a single covariance matrix, the confidence term $\log(1/\delta)$ is *additive with* d. This allows to take the union bound over an exponential in d number of events correponding to the centers of the epsilon net, while still preserving a near-linear in d sample complexity.

As discussed in Section 3, the main technical challenge when trying to extend our results to heavy-tailed distributions is posed by Assumption D2, which for Theorems 4.1–4.2 gets strengthened to Assumption D2*. To get rid of it, one could replace the empirical Hessians $\widehat{\mathbf{H}}(\theta)$ by some estimator $\bar{\mathbf{H}}(\theta)$ that estimates $\mathbf{H}(\theta)$ with high confidence in the positive-semidefinite sense (cf. (33)) under weak moment assumptions. Given such estimators, we can essentially repeat the covering argument in the analysis of Theorems 4.1–4.2, replacing the Hessian estimate in any $\theta \in \Theta_r(\theta_*)$ (with $r = 1$ or $r = 1/\sqrt{\rho}$) with the estimate $\bar{\mathbf{H}}(\theta')$ in the closest center θ' of the cover, and replacing empirical risk minimization with a version of stochastic quasi-Newton algorithm with $\bar{\mathbf{H}}(\theta')$ as the Hessian oracle for $\mathbf{H}(\theta)$. Unfortunately, the only known to us estimator that provably satisfies a high-confidence affine-invariant bound under weak moment assumptions is the one from [OR19], and its sample complexity scales as

$$K^2 d \log(e/\delta),$$

i.e., the confidence term enters *multiplicatively* with d. After taking the union bound over $d^{O(d)}$ events, this bound becomes *quadratic* in d. While this is sufficient to extend Theorems 3.1–3.2, the argument in Theorems 4.1–4.2 is destroyed. Thus, extending the latter theorems, and obtaining near-linear sample complexity, has to rely on $\preccurlyeq$-type covariance estimation with additive confidence, cf. (37). The closest in this direction is the recent work [MZ18] which establishes a high-probability bound in the operator norm, $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| \le c\|\boldsymbol{\Sigma}\|$, holding with probability $\ge 1-\delta$ when $n \ge C(\kappa)[\boldsymbol{r}(\boldsymbol{\Sigma}) + \log(1/\delta)]$, where $C(\kappa)$ is a constant depending only on the kurtosis, and $\boldsymbol{r}(\boldsymbol{\Sigma}) := \mathrm{Tr}(\boldsymbol{\Sigma})/\|\boldsymbol{\Sigma}\| \le d$ is the effective rank. Unfortunately, it is challenging to apply this result in our context, since the operator-norm bounds cannot be translated to $\preccurlyeq$-type guarantees akin to (33) when the estimator is not affine-equivariant. Some progress in this direction has recently been obtained in [OR19]; see [OR19, Sec. 2.3] for a detailed discussion.

5 High-dimensional setup

Our next goal is to extend the results obtained so far to the high-dimensional setting. Namely, we assume that $\Theta = \mathbb{R}^d$ with $d \gg n$, and that the optimal parameter θ_* is *sparse*, i.e., the number of non-zero components of θ_* is at most $\mathtt{s} \ll d$. Note that if the support $\mathcal{S}$ of θ_* was known, a reasonable estimator could be obtained by replacing X with its projection $X_{\mathcal{S}}$ on $\mathcal{S}$, and minimizing the empirical risk on $\mathcal{S}$. As in the case of quadratic loss, and the classical Lasso estimator, this would lead to the improvement over the results of Section 3–4: the ambient dimension d would be replaced with $\mathtt{s}$, and d_{eff} with the quantity $\mathrm{Tr}(\mathbf{H}_{\mathcal{S}}^{-1}\mathbf{G}_{\mathcal{S}})$ where $\mathbf{G}_{\mathcal{S}} = \mathbb{E}[\ell'(Y, X_{\mathcal{S}}^\top\theta_*)X_{\mathcal{S}}X_{\mathcal{S}}^\top]$ and $\mathbf{H}_{\mathcal{S}} = \mathbb{E}[\ell''(Y, X_{\mathcal{S}}^\top\theta_*)X_{\mathcal{S}}X_{\mathcal{S}}^\top]$. However, in reality $\mathcal{S}$ is unknown, and the common recommendation is to use the ℓ_1-penalized M-estimator, given by

$$\widehat{\theta}_{\lambda,n} \in \operatorname*{Argmin}_{\theta\in\mathbb{R}^d} L_n(\theta) + \lambda\|\theta\|_1. \tag{38}$$

In the case of quadratic loss, it is well-known that the risk of the ℓ_1-penalized estimator, when measured in terms of the ℓ_1-loss or the "prediction" loss corresponding to the design covariance matrix, is within a logarithmic in d factor from the "ideal" risk of the projection oracle, provided that the penalization parameter λ is appropriately chosen, and the design is near-isotropic and subgaussian – see, e.g., [Tib96], [CT07], [BRT09], [JN11]. While the statistical theory for the quadratic loss is almost

complete, this is not yet the case for general M-estimators. Here our goal is to partially close this gap, providing analogues of Theorems 3.1 and 3.2 in the high-dimensional setting. These results extend those obtained in [Bac10] for the logistic loss using pseudo self-concordance, and are close to those proved in [vdGM12]; we discuss the connections with these works in the end of this section. Finally, notice that we do not prove analogues of Theorems 4.1–4.2, which would have resulted in a near-linear, rather than quadratic, dependency of the critical sample size from $\mathtt{s}$. We leave such extensions for future work.

We now introduce the final assumption complimentary to Assumption C.

Assumption C*. *One has $\boldsymbol{\Sigma} = \mathbf{I}$. Moreover, for some $\varkappa_1, \varkappa_2 > 0$ it holds*

$$\mathbf{G} \preccurlyeq \varkappa_1 \mathbf{I}, \quad \mathbf{H} \preccurlyeq \varkappa_2 \mathbf{I}.$$

Together, Assumptions C and C* imply the bounds in operator norm:

$$\|\mathbf{G}\|_\infty \le \varkappa_1, \quad \|\mathbf{H}\|_\infty \le \varkappa_2, \quad \|\mathbf{H}^{-1}\|_\infty \ge 1/\rho.$$

Moreover, we can reasonably expect that in the ill-specified case, $\mathbf{G} \succcurlyeq \mathbf{H}$, which is a stronger version of the natural inequality $d_{\text{eff}} \ge d$. When this is the case, the eigenvalues of both $\mathbf{H}$ and $\mathbf{G}$ belong to the interval $[\rho^{-1}, \overline{\varkappa}]$ where $\overline{\varkappa} := \max(\varkappa_1, \varkappa_2)$. Then, the product

$$Q := \rho \overline{\varkappa}$$

can be considered as the condition number of the estimation problem at hand. In particular, we are about to see that the excess risk bounds, as well the bounds for the critical sample size, get inflated by Q in the high-dimensional regime. This reflects the requirement that the problem should be well-conditioned with respect to the *standard* coordinate basis, since both ℓ_0-"norm" and ℓ_1-norm depend on the choice of the basis. Some further remarks are given below.

- Similarly to the bound (25), we can always bound $\varkappa_1$ and $\varkappa_2$:

$$\varkappa_1 \le \sup_{(y,\eta) \in \mathcal{Y} \times \mathbb{R}} |\ell'(y, \eta)|, \quad \varkappa_2 \le \sup_{(y,\eta) \in \mathcal{Y} \times \mathbb{R}} \ell''(y, \eta).$$

  Arguably, these bounds are more informative than the bound (25) for ρ, as they involve the suprema of the loss derivatives (e.g., the right-hand sides are constants for pseudo-Huber and logistic losses).

- Correlated designs can also be considered, but this would lead to the inflation of the bounds by the condition number of $\boldsymbol{\Sigma}$. This is natural, as ℓ_1-regularization fixes the basis, and the estimator is not affine-invariant.

The next result characterizes the statistical properties of the ℓ_1-penalized M-estimator (38) with a canonically self-concordant loss, extending Theorem 3.1.

Theorem 5.1. *Assume SCa, D0, D1, D2, C, C*, and $|\theta_*|_0 \le \mathtt{s}$.*

1. *Whenever*

$$n \gtrsim \max\left\{\rho \varkappa_2 K_2^4 \mathtt{s} \log(ed/\delta)\,,\ \rho^2 \varkappa_1 K_0^2 K_1^2 \mathtt{s}^2 \log(edn/\delta)\right\}, \tag{39}$$

   and the regularization parameter satisfies

$$K_1 \sqrt{\frac{\varkappa_1 \log(ed/\delta)}{n}} \lesssim \lambda \lesssim \frac{1}{\rho K_0 \mathtt{s} \sqrt{\log(edn/\delta)}}, \tag{40}$$

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

$$\|\widehat{\theta}_{\lambda,n} - \theta_*\|_1 \lesssim \rho \mathtt{s} \lambda, \quad \|\widehat{\theta}_{\lambda,n} - \theta_*\|_{\mathbf{H}}^2 \lesssim \rho \mathtt{s} \lambda^2. \tag{41}$$

2. *Define* $\mathcal{E} := \{\|X\|_\infty \lesssim K_0\sqrt{\log(ed/\delta)}\}$. *Then,* $\mathbb{P}(\mathcal{E}) \geq 1-\delta$, *and whenever*

$$\delta \lesssim \left(\frac{\lambda}{K_1\sqrt{\varkappa_1 \log(ed)}}\right)^{1+\frac{1}{\log(d)}},$$

the restricted risk $L_{\mathcal{E}}(\theta) := \mathbb{E}[\ell_Z(\theta)\mathbb{1}_{\mathcal{E}}(X)]$ *w.p. at least* $1-\delta$ *satisfies*

$$L_{\mathcal{E}}(\widehat{\theta}_{\lambda,n}) - L_{\mathcal{E}}(\theta_*) \lesssim \rho \mathtt{s} \lambda^2. \tag{42}$$

Clearly, the right choice of λ is the one attaining the lower bound in (40):

$$\lambda \approx K_1\sqrt{\frac{\varkappa_1 \log(ed/\delta)}{n}}$$

This choice is always possible since the left-hand side in (40) is upper-bounded with the right-hand side due to the second bound in (39). With such λ, both the prediction error and the (restricted) excess risk $L_{\mathcal{E}}(\widehat{\theta}_{\lambda,n}) - L_{\mathcal{E}}(\theta_*)$ are at most

$$O\left(\frac{Q\mathtt{s}\log(ed/\delta)}{n}\right)$$

whenever $n \gtrsim \max(Q\mathtt{s}, \rho Q\mathtt{s}^2)\log(ed/\delta)$, ignoring the dependence on the subgaussian constants. Thus, in the case of pseudo self-concordant losses, d and d_{eff} both get replaced with $\mathtt{s}$, at the expense of extra $O(Q\log d)$ factor in the bounds.

Next we state a version of Theorem 5.1 for canonically self-concordant losses.

Theorem 5.2. *Assume SCb, D1, D2, C, C*, and* $|\theta_*|_0 \leq \mathtt{s}$.

1. *Whenever*

$$n \gtrsim \max\left\{\rho\varkappa_2 K_2^4 \mathtt{s}\log(ed/\delta),\ \rho^2\varkappa_1\varkappa_2 K_1^2 K_2^2 \mathtt{s}^2 \log(edn/\delta)\right\} \tag{43}$$

and the regularization parameter satisfies

$$K_1\sqrt{\frac{\varkappa_1\log(ed/\delta)}{n}} \lesssim \lambda \lesssim \frac{1}{\rho K_2 \mathtt{s}\sqrt{\varkappa_2\log(edn/\delta)}}, \tag{44}$$

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

$$\|\widehat{\theta}_{\lambda,n} - \theta_*\|_1 \lesssim \rho\mathtt{s}\lambda, \quad \|\widehat{\theta}_{\lambda,n} - \theta_*\|_{\mathbf{H}}^2 \lesssim \rho\mathtt{s}\lambda^2. \tag{45}$$

2. *The event* $\mathcal{E} := \{\|\widetilde{X}\|_\infty \lesssim K_2\sqrt{\varkappa_2\log(ed/\delta)}\}$ *satisfies* $\mathbb{P}(\mathcal{E}) \geq 1-\delta$. *Moreover, whenever*

$$\delta \lesssim \left(\frac{\lambda}{K_1\sqrt{\varkappa_1\log(ed)}}\right)^{1+\frac{1}{\log(d)}},$$

the restricted risk $L_{\mathcal{E}}(\theta) := \mathbb{E}[\ell_Z(\theta)\mathbb{1}_{\mathcal{E}}(X)]$ *w.p. at least* $1-\delta$ *satisfies*

$$L_{\mathcal{E}}(\widehat{\theta}_{\lambda,n}) - L_{\mathcal{E}}(\theta_*) \lesssim \rho\mathtt{s}\lambda^2. \tag{46}$$

Comparison of Theorems 5.1 and 5.2. The usual gain of ρ that we have observed so far for canonically viz. pseudo self-concordant losses is not preserved in ℓ_1-regularized estimators. Instead, the second bound in (39) and the upper bound in (40) get inflated with $\varkappa_2$, and the critical sample size, given the "ideal" choice of the regularization parameter corresponding to the lower bound in (44), becomes $n \gtrsim \max(Q\mathtt{s}, Q^2\mathtt{s}^2)\log(ed/\delta)$. Essentially, the reason for that is that ℓ_1-regularization does not "know" anything about the matrices $\mathbf{H}$ and $\mathbf{H}_n$, and, in a sense, violates the affine-invariant structure of the proofs for non-regularized M-estimators. This seems to be a fundamental problem with ℓ_1-regularization, rather than the artifacts of our proofs, since ℓ_1-regularized M-estimators are *themselves* not affine-invariant. As such, we believe the additional factors of Q and Q^2 to be unimprovable in the high-dimensional setup without further assumptions.

Comparison with prior work. Theorem 5.1 extends the result of [Bac10, Theorem 5] for logistic regression with fixed design, obtained using the pseudo self-concordance of the logistic loss. While the established error bounds are similar, our results have important novelties. First, we analyze the random-design setting, whereas [Bac10] assumes fixed design. Second, the result of [Bac10] requires larger sample size, scaling with the product of $\mathtt{s}$ and R^2 where R is an upper bound on $\|X\|_2$. Typically, R scales as $\Omega(\sqrt{d})$ (e.g., this is the case where the design is pre-generated by sampling from a subgaussian distribution), thus [Bac10] essentially proves the bound $O(\mathtt{s}d)$ for the critical sample size.

On the other hand, our results can be compared to those in [vdGM12] who establish the rate $O(\lambda\mathtt{s})$ for the ℓ_1-error and $O(\lambda^2\mathtt{s})$ for the prediction error (see their Theorems 5.2 and 7.3), addressing a larger class of models including GLMs with non-canonical link functions, and general convex robust losses. However, in order to control the precision of the local quadratic approximations of the risk, the authors of [vdGM12] assume that $\ell''(Y, X^\top\theta_*)$ is bounded from below (Conditions A4 and B), which can only be guaranteed by assuming that θ_* is bounded in ℓ_1-norm. Thus, their results do not address the case of unbounded parameter. Remarkably, these results similarly require the sample size to scale as $\Omega(\mathtt{s}^2\log d)$.

Remark 5.1. *In the proofs of Theorems 5.1–5.2, matrices $\mathbf{H}$ and $\mathbf{H}_n$ only interact with residual Δ which with high probability satisfies the restricted subspace condition (105). Hence, we can strengthen the result, replacing Assumption C and the inequality $\mathbf{H} \preccurlyeq \varkappa_2\mathbf{I}$ in Assumption C* with the requirement that*

$$\|\Delta\|_2^2/\rho \le \|\Delta\|_{\mathbf{H}}^2 \le \varkappa_2\|\Delta\|_2^2$$

in the case where $\Delta \in \mathbb{R}^d$ is approximately sparse*, i.e., satisfies $\|\Delta - [\Delta]_{\mathtt{s}}\|_1 \le 3\|[\Delta]_{\mathtt{s}}\|_1$, where $[\Delta]_{\mathtt{s}}$ is the projection of Δ to the span of its $\mathtt{s}$ largest coordinates. This observation can be exploited to accelerate computation of the estimator (38) when using proximal Newton-type methods (see [LSS14]) via* Hessian sketching*, i.e., by replacing the estimates $\mathbf{H}_n(\theta)$ with the estimates $\mathbf{H}_m(\theta) := \frac{1}{m}\sum_{j=1}^m \widetilde{X}_j(\theta)\widetilde{X}_j(\theta)^\top$ computed from a small subsample.*

We defer further discussion of related work on ℓ_1-regularized M-estimators to Section 7.

6 Numerical experiments

We now present two numerical experiments that illustrate our theoretical results.[6]

Critical sample size grows linearly with model dimensionality. Here the point is to illustrate the results in Section 4, namely Theorems 4.1–4.2. Recall that, in a nutshell, these results state that the fast $O(d/n)$ rate for the excess risk becomes available starting from the critical sample size which is $O(d_{\text{eff}} \vee d)$, where $O(\cdot)$ hides factors depending on the distribution-dependent constants $K_0, K_1, \bar{K}_2, \rho$

[6] All our codes are available online at `http://github.com/ostrodmit/self-concordant`.

arising in Assumptions D0, D1, D2*, and C. In our first experiment (see Fig. 3), we empirically demonstrate that the critical sample size indeed scales linearly with the parameter dimension. For growing sample size $n = 10^k$, $k \in [1, 3]$, we generate an i.i.d. sample $(X_i, Y_i)_{i=1}^n$ with standard Gaussian design $X_i \sim \mathcal{N}(0, \mathbf{I}_d)$ and conditional distribution of the (binary) label given by $\mathbb{P}[Y_i = 1] = 1/(1 + \exp(-X_i^\top \theta_*))$ (i.e., such that the logistic model is well-specified) or by $\mathbb{P}[Y_i = 1] = 1 - \phi(X^\top \theta^*)$, where $\phi(\cdot)$ is the standard Gaussian c.d.f., which corresponds to the probit regression. Thus, the logistic model for $Y|X$ is well-specified in the second case We take $\theta_* = \mathbb{1}_d/\sqrt{d}$ (thus $\|\theta_*\|_2 = 1$) and consider the following three quantities for $d \in \{8, 16, 32, 64\}$:

1. Excess risk $L(\widehat{\theta}_n) - L(\theta_*)$ of the logistic regression estimator, i.e., for the M-estimator with the logistic loss $\ell(y, \eta) = \log(1 + e^\eta) - y\eta$.

2. Excess risk $L^{\mathsf{SC}}(\widehat{\theta}_n^{\mathsf{SC}}) - L^{\mathsf{SC}}(\theta_*^{\mathsf{SC}})$ for the M-estimator with loss (22) – canonically self-concordant analogue of the logistic loss proposed in Sec. 2.1. Here $L^{\mathsf{SC}}(\theta) := \mathbb{E}[\ell^{\mathsf{SC}}(Y, X^\top \theta)]$ with $\ell^{\mathsf{SC}}(y, \eta)$ given by (22); θ_*^{SC} minimizes $L^{\mathsf{SC}}(\theta)$ and might be different from θ_*. Note that $\ell^{\mathsf{SC}}(y, \cdot)$ and $\ell(y, \cdot)$ have the same second-order Taylor expansion around $\eta = 0$ (see Fig.1).

3. Excess risk $L(\widehat{\theta}_n^{\mathsf{SC}}) - L(\theta_*)$ that evaluates $\widehat{\theta}_n^{\mathsf{SC}}$ as a surrogate estimator.

In all three cases, we approximate the excess risk via a test sample with $N = 10^4$ observations, and we compute θ_*^{SC} by running fmincon optimization routine in Matlab (we use the constraint $\|\theta\|_2 \le 2$ to avoid numerical instabilities). Then, for each value of d and the three notion of excess risk, we plot the excess risk against the sample size in the $\log_{10} - \log_{10}$ scale. The experiment is repeated $T = 800$ times, and the averaged curve is then plotted along with a 3σ-confidence interval.

The results are shown in Fig. 3. We can distinctively see the elbow effect: the initial slow convergence rate (slope around $-1/2$ on the log-log scale) changes to the fast rate (slope -1) for larger sample size. This is observed for all three curves, all values of d, and both conditional distributions of Y.

- For the logistic distribution, $\widehat{\theta}_n$ outperforms $\widehat{\theta}_n^{\mathsf{SC}}$ in the fast rate zone (i.e., with sample sizes above the critical level) in terms of their corresponding "native" risks as well as the logistic risk. This is expected: while $\widehat{\theta}_n$ is well-specified, estimator $\widehat{\theta}_n^{\mathsf{SC}}$ has to pay for model misspecification, and its excess risk depends on d_{eff} rather than d (cf. (34) in Theorem (4.1)). Meanwhile, for smaller sample sizes $L^{\mathsf{SC}}(\widehat{\theta}_n^{\mathsf{SC}}) - L^{\mathsf{SC}}(\theta_*^{\mathsf{SC}})$ is smaller than the other two excess risks. This seems to be simply due to $\ell^{\mathsf{SC}}(y, \eta)$ being smaller than $\ell(y, \eta)$ away from $\eta = 0$ (cf. Fig. 1).

- In the case of probit distribution, both estimators are misspecified, and turn out to have very close performance in terms of all three excess risks.

Finally, and most importantly, we see that the "elbow" on the curves moves to the right *in (roughly) constant increments* as we increase d geometrically. This is what we expect: according to Theorems 4.2–4.1, the critical sample size grows linearly with d or d_{eff} in the misspecified case (and here d_{eff} is itself linear in d).

Critical sample size growing as e^{RD} for "bad" design distributions. Here we empirically investigate the dependency of constants $K_0, K_1, \bar{K}_2, \rho$ from the norm $D = \|\theta_*\|_{\boldsymbol{\Sigma}}$ of the population risk minimizer. Recall that in Appendix D we provide polynomial bounds in the case of logistic regression with Gaussian design. However, for certain (quite artificial) distributions of design the dependency might be exponential as implied by the results of [HKL14]. In this experiment, we consider the adversarial distribution proposed in [HKL14, Sec. 3.2], in which $X \in \mathbb{R}^2$ is supported on three points with carefully chosen probabilities (see [HKL14, Figs. 3-4]) and $Y \equiv 1$. The authors prove the $\Omega(1/\sqrt{n})$ lower bound (and hence the absence of fast rate) for the excess risk long as $n \lesssim e^{RD}$, where $R = \|X\|_{\boldsymbol{\Sigma}^{-1}}$. We empirically discover a similar phenomenon for the self-concordant loss (22). To this end, we follow a similar protocol

as in the previous experiment but generate the pairs (X_i, Y_i) according to the distribution in [HKL14] and linearly increase D while fixing $d = 2$. The experiment is repeated $T = 1600$ times for sample sizes $n \in [10^1, 10^4]$, and the population risk is approximated via a test sample with size $N = 5 \cdot 10^4$. We then plot the same three dependencies as in the previous experiment (again in the log-log scale) for $D \in \{1, 3, 5, 7\}$.

The results are presented in Fig. 4. For small sample sizes the curves oscillate, which seems to be due to the special low-dimensional structure of the design distribution. However, the upper envelope of the curve clearly exhibits the same "elbow" effect as before: the slope changes from roughly $-1/2$ to -1 for large sample sizes. Moreover, the horizontal location of the elbow moves in nearly uniform increments as we change D linearly, precisely as expected from the theory in [HKL14]. We also note that the "transfer" risk $L(\widehat{\theta}_n^{\mathsf{SC}}) - L(\theta_*)$ converges to a non-zero value, which shows that $\theta_*^{\mathsf{SC}} \neq \theta_*$ for the distribution considered here.

7 Related work

Self-concordant analysis of logistic regression. Our approach is inspired by [Bac10], and we reuse and extend some of their technical results in our Propositions B.3–B.4. However, our results and analysis are crucially different from those in [Bac10] in several ways. First, we address the random-design setting, whereas in [Bac10] the design is fixed. Second, [Bac10] considers only pseudo self-concordant losses, focusing on logistic regression, whereas we also provide results for canonically self-concordant losses, and, crucially, compare the two cases. Third, we obtain similar results for ill-specified models, whereas [Bac10] only establishes a slow rate in this case. Finally, and most importantly, while we use very similar tools to those in [Bac10], the "core" of our analysis is more direct. Namely, [Bac10] studies the minimizer $\widehat{\theta}_{\lambda,n}$ of *the ℓ_2-penalized empirical risk* with strictly positive regularization parameter λ, and moreover, imposes some technical condition on the minimal magnitude of λ, see their Eq. (13). Upon close inspection, this condition implies

$$n \gtrsim \rho \cdot \mathtt{df}_\lambda^2, \quad \mathtt{df}_\lambda := \mathrm{Tr}[\mathbf{H}(\mathbf{H} + \lambda \mathbf{I})^{-1}], \tag{47}$$

where the *degrees of freedom* parameter $\mathtt{df}_\lambda$ replaces d in the ℓ_2-penalized setting. This, in turn, allows to carry out an argument analogous to ours, but applying Proposition B.4 to the *regularized* empirical risk. However, ℓ_2-penalization makes the analysis much more involved, as it rests on the comparison of the regularized risks, and accordingly, relates θ_* and $\widehat{\theta}_{\lambda,n}$ through the intermediate point – the minimizer θ_λ of the regularized average risk. The extra condition in [Bac10], which makes this analysis possible, is non-trivial, and requires some fine balance between the regularization parameter, sample size, and various types of degrees of freedom and biases. We manage to circumvent these difficulties for the plain ERM, including the ill-specified case, by realizing that the only condition needed to carry out the argument based on self-concordance, in the non-regularized case, is the sufficient sample size.

Self-concordant analysis and improper algorithms. Another relevant work is [Bac14] which studies logistic regression with random design, but analyzes an estimate computed by stochastic approximation with averaging. While this estimator is more advantageous from the computational standpoint, the control of the distance to the optimum is more involved (see [Bac14, Proposition 7]) which leads to the suboptimal risk bound

$$\mathbb{E}_n[L(\widehat{\theta})] - L(\theta_*) \lesssim \frac{R^2(R^4 D_0^4 + 1)}{\mu n}, \tag{48}$$

where μ is the minimal eigenvalue of $\mathbf{H}$, R is an upper bound for $\|X\|_2$ and $\sup_{\theta \in \Theta} \|\nabla \ell_Z(\theta)\|_2$, and $D_0 := \|\theta_0 - \theta_*\|_2$ is the initial ℓ_2-distance from the optimum (in fact, if D_0 is known up to a constant factor, $R^4 D_0^4$ in (48) can be replaced with $R^2 D_0^2$). The bound (48) reflects the fact that gradient

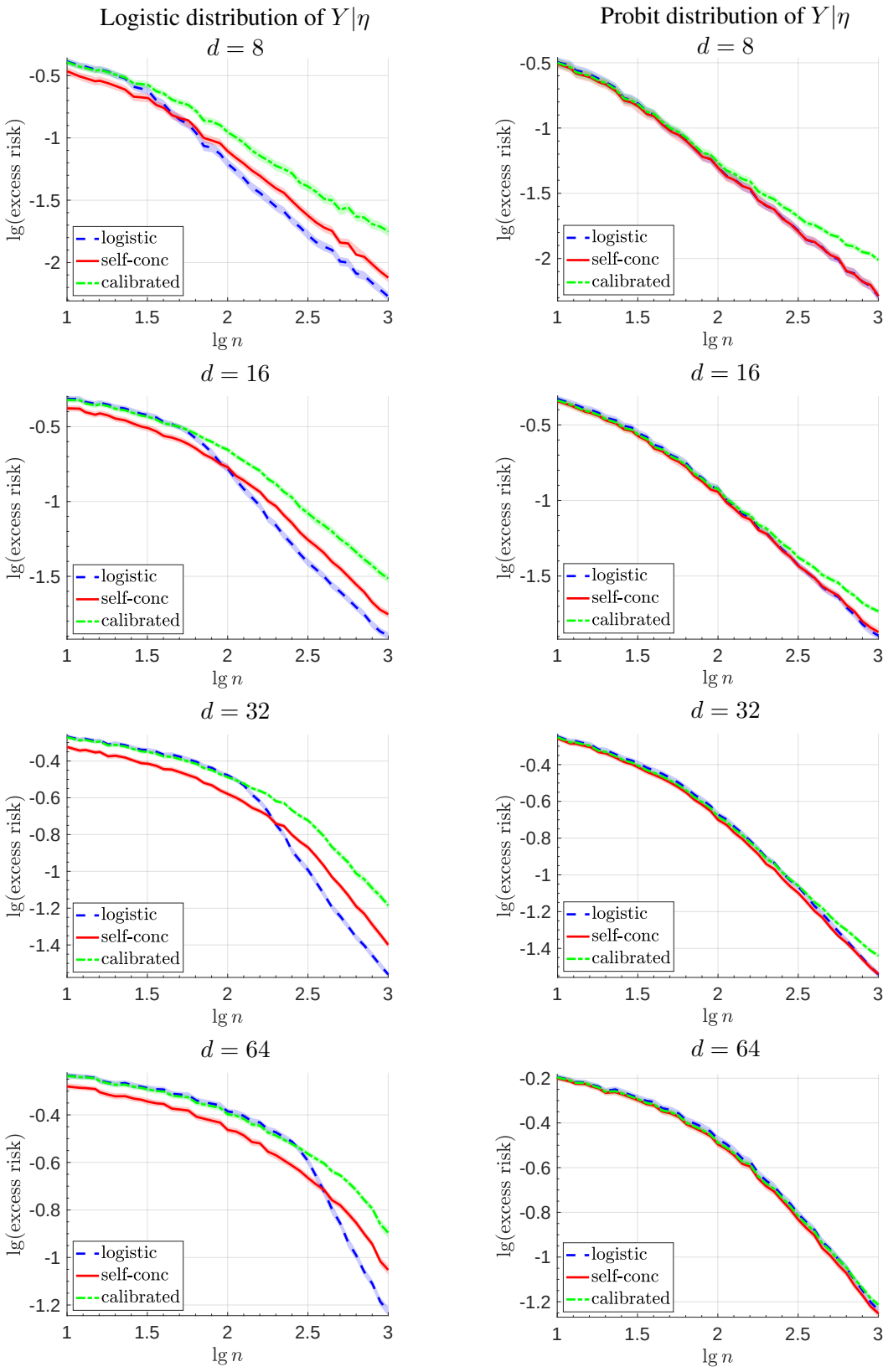

Figure 3: The comparison of two M-estimators: with the logistic loss (estimator $\widehat{\theta}_n$) and its canonically self-concordant analogue (22) (estimator $\widehat{\theta}_n^{\mathsf{SC}}$) in the first experiment. *"Logistic"*, *"self-conc"* and *"calibrated"* correspond to the three notions of excess risk: the "native" risks for $\widehat{\theta}_n, \widehat{\theta}_n^{\mathsf{SC}}$ and the "transfer" risk for $\widehat{\theta}_n^{\mathsf{SC}}$ with logistic loss (see p. 22).

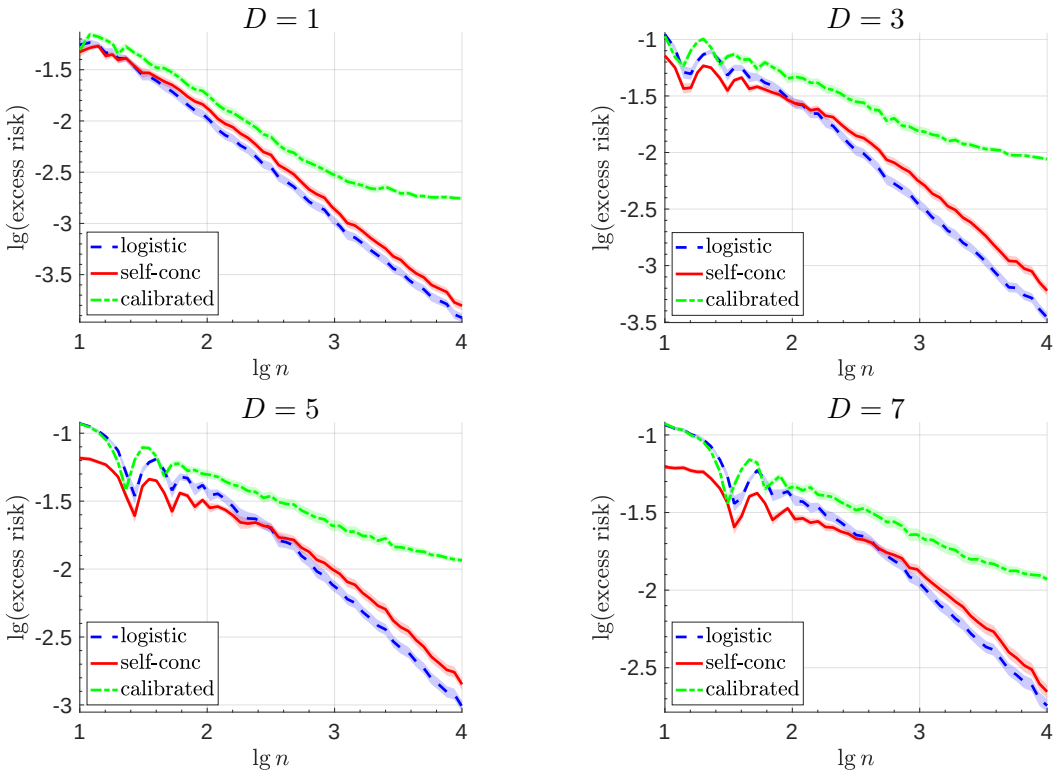

Figure 4: The comparison of M-estimators $\widehat{\theta}_n, \widehat{\theta}_n^{\mathsf{SC}}$ in the second experiment, using the adversarial data distribution from [HKL14] (see p. 22 for more details). *"Logistic"*, *"self-conc"* and *"calibrated"* correspond to the same three notions of excess risk as in the first experiment (see Fig. 3).

descent trajectory is not affine-invariant, hence the distances are not "measured" in terms of the natural norm $\|\cdot\|_{\mathbf{H}}$. For the *natural gradient*, that is, gradient descent on the tranformed problem $\tilde{\theta} = \mathbf{H}^{1/2}\theta$, factor μ would disappear from (48), but R would be replaced with $\max(d_{\text{eff}}, \rho \cdot d)$, and D_0 with the initial prediction distance $\|\theta_0 - \theta_*\|_{\mathbf{H}}$, which would lead to a bound scaling as the cube of $\max(d_{\text{eff}}, \rho d)$. The follow-up work [BM13] studies a version of the quasi-Newton method, solving the quadratic subproblems via stochastic approximation. This allows to conduct affine-invariant analysis of the outer loop, and results in

$$\mathbb{E}_n[L(\widehat{\theta})] - L(\theta_*) \lesssim \frac{\rho^2(R^4 D_0^4 + 1)\max(d_{\text{eff}}, \rho d)}{n} \tag{49}$$

whenever $n \gtrsim (R^4 D_0^4 + 1)$. It should be noted that the curvature parameter ρ that appears in these results, as well as in our results for pseudo self-concordant losses, is *problem-dependent*. In particular, it depends on the true distribution $\mathcal{P}$ of the data, and can be very large if this distribution is chosen adversarially. By constructing such an adversarial distribution, [HKL14] prove a *lower bound* $\Omega(\sqrt{RD/n})$, i.e., for the excess risk of any algorithm, in logistic regression in the finite-sample regime $n = O(e^{RD})$. This implies that ρ grows super-polynomially in RD for this distribution. Notably, the lower bound of [HKL14] is not applicable in the setting of *improper prediction*, where one is allowed to estimate $\eta_* := X^\top \theta_*$ with any predictor $\widehat{\eta} : X \mapsto \mathbb{R}$, not necessarily with a linear one. Making such an observation, [FKL+18] recently proposed an improper estimator which attains the excess risk $O(d/n)$ up to logarithmic factors in RD, n, and $1/\delta$. Their estimator reduces to Vovk's Aggregating Algorithm [Vov98] for online convex optimization, combined with a simple "boosting the confidence" scheme proposed in [Meh17].

Non-parametric setup and ℓ_2-regularization. After the arXiv preprint of this work began circulating,[7] our analysis was extended in [MFOBR19] to M-estimators with ℓ_2-regularization, including the case of infinite-dimensional parameter. The reader might refer to [MFOBR19] for an overview of related work in this direction. In a nutshell, [MFOBR19] proves asymptotically near-optimal bounds $\widetilde{O}(d_{\text{eff}}/n)$ on the "variance" term corresponding to the excess risk $L(\widehat{\theta}_{\lambda,n}) - L(\theta_\lambda)$, and the additional "bias" term $L(\theta_\lambda) - L(\theta)$, under condition (47), and without extra conditions in the vein of those in [Bac10]. Moreover, it is shown that the classical source and capacity conditions [CDV07], known to lead to faster non-parametric rates in ridge regression, can be extended to M-estimators with self-concordant losses. However, [MFOBR19] does not extend our improved results with near-linear sample size (Theorems 4.1–4.2) to the ℓ_2-penalized case. We believe that such extension is possible by replacing Theorem A.2 with a similar result for regularized covariance matrices such as [KL17, Thm. 9]. This, however, would be of little practical interest, since typically under source condition, $\mathtt{df}_\lambda$ is a constant depending only on the rate of decay of the eigenvalues of $\mathbf{H}$.

Quasi-Newton algorithms. We also mention in passing that recently there has been a surge of interest in stochastic quasi-Newton methods applied to the finite-sum setting with self-concordant losses, see, e.g., [ZL15], [ZGG17]. However, none of these works establishes generalization bounds for the associated estimator. In fact, such bounds have recently been established in the work [MFBR19] for a certain (globally convergent) ad-hoc quasi-Newton scheme. These generalization bounds are similar to those established in [MFOBR19] for the exact ERM, with similar criticism.

Empirical processes. The use of empirical processes in the context of parametric estimation was pioneered in [Spo12], which has been one of the main inspirations for our work. Apart from the technical difficulty in the proofs, our main critique of [Spo12] is the *global* conditions they impose – most importantly, they require $\nabla L_n(\theta)$ to be subgaussian uniformly over the whole parametric set Θ. As can be seen from the proof of Proposition D.1 in Appendix D, verification of such global conditions can be quite technical; moreover, such conditions can in fact be way more restrictive than their local counterparts (see, e.g., the bounds in Proposition D.1 which degrade drastically when $\|\theta_*\|_{\mathbf{\Sigma}} \gg 1$, and which are sharp as can be seen from the analysis).

Recently we learned about the work [MBM18] that applies empirical processes to study constrained empirical risk minimization with smooth *non-convex* losses. (Its preprint came out after that of our work.) Essentially, [MBM18] proves that in the regime $n \gtrsim d \log d$ (resp. $n \gtrsim \mathtt{s} \log d$ in the high-dimensional setup), the sample gradients $L_n(\theta)$ and Hessians $\nabla^2 L_n(\theta)$ uniformly converge to their population counterparts, assuming that $\nabla L_n(\theta)$ is subgaussian, and $\nabla^2 L_n(\theta)$ is subexponential on the whole domain Θ. This allows to establish correspondence between the stationary points of $L(\cdot)$ and $L_n(\cdot)$. While the focus of [MBM18] is different, we expect that one may also prove asymptotically optimal rates for the excess risk in this regime, i.e., prove "local" analogues of our improved results (cf. Section 4), by localizing a minimizer $\widehat{\theta}_n$ to the unit Dikin ellipsoid of the associated population risk minimizer θ_*.[8] However, [MBM18] has the same limitations as [Spo12], requiring "global" conditions on the tails of $\nabla L_n(\theta)$ and $\nabla^2 L_n(\theta)$. Similar criticism applies to the literature on ℓ_1-regularized M-estimators in high dimensions ([vdGM12, NRWY12, Loh17]); we discuss these results in more detail in Section 5.

Further related work on ℓ_1-regularized M-estimators. Interestingly, [ZWJ17] showed that, in absense of restricted-eigenvalue (RE) type conditions imposed on the (fixed) design matrix, decomposable regularizers only lead to slow $O(1/\sqrt{n})$ rates, even with quadratic loss. Hence, the light-tailed design condition that we impose appears to be necessary when ℓ_1-regularized M-estimators are considered.

[7] This happened in October 2018.

[8] This, however, would require restating their Assumptions 1-3 in affine-invariant manner.

Not directly relevant to our results here, we note that, whenever the computational considerations are not important, the ℓ_1-regularization can perform worse than other types of regularization. In fact, ℓ_0-regularized estimators are known to achieve $O(\mathtt{s}/n)$ rate for the prediction error without RE-type conditions (in the fixed-design setup) [ZWJ17]. Moreover, there are other (non-convex) penalties that have favorable statistical properties without incoherence, see, e.g., [LW17, LW15, Loh17, LW11].

After the preprint of this work had been publicized, the statistical performance of regularized M-estimators has been studied in the asymptotic regime $n, d \to \infty$ with $d/n = c$, including the "high-dimensional" case with $c \gg 1$ – see [TAH18, SC19].

8 Conclusion

Our work sheds light on the mechanism behind the transition to the fast-rate regime in M-estimators with smooth losses. Our analysis allows to deal with M-estimators with losses satisfying self-concordance-type assumptions, including logistic regression, other generalized linear models, and robust regression. Self-concordance assumptions allow to control the precision of the local quadratic approximations of empirical risk. Simple analysis under minimal assumptions leads to a fast-rate guarantee for large sample sizes – larger (in order) than $d \cdot d_{\text{eff}}$, where d_{eff} is the effective dimensionality of the parametric model. However, a refined analysis under slightly stronger assumptions leads to the $O(\max\{d_{\text{eff}}, d \log d\})$ sample size threshold. This is done through a combination of self-concordance with a covering argument, allowing to control the uniform deviations of the empirical risk Hessian in the Dikin ellipsoid around the population risk minimizer. We also extend the analysis to ℓ_1-regularized M-estimators in the high-dimensional regime. Finally, we verify the empirical performance of a canonically self-concordant analogue of the logistic loss in numerical experiments.

Acknowledgments

The first author has been supported by the ERCIM Alain Bensoussan Fellowship while working on this project. The second author acknowledges the support of the European Research Council (grant SEQUOIA 724063).

A Probabilistic tools

A.1 Subgaussian distributions

We recall the definition of subgaussian norm for random variables $\xi \in \mathbb{R}$:

$$\|\xi\|_{\psi_2} := \inf\left\{\sigma > 0 : \ \mathbb{E}[e^{\xi^2/\sigma^2}] \le e\right\}.$$

The lemma below provides equivalent definitions of the subgaussian norm.

Lemma A.1 ([Ver12, Lemma 5.5])**.** *There exists an absolute constant $c > 0$ such that $\|\xi\|_{\psi_2} \le \sigma$ is equivalent to either of the following:*

1. *Subgaussian tails: for any $t \ge 0$, $\mathbb{P}\left\{|\xi| > t\right\} \le \exp\left(1 - ct^2/\sigma^2\right)$.*
2. *Subgaussian moments: for any $p \ge 1$, $\mathbb{E}[|\xi|^p]^{1/p} \le c\sigma\sqrt{p}$.*

Moreover, if $\mathbb{E}[\xi] = 0$, each of these properties is equivalent to the moment bound

$$\mathbb{E}\exp(t\xi) \le \exp(c\sigma^2 t^2).$$

Following [Ver12], we define the $\|\cdot\|_{\psi_2}$-norm of a random vector as follows:

$$\|Z\|_{\psi_2} := \sup_{\theta \in \mathcal{S}^{d-1}} \|\langle Z, \theta \rangle\|_{\psi_2}, \tag{50}$$

where $\mathcal{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$. Note that this is indeed a norm; in particular, it satisfies the triangle inequality: $\|Z_1 + Z_2\|_{\psi_2} \le \|Z_1\|_{\psi_2} + \|Z_2\|_{\psi_2}$ for any pair of random vectors Z_1, Z_2. Another elementary property is that $\|\mathbf{J}Z\|_{\psi_2} \le \|\mathbf{J}\|_\infty \|Z\|_{\psi_2}$ for arbitrary matrix $\mathbf{J}$. Some well-known properties of subgaussian random vectors are summarized in the following lemmas.

Lemma A.2. *Let the entries of $Z \in \mathbb{R}^d$ satisfy $\|Z_i\|_{\psi_2} \le K$, $i \in [d]$. Then, with probability at least $1-\delta$,*

$$\|Z\|_\infty \lesssim K\sqrt{\log(ed/\delta)}.$$

Proof. The claim follows from Item 1 of Lemma A.1 by the union bound. ∎

Next we give a bound for the p-th moment of $\|Z\|_\infty$. Although this bound is loose for any fixed p, we only use it in the regime $p \approx \log d$ where it is tight.[9]

Lemma A.3. *In the assumptions of the previous lemma, for any $p \ge 1$ it holds*

$$\mathbb{E}[\|Z\|_\infty^p]^{1/p} \lesssim K d^{1/p}\sqrt{p}.$$

Proof. Using the bound from Lemma A.2, we obtain

$$\begin{aligned}\mathbb{E}[\|Z\|_\infty^p] &= \int_0^\infty \mathbb{P}\left\{\|Z\|_\infty \ge u\right\} \mathrm{d}(u^p) \le ed\int_0^\infty e^{-\frac{c^2u^2}{K^2}}\mathrm{d}(u^p)\\ &\le ed\left(\frac{K}{c}\right)^p \frac{p}{2}\Gamma\left(\frac{p}{2}\right) \le ed\left(\frac{K}{c}\right)^p \frac{p}{2}\left(\frac{p}{2}\right)^{p/2} = \frac{d(K\sqrt{p})^p ep}{2(c\sqrt{2})^p}.\end{aligned}$$

We obtain the claim by extracting the p-th root and doing simple estimates. ∎

Lemma A.4 (Hoeffding-type inequality, follows from [Ver12, Lemma 5.9] via (50))**.** *Let $Z_1, ..., Z_n$ be i.i.d. random vectors, then one has $\|\sum_{i=1}^n Z_i\|_{\psi_2}^2 \lesssim \sum_{i=1}^n \|Z_i\|_{\psi_2}^2$.*

The next result shows that the $\|\cdot\|_{\psi_2}$-norm is stable under affine transforms.

Lemma A.5 (Subgaussian norm after affine transform and decorrelation)**.** *Suppose that $X \in \mathbb{R}^d$ satisfies $\mathbb{E}[X] = 0$, $\mathbf{\Sigma} := \mathbb{E}[XX^\top]$, and $\|\mathbf{\Sigma}^{-1/2}X\|_{\psi_2} \le K$. Then for any $A \in \mathbb{R}^{d\times d}$, $b \in \mathbb{R}^d$, vector $\widehat{X} = AX + b$ satisfies*

$$\|\widehat{\mathbf{\Sigma}}^{-1/2}\widehat{X}\|_{\psi_2} \lesssim K, \quad \text{where} \quad \widehat{\mathbf{\Sigma}} = \mathbb{E}[\widehat{X}\widehat{X}^\top].$$

Proof. The quantity $\mathbf{\Sigma}^{-1/2}X$ is invariant with respect to linear transforms, so it only remains to treat the case $\widehat{X} = X + b$. Now, in this case, $\widehat{\mathbf{\Sigma}} = \mathbf{\Sigma} + bb^\top$, and

$$\|\widehat{\mathbf{\Sigma}}^{-1/2}\widehat{X}\|_{\psi_2} \le \|\widehat{\mathbf{\Sigma}}^{-1/2}X\|_{\psi_2} + \|\widehat{\mathbf{\Sigma}}^{-1/2}b\|_{\psi_2} \le \|\widehat{\mathbf{\Sigma}}^{-1/2}X\|_{\psi_2} + \|\widehat{\mathbf{\Sigma}}^{-1/2}b\|_2.$$

Since $\widehat{\mathbf{\Sigma}} \succcurlyeq \mathbf{\Sigma}$, we have $\|\widehat{\mathbf{\Sigma}}^{-1/2}X\|_{\psi_2} \le \|\mathbf{\Sigma}^{-1/2}X\|_{\psi_2} \le K$. On the other hand,

$$\|\widehat{\mathbf{\Sigma}}^{-1/2}b\|_2^2 = b^\top\widehat{\mathbf{\Sigma}}^{-1}b^\top \le 1.$$

by the Sherman-Morrison formula. Finally, note that $K \gtrsim 1$, as follows from the inequality $\mathbb{E}[\xi^4] \ge (\mathbb{E}[\xi^2])^2$ applied to $\xi = \langle u, X\rangle$, and Item 2 of Lemma A.1. ∎

[9] Tight bounds for all moments can be obtained via the Chernoff method combined with the general Orlicz norms $\|\cdot\|_{\psi_\alpha}$ with $\alpha = 2/p$, see, e.g., [Pol90]. It is beyond the scope of this paper.

A.2 Quadratic forms of subgaussian vectors

We call random vector $Z \in \mathbb{R}^d$ *isotropic* if $\mathbb{E}[Z] = 0$ and $\mathbb{E}[ZZ^\top] = \mathbf{I}_d$. The following result is a deviation bound for quadratic forms of isotropic subgaussian random vectors. It is obtained from [HKZ12b, Theorem 2.1] using the isotropicity assumption which allows to get rid of the K^2 factor ahead of $\mathrm{Tr}(\mathbf{J})$.

Theorem A.1. *Let $Z \in \mathbb{R}^d$ be an isotropic random vector with $\|Z\|_{\psi_2} \le K$, and let $\mathbf{J} \in \mathbb{R}^{d\times d}$ be positive semidefinite. Then,*

$$\mathbb{P}\left\{\|Z\|_{\mathbf{J}}^2 - \mathrm{Tr}(\mathbf{J}) \ge t\right\} \le \exp\left(-c\min\left\{\frac{t^2}{K^2\|\mathbf{J}\|_2^2}, \frac{t}{K\|\mathbf{J}\|_\infty}\right\}\right).$$

In other words, with probability at least $1-\delta$ it holds

$$\|Z\|_{\mathbf{J}}^2 - \mathrm{Tr}(\mathbf{J}) \lesssim K^2\left(\|\mathbf{J}\|_2\sqrt{\log(1/\delta)} + \|\mathbf{J}\|_\infty \log(1/\delta)\right).$$

Corollary A.1. *We obtain a deviation bound for the ℓ_2-norm of the projection of an isotropic subgaussian vector Z onto an arbitrary direction $u \in \mathbb{R}^d$: with probability at least $1-\delta$ it holds*

$$|\langle u, Z\rangle| \lesssim \|u\|_2 K\sqrt{\log(e/\delta)}. \tag{51}$$

This follows, through some elementary algebra, by applying Theorem A.1 to the rank-one matrix $\mathbf{J} = uu^\top$ which satisfies $\|\mathbf{J}\|_\infty = \|\mathbf{J}\|_2 = \mathrm{Tr}(\mathbf{J}) = \|u\|_2^2$.

The next result follows from Theorem A.1 since $\|\mathbf{J}\|_\infty \le \|\mathbf{J}\|_2 \le \mathrm{Tr}(\mathbf{J})$.

Corollary A.2. *Under the premise of Theorem A.1, $\zeta = \|Z\|_{\mathbf{J}}$ is subgaussian:*

$$\mathbb{P}\left\{\zeta \ge c(1+t)K\sqrt{\mathrm{Tr}(\mathbf{J})}\right\} \le \exp(-t^2).$$

As a consequence, $\mathbb{P}\left\{\zeta \ge cuK\sqrt{\mathrm{Tr}(\mathbf{J})}\right\} \le \exp\left(c_1 - u^2/c_2\right)$, so that

$$\|\zeta\|_{\psi_2} \le cK\sqrt{\mathrm{Tr}(\mathbf{J})}.$$

A.3 Sample covariance matrices

Next we focus on sample second-moment matrices of subgaussian vectors.

Theorem A.2 ([Ver12, Theorem 5.39])**.** *Assume that the random vector $\widetilde{X} \in \mathbb{R}^d$ satisfies $\mathbb{E}[\widetilde{X}\widetilde{X}^\top] = \mathbf{H}$ and $\|\mathbf{H}^{-1/2}\widetilde{X}\|_{\psi_2} \le K$. Let $\mathbf{H}_n = \frac{1}{n}\sum_{i=1}^n \widetilde{X}_i\widetilde{X}_i^\top$ where $\widetilde{X}_1, ..., \widetilde{X}_n$ are independent copies of $\widetilde{X}$. Whenever*

$$n \gtrsim K^4(d + \log(1/\delta)),$$

with probability at least $1-\delta$ it holds

$$\|\Delta\|_{\mathbf{H}}^2/2 \le \|\Delta\|_{\mathbf{H}_n}^2 \le 2\|\Delta\|_{\mathbf{H}}^2, \quad \forall \Delta \in \mathbb{R}^d. \tag{52}$$

Next we present an extension of this result to the high-dimensional setting.

Theorem A.3 ([Zho09, Theorem 1.6])**.** *Let $\mathbf{H}$, $\mathbf{H}_n$, and $\widetilde{X}$ be as in the previous theorem, and suppose that $\mathbf{H}$ satisfies the $(\rho, \varkappa, \mathtt{s})$-restricted eigenvalues (RE) condition for some $\rho, \varkappa > 0$ and $\mathtt{s} \le d$. Namely, for any $\Delta \in \mathbb{R}^d$ such that $\|\Delta_{\mathcal{S}_c}\|_1 \le 3\|\Delta_{\mathcal{S}}\|_1$, where $\mathcal{S}$ is the subspace of $\mathbb{R}^d$ correponding to $\mathtt{s}$ largest coordinates of Δ, and $\mathcal{S}_c$ is the complement of $\mathcal{S}$, it holds*

$$\|\Delta\|_2^2/\rho \le \|\Delta\|_{\mathbf{H}}^2 \le \varkappa\|\Delta\|_2^2.$$

Then, whenever

$$n \gtrsim \rho\varkappa K^4 \mathtt{s}\log(ed/\delta),$$

it holds that with probability $\ge 1-\delta$, for any $\Delta \in \mathbb{R}^d$ satisfying the RE condition,

$$\|\Delta\|_{\mathbf{H}}^2/2 \le \|\Delta\|_{\mathbf{H}_n}^2 \le 2\|\Delta\|_{\mathbf{H}}^2.$$

B Technical results for self-concordant (-like) functions

Here we summarizing the technical results related to self-concordant-like functions. These results are used later on to control the population and empirical risks $L(\theta)$, $L_n(\theta)$ in the proofs in Appendix C. In what follows, we fix $\theta_0, \theta_1 \in \Theta$, and let $\theta_t := \theta_0 + t(\theta_1 - \theta_0)$, $t \in [0,1]$. We define functions $\phi(\cdot)$, $\phi_n(\cdot)$ on $[0,1]$ by

$$\phi(t) := L(\theta_t), \quad \phi_n(t) := L_n(\theta_t). \tag{53}$$

The next result follows from the assumptions of Section 2.1 via the chain rule.

Proposition B.1. *Suppose that $\ell_z(\cdot)$ is convex, and $\ell'''_z(\cdot)$ exists on Θ.*

(a) *If Assumption SCa is satisfied, then for any $t \in [0,1]$, one has*

$$|\phi'''_n(t)| \le \phi''_n(t) \max_{i\in[n]} |\langle X_i, \theta_1 - \theta_0\rangle|, \tag{54}$$

$$|\phi'''(t)| \le \phi''(t) \sup_{x\in\mathcal{X}} |\langle x, \theta_1 - \theta_0\rangle|. \tag{55}$$

(b) *If Assumption SCb is satisfied instead, then for any $t \in [0,1]$, one has*

$$|\phi'''_n(t)| \le \phi''_n(t) \big[\max_{i\in[n]} \phi''_{Z_i}(t)\big]^{1/2}, \tag{56}$$

$$|\phi'''(t)| \le \phi''(t) \big[\sup_{z\in\mathcal{Z}} \phi''_z(t)\big]^{1/2}. \tag{57}$$

Proof. Recall that $\theta_t = \theta_0 + t(\theta_1 - \theta_0)$ for $t \in [0,1]$, and denote $\Delta := \theta_1 - \theta_0$. Differentiating under the expectation, we obtain that the derivatives of $\phi(t) = L(\theta_t)$ and $\phi_n(t) = L_n(\theta_t)$ are given by

$$\phi^{(p)}(t) = \mathbb{E}[\ell^{(p)}(Y, \langle X, \theta_t\rangle)\langle X, \Delta\rangle^p], \tag{58}$$

$$\phi_n^{(p)}(t) = \frac{1}{n}\sum_{i\in[n]} \ell^{(p)}(Y_i, \langle X_i, \theta_t\rangle)\langle X_i, \Delta\rangle^p. \tag{59}$$

This holds for $p \le 3$ due to the basic smoothness assumption. Now, let Assumption SCa be satisfied. Using (58) with $p \in \{2,3\}$, we get

$$|\phi'''(t)| \le \mathbb{E}[|\ell'''(Y, \langle X, \theta_t\rangle)||\langle X, \Delta\rangle|^3] \le \mathbb{E}[\ell''(Y, \langle X, \theta_t\rangle)\langle X, \Delta\rangle^2] \sup_{x\in\mathcal{X}} |\langle x, \Delta\rangle|,$$

arriving at (55). Analogously, we obtain (54) from (59), replacing $\mathcal{X}$ with the set $\{X_1, ..., X_n\}$. On the other hand, if Assumption SCb is satisfied instead,

$$\begin{aligned}|\phi'''(t)| &\le \mathbb{E}[|\ell'''(Y, \langle X, \theta_t\rangle)||\langle X, \Delta\rangle|^3] \le \mathbb{E}[\ell''(Y, \langle X, \theta_t\rangle)^{3/2}|\langle X, \Delta\rangle|^3] \\ &\le \mathbb{E}[\ell''(Y, \langle X, \theta_t\rangle)\langle X, \Delta\rangle^2] \sup_{x,y\in\mathcal{X}\times\mathcal{Y}} \Big\{\sqrt{\ell''(y, \langle x, \theta_t\rangle)}|\langle x, \Delta\rangle|\Big\},\end{aligned}$$

implying (57). We obtain (56) by replacing $\mathbb{E}[\cdot]$ with sample averaging. ∎

The next proposition, whose proof follows [Nes13], allows to control the second derivative of the loss when it is restricted to a straight line.

Proposition B.2. *Suppose $g : \mathbb{R} \to \mathbb{R}$ is differentiable, non-negative, and*

$$|g'(t)| \le 2c[g(t)]^{3/2}, \qquad \forall t \in \mathbb{R}^{(+)} : c|t|\sqrt{g(0)} \le 1$$

for some $c \ge 0$. Then, for any $t \in \mathbb{R}^{(+)}$ such that $c|t|\sqrt{g(0)} \le 1$, it holds

$$\frac{g(0)}{(1 + c|t|\sqrt{g(0)})^2} \le g(t) \le \frac{g(0)}{(1 - c|t|\sqrt{g(0)})^2}.$$

Proof. We first treat the case $g(0) > 0$. Consider the segment

$$T_0 = \big[-1/c\sqrt{g(0)}, 1/c\sqrt{g(0)} \big],$$

and assume that $g(t) > 0$ on the whole T_0. Then, we can define $\psi(t) := 1/\sqrt{g(t)}$ on T_0, and the premise of the proposition translates to $|\psi'(t)| \leq c$. Now, let $t \in T_0$ be positive without loss of generality. Integrating from 0 to t, we get

$$-ct \leq 1/\sqrt{g(t)} - 1/\sqrt{g(0)} \leq ct.$$

Multiplying by the product $\sqrt{g(t)g(0)} > 0$, and rearranging the terms, we prove the claim in the case where $g(t)$ does not vanish on T_0 (the case of negative t is treated analogously). Now, let $t_0 \in T_0$ be the leftmost zero of $g(t)$ on $T_0 \cup \mathbb{R}^+$ (recall that we still assume $g(0) > 0$). Then the preceding argument is valid for any $t \in [0, t_0]$, which implies that $g(t_0) > 0$, thus yielding a contradiction. This argument can be repeated for negative t, taking t_0 to be the rightmost negative zero of $g(t)$ on T_0. Hence, $g(0) > 0$ in fact implies that $g(t) > 0$ on the whole T_0.

Finally, assume that $g(0) = 0$. Then if $g(t) \equiv 0$ on the whole T_0, we are done. Otherwise, there is a point $t' \in T_0$ in which $g(t') > 0$. W.l.o.g. assume that $t' > 0$, let t_0 be the rightmost zero of $g(t)$ on $T_0 \cup \mathbb{R}^+$, and take a pair of points $t_1, t_2 \in T_0$ such that $t_0 < t_1 < t_2$. Integrating $\psi'(t)$ from t_1 to t_2, we get

$$-c(t_2 - t_1) \leq 1/\sqrt{g(t_2)} - 1/\sqrt{g(t_1)} \leq c(t_2 - t_1),$$

which, after the mutiplication by $\sqrt{g(t_1)g(t_2)}$ and rearrangement, results in

$$g(t_1) \geq \frac{g(t_2)}{1 + (t_2 - t_1)\sqrt{g(t_2)}}.$$

When $t_1 \to t_0$, by continuity of $g(t)$ we get a contradiction with $g(t_0) = 0$. ∎

The next proposition describes the local properties of multivariate functions whose restrictions to line segments behave as pseudo self-concordant functions (Case **(a)**), or similarly but with a weaker control of the third derivative (Case **(b)**). Case **(a)** repeats [Bac10, Proposition 1], and suffices for pseudo self-concordant losses; Case **(b)** allows to treat canonically self-concordant losses.

Proposition B.3. *Let $F : \Theta \to \mathbb{R}$ be a convex C^3-mapping, fix $\theta_0, \theta_1 \in \Theta$, and let $\phi_F(t) := F(\theta_t)$, $\theta_t := \theta_0 + t(\theta_1 - \theta_0)$. Assume that $\mathbf{H}_0 := \nabla^2 F(\theta_0) \succ 0$. Finally, for some $W \in \mathbb{R}^d$, define*

$$S := |\langle W, \theta_1 - \theta_0 \rangle|.$$

(a) *[Bac10, Proposition 1]. Suppose that $\phi_F(t)$ satisfies*

$$|\phi_F'''(t)| \leq S\phi_F''(t), \quad 0 \leq t \leq 1, \quad \textit{then,}$$

$$F(\theta_1) - F(\theta_0) - \nabla F(\theta_0)^\top(\theta_1 - \theta_0) \leq \tfrac{e^S - S - 1}{S^2}\|\theta_1 - \theta_0\|^2_{\mathbf{H}_0}, \tag{60}$$

$$F(\theta_1) - F(\theta_0) - \nabla F(\theta_0)^\top(\theta_1 - \theta_0) \geq \tfrac{e^{-S} + S - 1}{S^2}\|\theta_1 - \theta_0\|^2_{\mathbf{H}_0}. \tag{61}$$

(b) *Suppose that $\theta_{1/S} \in \Theta$, and $\phi_F(t)$ satisfies, instead,*

$$|\phi_F'''(t)| \leq \tfrac{S}{1 - St}\phi_F''(t), \quad 0 \leq t < 1/S. \quad \textit{Then}$$

$$\tfrac{1}{3S^2}\|\theta_1 - \theta_0\|^2_{\mathbf{H}_0} \leq F(\theta_{1/S}) - F(\theta_0) - \tfrac{1}{S}\nabla F(\theta_0)^\top(\theta_1 - \theta_0) \leq \tfrac{1}{S^2}\|\theta_1 - \theta_0\|^2_{\mathbf{H}_0}. \tag{62}$$

Moreover, if $S < 1$, we have

$$\tfrac{1}{2+S}\|\theta_1 - \theta_0\|^2_{\mathbf{H}_0} \leq F(\theta_1) - F(\theta_0) - \nabla F(\theta_0)^\top(\theta_1 - \theta_0) \leq \tfrac{1}{2-S}\|\theta_1 - \theta_0\|^2_{\mathbf{H}_0}. \tag{63}$$

Proof. We first treat the one-dimensional case, extending Proposition B.2.

Lemma B.1 (Lemma 1 in [Bac10])**.** *Let $g : [0,1] \to \mathbb{R}$ be a three times differentiable and convex function such that $g''(0) > 0$, and for some $S \geq 0$,*

$$|g'''(t)| \leq S g''(t), \quad 0 \leq t \leq 1.$$

Then, for any $0 \leq t \leq 1$, one has

$$\tfrac{e^{-St}+St-1}{S^2} g''(0) \leq g(t) - g(0) - g'(0)t \leq \tfrac{e^{St}-St-1}{S^2} g''(0), \quad 0 \leq t \leq 1. \tag{64}$$

Proof. First assume that $g''(t) > 0$ on $[0,1]$. Then, the premise of the lemma implies that $-S\mathrm{d}t \leq \mathrm{d}\log g''(t) \leq S\mathrm{d}t$ for $0 \leq t \leq 1$. Integrating this, we get

$$g''(0)e^{-St} \leq g''(t) \leq g''(0)e^{St}. \tag{65}$$

Two more integrations successively give

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

and then (64). Now, let $t_0 \in (0,1]$ be the leftmost zero of $g''(t)$. Then, the preceding argument can be applied on $[0, t_0]$, yielding a contradiction due to the left inequality in (65). ∎

Lemma B.2. *Let $g : [0,1] \to \mathbb{R}$ be a three times differentiable and convex function such that $g''(0) > 0$, and for some $S \geq 0$,*

$$|g'''(t)| \leq \tfrac{S}{1-t} g''(t), \quad 0 \leq t < 1.$$

Then, for any $0 \leq t \leq 1$, one has

$$\begin{aligned} g(t) - g(0) - g'(0)t &\geq \tfrac{(1-t)^{2+S}+(2+S)t-1}{(1+S)(2+S)} g''(0), \\ g(t) - g(0) - g'(0)t &\leq \tfrac{(1-t)^{2-S}+(2-S)t-1}{(1-S)(2-S)} g''(0), \end{aligned} \tag{66}$$

where the upper bound holds whenever $S < 1$ for any $t \in [0,1)$, and whenever $S < 2$ when $t = 1$. In particular, taking $t = 1$, we have

$$\tfrac{1}{2+S} g''(0) \leq g(1) - g(0) - g'(0) \leq \tfrac{1}{2-S} g''(0).$$

Proof. W.l.o.g., we assume $g''(t) > 0$; the general case can be treated as in Lemma B.1. We follow the same steps as in Lemma B.1: after the first integration,

$$(1-t)^S g''(0) \leq g''(t) \leq (1-t)^{-S} g''(0). \tag{67}$$

Integrating two more times, and assuming $S < 1$ for the upper bound, we get

$$\tfrac{1-(1-t)^{1+S}}{1+S} g''(0) \leq g'(t) - g'(0) \leq \tfrac{1-(1-t)^{1-S}}{1-S} g''(0)$$

and then (66). When $t = 1$, the term $(1-S)$ vanishes from the denominator of the right-hand side of (66), hence in this case we can take $S < 2$. ∎

Lemma B.3. *Let $g : [0,1] \to \mathbb{R}$ be a three times differentiable and convex function such that $g''(0) > 0$, and for some $S \geq 0$,*

$$|g'''(t)| \leq \tfrac{S}{1-St} g''(t), \quad 0 \leq t < 1/S.$$

Then, for any $0 \leq t \leq 1/S$, one has

$$\left(\tfrac{t^2}{2} - \tfrac{St^3}{6}\right) g''(0) \leq g(t) - g(0) - g'(0)t \leq \tfrac{St+(1-St)\log(1-St)}{S^2} g''(0). \tag{68}$$

In particular, taking $t = 1/S$, we have

$$\frac{g''(0)}{3S^2} \leq g(1/S) - g(0) - \frac{g'(0)}{S} \leq \frac{g''(0)}{S^2}.$$

Proof. We again assume w.l.o.g. that $g''(t) > 0$. Integrating three times, we get

$$(1 - St)g''(0) \le g''(t) \le \tfrac{1}{1-St} g''(0),$$

then

$$\left(t - \tfrac{St^2}{2}\right) g''(0) \le g'(t) - g'(0) \le \left(-\tfrac{\log(1-St)}{S}\right) g''(0), \tag{69}$$

implying (68). The last claim follows by continuity of $f(u) = u \log u$ at 0. ■

Proof of the proposition Case **(a)**, the first statement of Case **(b)**, and the second statement of Case **(b)** follow, correspondingly, from Lemmas B.1, B.3, and B.2 applied to $g(t) = \phi_F(t)$ and using that $g(t) = F(\theta_t)$, $g'(0) = \langle F'(\theta_0), \theta_1 - \theta_0 \rangle$, and $g''(0) = \|\theta_1 - \theta_0\|^2_{\mathbf{H}_0}$. Note that the inner-product structure of S does not play a role here, but is used in Proposition B.4. ■

The next result describes the behavior of (pseudo) self-concordant functions close to the optimum. Case **(a)** corresponds to [Bac10, Proposition 2]. The argument for Case **(b)** appears to be new, and is of independent interest. We note that a very similar argument was independently invented by U. Marteau-Ferey in [MFOBR19].

Proposition B.4. *Suppose that one of the Cases* **(a)**–**(b)** *in Proposition B.3 holds with fixed θ_0, all $\theta_1 \in \Theta$, and $W \in \mathbb{R}^d$ which can depend on θ_1. Whenever*

$$\|W\|_{\mathbf{H}_0^{-1}} \|\nabla F(\theta_0)\|_{\mathbf{H}_0^{-1}} \le 1/4,$$

function $F(\theta)$ has a unique minimizer $\tilde\theta \in \Theta$, and $\|\tilde\theta - \theta_0\|_{\mathbf{H}_0} \le 4\|\nabla F(\theta_0)\|_{\mathbf{H}_0^{-1}}$.

The key message of Proposition B.4 is that the *local* information about $F(\cdot)$ at one point efficiently amounts to the *global* information about how close is this point to the optimum. When applied to the *empirical risk* with $\theta_0 = \theta_*$ and $\tilde\theta = \widehat\theta_n$, this proposition allows to localize $\widehat\theta_n$ using that the quantity $\|\nabla L_n(\theta_*)\|^2_{\mathbf{H}^{-1}}$ decreases at rate $O(d_{\text{eff}}/n)$ under the i.i.d. assumption.

Proof. Note that from (65), (67), or (69), depending on the case, it follows that $\nabla^2 F(\theta) \succ 0$ for any $\theta \in \Theta$, hence the minimum $\tilde\theta$ is unique provided that it exists. Now, consider the level set

$$\Theta_F(F(\theta_0)) := \{\theta \in \Theta : F(\theta) \le F(\theta_0)\}.$$

Let $\theta_1 \in \Theta_F(F(\theta_0))$ be arbitrary, and $r = \|\theta_1 - \theta_0\|_{\mathbf{H}_0}$. Denote $\nu := \|\nabla F(\theta_0)\|_{\mathbf{H}_0^{-1}}$ and $R := \|W\|_{\mathbf{H}_0^{-1}}$; note that $S \le Rr$. We now treat all cases of Proposition B.3.

Case (a). By (61), we have

$$F(\theta_1) \ge F(\theta_0) + \langle \nabla F(\theta_0), \theta_1 - \theta_0 \rangle + \tfrac{e^{-Rr} - 1 + Rr}{R^2 r^2} r^2 \ge F(\theta_0) - \nu r + \tfrac{e^{-Rr} - 1 + Rr}{R^2},$$

where we first used that $u \mapsto (e^{-u} - 1 + u)/u$ is a decreasing function, and then the Cauchy-Schwarz inequality. Denoting $u = Rr$, we arrive at

$$e^{-u} - 1 + u \le \nu R u. \tag{70}$$

By the premise, we know that $\nu R \le 1/2$, hence $e^{-u} - 1 + u/2 \le 0$. We can check numerically that this implies $u \le 2$; moreover, one has $e^{-u} - 1 + u \ge u^2/4$ for such u. Plugging this back into (70), we arrive at $u \le 4\nu R$, that is, $\|\theta_1 - \theta_0\|_{\mathbf{H}_0} \le 4\nu$. In other words, the level set $\Theta_F(F(\theta_0))$ is compact and belongs to the $\|\cdot\|_{\mathbf{H}_0}$-ball of radius 4ν centered at θ_0. Hence, the minimum $\tilde\theta$ exists and belongs to the same ball; it is also unique since $F(\theta) \succ 0$.

Case (b) with $S < 1$. By the lower bound in (63), we have

$$F(\theta_1) \geq F(\theta_0) + \langle \nabla F(\theta_0), \theta_1 - \theta_0 \rangle + \tfrac{1}{2+Rr} r^2 \geq F(\theta_0) - \nu r + \tfrac{1}{2+Rr} r^2,$$

where we used that $u \mapsto 1/(2+u)$ is a decreasing function on $\mathbb{R}^+$. Whence,

$$\frac{u}{u+2} \leq \nu R,$$

where $u := Rr$. Since $\nu R \leq 1/2$, we have $u \leq 2$. Thus, we get $r \leq 4\nu$ as required.

Case (b) with arbitrary $S \geq 0$. First assume that $Rr \geq S \geq 1$. Then, $\theta_{1/S}$ belongs to the segment $[\theta_0, \theta_1]$ and to Θ. Whence $F(\theta_{1/S}) \leq F(\theta_0)$ by convexity of $\Theta_F(F(\theta_0))$. On the other hand, from the lower bound in (62) we have

$$F(\theta_{1/S}) \geq F(\theta_0) - \frac{\nu r}{S} + \frac{r^2}{3S^2}.$$

Whence $\nu \geq \frac{r}{3S} \geq \frac{1}{3R}$, and we arrive at the contradiction. Thus, the only possibility is that $S < 1$, in which case the statement has already been proved. ■

B.1 Properties of pseudo-Huber loss (21)

We can check that the Fenchel dual of $\phi : (-1, 1) \to \mathbb{R}$ defined in (20) is indeed $\varphi(t)$, cf. (21), by solving a quadratic equation. Since ϕ is a barrier on $(-1, 1)$, we have $|\varphi'(t)| < 1$ for any $t \in \mathbb{R}$. Now, we have $\phi'(\varphi'(t)) = t$ for $t \in \mathbb{R}$, see, e.g., [Roc70]. Differentiating this identity, we obtain

$$\phi''(\varphi'(t)) \cdot \varphi''(t) = 1. \tag{71}$$

Clearly, the Fenchel dual of an even function is also even, hence $\varphi'(0) = 0$, and $\varphi''(0) = 1/\phi''(0)$. Differentiating once again, we get

$$\phi'''(\varphi'(t)) \cdot [\varphi''(t)]^2 + \phi''(\varphi'(t)) \cdot \varphi'''(t) = 0,$$

whence, using that $\phi''(u) > 0$ for any $u \in (-1, 1)$,

$$|\varphi'''(t)| = \frac{|\phi'''(\varphi'(t))|}{\phi''(\varphi'(t))} [\varphi''(t)]^2.$$

Whence, if $|\phi'''(u)| \leq c[\phi''(u)]^{3/2}$, we get that $|\phi'''(u)| \leq c[\phi''(u)]^{3/2}$ via (71). ■

C Proofs of theorems

C.1 Proof of Theorem 3.1

1°. Recall that $\mathbf{H} = \nabla^2 L(\theta_*)$, and let $\mathbf{H}_n := \nabla^2 L_n(\theta_*)$. Note that due to Assumption D2 and the first bound on n in the premise of the theorem, we can apply Theorem A.2 to $\mathbf{H}_n$ and $\mathbf{H}$. Thus, with probability at least $1 - \delta$ we have

$$\tfrac{1}{2}\mathbf{H} \preccurlyeq \mathbf{H}_n \preccurlyeq 2\mathbf{H}. \tag{72}$$

On the other hand, we can prove (28) using Assumption D1. Indeed, the vectors

$$\nabla \ell_{Z_i}(\theta_*) = \ell'(Y_i, X_i^\top \theta_*) X_i, \quad i \in [n],$$

are independent, zero mean and with covariance $\mathbf{G}$. Hence, the random vectors $\mathbf{G}^{-1/2}\nabla\ell_{Z_i}(\theta_*)$. $i \in [n]$, are independent and isotropic (have zero mean and unit covariance). Moreover, by Assumption D1, $\|\mathbf{G}^{-1/2}\nabla\ell_{Z_i}(\theta_*)\|_{\psi_2} \le K_1$. Hence, by Lemma A.4 about the subgaussian norm of the sum of i.i.d. random vectors, we have that the random vector $V_n := \sqrt{n}\mathbf{G}^{-1/2}\nabla L_n(\theta_*)$, is isotropic, satisfies $\|V_n\|_{\psi_2} \lesssim K_1$, and, moreover,

$$\|\nabla L_n(\theta_*)\|^2_{\mathbf{H}^{-1}} = \tfrac{1}{n}\|V_n\|^2_{\mathbf{J}} \;\text{ with }\; \mathbf{J} := \mathbf{G}^{1/2}\mathbf{H}^{-1}\mathbf{G}^{1/2}. \tag{73}$$

Using that $\|\mathbf{J}\|_\infty \le \|\mathbf{J}\|_2 \le \mathrm{Tr}(\mathbf{J}) = d_{\mathrm{eff}}$, by Theorem A.1, we arrive at (28).

2°. Our next goal is proving (29). Let $\mu := \mathbb{E}[X]$ and $\mathbf{\Sigma}_o := \mathbb{E}[(X-\mu)(X-\mu)^\top]$ so that $\mathbf{\Sigma} = \mathbf{\Sigma}_o + \mu\mu^\top$. Denoting $\mathbf{Q} = \mathbf{\Sigma}_o^{1/2}\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_o^{1/2}$, we have

$$\begin{aligned}\|X_i\|^2_{\mathbf{\Sigma}^{-1}} &= \|X_i-\mu\|^2_{\mathbf{\Sigma}^{-1}} + 2\langle \mathbf{\Sigma}^{-1/2}\mu, \mathbf{\Sigma}^{-1/2}(X_i-\mu)\rangle + \|\mu\|^2_{\mathbf{\Sigma}^{-1}} \\ &= \|\mathbf{\Sigma}_o^{-1/2}(X_i-\mu)\|^2_{\mathbf{Q}} + 2\langle \mathbf{Q}^{1/2}\mathbf{\Sigma}_o^{-1/2}\mu, \mathbf{Q}^{1/2}\mathbf{\Sigma}_o^{-1/2}(X_i-\mu)\rangle + \|\mathbf{\Sigma}^{-1/2}\mu\|_2^2.\end{aligned} \tag{74}$$

By construction, $\mathbf{\Sigma}_o^{-1/2}(X_i-\mu)$ is isotropic. Moreover, $\|\mathbf{\Sigma}_o^{-1/2}(X_i-\mu)\|_{\psi_2} \lesssim K_0$ due to Assumption D0 and Lemma A.5. Note that $\|\mathbf{Q}\|_2 \le \mathrm{Tr}(\mathbf{Q}) \le d$ and $\|\mathbf{Q}\|_\infty \le 1$. Hence, by Theorem A.1, with probability at least $1-\delta$ one has

$$\|\mathbf{\Sigma}_o^{-1/2}(X_i-\mu)\|^2_{\mathbf{Q}} \lesssim K_0^2 d\big[\sqrt{\log(e/\delta)} + \log(1/\delta)\big] \lesssim K_0^2 d\log(e/\delta)\,.$$

Now, the second term in the right-hand side of (74) can be controlled as follows:

$$\begin{aligned}|\langle \mathbf{Q}^{1/2}\mathbf{\Sigma}_o^{-1/2}\mu, \mathbf{Q}^{1/2}\mathbf{\Sigma}_o^{-1/2}(X_i-\mu)\rangle| &\le \|\mathbf{Q}\|_\infty^{1/2}\|\mathbf{Q}^{1/2}\mathbf{\Sigma}_o^{-1/2}\mu\|_2\|\mathbf{\Sigma}_o^{-1/2}(X_i-\mu)\|_2 \\ &= \|\mathbf{Q}\|_\infty^{1/2}\|\mathbf{\Sigma}^{-1/2}\mu\|_2\|\mathbf{\Sigma}_o^{-1/2}(X_i-\mu)\|_2 \\ &\le \|\mathbf{\Sigma}^{-1/2}\mu\|_2\|\mathbf{\Sigma}_o^{-1/2}(X_i-\mu)\|_2 \\ &\lesssim K_0\sqrt{d\log(e/\delta)}\|\mathbf{\Sigma}^{-1/2}\mu\|_2,\end{aligned}$$

where the last inequality holds with probability $\ge 1-\delta$ by Corollary A.1. Finally,

$$\|\mathbf{\Sigma}^{-1/2}\mu\|_2^2 \le \mu^\top\mathbf{\Sigma}^{-1}\mu = \mu^\top(\mathbf{\Sigma}_o+\mu\mu^\top)^{-1}\mu \le 1.$$

Combining these results with the union bound, (72), and Assumption C, we have

$$\max_{i\in[n]}\|X_i\|^2_{\mathbf{H}_n^{-1}} \lesssim \rho K_0^2 d\log(en/\delta)\,, \quad \forall i\in[n] \tag{75}$$

with probability $\ge 1-\delta$. Now, (28), (72), and the 2nd bound in (27) imply that

$$\max_{i\in[n]}\|X_i\|^2_{\mathbf{H}_n^{-1}}\|\nabla L_n(\theta_*)\|^2_{\mathbf{H}_n^{-1}} \le c. \tag{76}$$

Now, putting $c = 1/4$, this results in (29). Indeed, invoking the bound (54) of Proposition B.1, we see that $L_n(\cdot)$ falls into Case (**a**) of Proposition B.3 with $\theta_0 = \theta_*$, $\mathbf{H}_0 = \mathbf{H}_n$, and $W(\theta) = X_{j(\theta)}$ for $j(\theta) \in \mathrm{Argmax}_{i\in[n]}|\langle X_i, \theta-\theta_*\rangle|$. Hence, we can apply Proposition B.4: clearly, $\|W(\theta)\|_{\mathbf{H}_n^{-1}} \le \max_{i\in[n]}\|X_i\|_{\mathbf{H}_n^{-1}}$ for all $\theta\in\Theta$, and then (76) with $c=1/4$ implies that the minimizer $\widehat{\theta}_n$ of $\widehat{L}_n(\cdot)$ is unique and satisfies $\|\widehat{\theta}-\theta_*\|^2_{\mathbf{H}_n} \le 4\|\nabla L_n(\theta_*)\|^2_{\mathbf{H}_n^{-1}}$. By (72), this results in (29).

3°. Let us now prove (30). To this end, consider the restricted risk $L_{\mathcal{E}_0}(\theta)$, fix two arbitrary points $\theta_0, \theta_1 \in \Theta$, and consider function $\phi_{\mathcal{E}_0}(t) := L_{\mathcal{E}_0}(\theta_t)$ where $\theta_t = \theta_0 + t(\theta_1-\theta_0)$ for $t\in[0,1]$. Differentiating $\phi_{\mathcal{E}_0}(t)$ three times (note that $\mathcal{E}_0$ does not depend on θ), we see that (55) can now be replaced with

$$|\phi'''_{\mathcal{E}_0}(t)| \le \phi''_{\mathcal{E}_0}(t)\sup_{x\in\mathcal{X}_{\mathcal{E}_0}}|\langle x, \theta_1-\theta_0\rangle|,$$

where $\mathcal{X}_{\mathcal{E}_0} := \{x \in \mathcal{X} : \|x\|_{\mathbf{H}^{-1}} \leq \sqrt{\rho}\mathfrak{B}_0\}$, with $\mathfrak{B}_0 := K_0\sqrt{d\log(e/\delta)}$, is the $(1-\delta)$-confidence set (under $\mathcal{E}_0$) for X. (We used Assumption C.) Besides, let us momentarily assume that the new Hessian $\mathbf{H}_{\mathcal{E}_0} := \nabla^2 L_{\mathcal{E}_0}(\theta_*)$ is invertible, and approximates $\mathbf{H}$ in the positive-semidefinite sense: for some constants $c, C > 0$,

$$c\mathbf{H} \preccurlyeq \mathbf{H}_{\mathcal{E}_0} \preccurlyeq C\mathbf{H}. \tag{77}$$

Later on, we will verify this under condition (31) on δ. Now, under (77), we can apply Case **(a)** of Proposition B.3 to $L_{\mathcal{E}}(\cdot)$ with $\theta_0 = \theta_*$, $\theta_1 = \widehat{\theta}_n$, $\mathbf{H}_0 = \mathbf{H}_{\mathcal{E}_0}$, and $W = W(\theta) \in \mathrm{Argmax}_{x\in\mathcal{X}_{\mathcal{E}_0}} |\langle x, \theta - \theta_*\rangle|$. Observe that $\|W(\theta)\|_{\mathbf{H}^{-1}} \leq \sqrt{\rho}\mathfrak{B}_0$, and let $r := \|\widehat{\theta}_n - \theta_*\|_{\mathbf{H}}^2$. By (60) combined with the Cauchy-Schwarz inequality,

$$L_{\mathcal{E}_0}(\widehat{\theta}_n) - L_{\mathcal{E}_0}(\theta_*) \lesssim \left(\frac{e^{\sqrt{\rho}\mathfrak{B}_0 r} - 1 - \sqrt{\rho}\mathfrak{B}_0 r}{\rho\mathfrak{B}_0^2 r^2}\right) r^2 + \nabla L_{\mathcal{E}_0}(\theta_*)^\top(\widehat{\theta}_n - \theta_*).$$

Now, observe that the term in the parentheses is at most a constant. Indeed, $\sqrt{\rho}\mathfrak{B}_0 r \lesssim 1$ follows from the combination of (27)–(29), and then $f(u) = e^u - 1 - u \lesssim u^2$ whenever $u \lesssim 1$ (in particular, $f(u) \leq u^2$ when $u \leq 1$). Thus,

$$L_{\mathcal{E}_0}(\widehat{\theta}_n) - L_{\mathcal{E}_0}(\theta_*) \lesssim r^2 + r\|\nabla L_{\mathcal{E}_0}(\theta_*)\|_{\mathbf{H}^{-1}}. \tag{78}$$

In order to prove (30), it remains to control $\|\nabla L_{\mathcal{E}_0}(\theta_*)\|_{\mathbf{H}^{-1}}$ and to verify (77).

4^o. To estimate the additional term in (78), consider the *complementary risk:*

$$L_{\mathcal{E}_0^c}(\theta_*) := \mathbb{E}[\ell_Z(\theta_*)\mathbb{1}_{\mathcal{E}_0^c}(X)],$$

where $\mathcal{E}_0^c$ is the complement of $\mathcal{E}_0$, so that $\mathbb{P}(\mathcal{E}_0^c) \leq \delta$. Note that, since $\nabla L(\theta_*) = 0$, we have $\nabla L_{\mathcal{E}_0}(\theta_*) = -\nabla L_{\mathcal{E}_0^c}(\theta_*)$, whence

$$\|\nabla L_{\mathcal{E}_0}(\theta_*)\|_{\mathbf{H}^{-1}} = \|\nabla L_{\mathcal{E}_0^c}(\theta_*)\|_{\mathbf{H}^{-1}}.$$

We now estimate $\|\nabla L_{\mathcal{E}_0^c}(\theta_*)\|_{\mathbf{H}^{-1}}$ through a technique inspired by the one in [Ver11, Section 1.3]. For any p, q such that $1/p + 1/q = 1$, we have by Hölder's inequality:

$$\|\nabla L_{\mathcal{E}_0^c}(\theta_*)\|_{\mathbf{H}^{-1}} \leq \mathbb{E}[\|\nabla\ell_Z(\theta_*)\|_{\mathbf{H}^{-1}}\mathbb{1}_{\mathcal{E}_0^c}] \leq \mathbb{E}[\|\nabla\ell_Z(\theta_*)\|_{\mathbf{H}^{-1}}^p]^{1/p}\delta^{1/q}, \tag{79}$$

Note that

$$\|\nabla\ell_Z(\theta_*)\|_{\mathbf{H}^{-1}}^2 = \|\mathbf{G}^{-1/2}\nabla\ell_Z(\theta_*)\|_{\mathbf{J}}^2$$

where $\mathbf{J} = \mathbf{G}^{1/2}\mathbf{H}^{-1}\mathbf{G}^{1/2}$, and $\mathbf{G}^{-1/2}\nabla\ell_Z(\theta_*)$ is isotropic and satisfies

$$\|\mathbf{G}^{-1/2}\nabla\ell_Z(\theta_*)\|_{\psi_2} \leq K_1.$$

Hence, by Corollary A.2, $\zeta := \|\mathbf{H}^{-1/2}\nabla\ell_Z(\theta_*)\|$ satisfies $\|\zeta\|_{\psi_2} \lesssim K_1\sqrt{d_{\mathrm{eff}}}$. As such, we can bound the moments of ζ using Lemma A.1:

$$\mathbb{E}[\|\nabla\ell_Z(\theta_*)\|_{\mathbf{H}^{-1}}^p]^{1/p} \lesssim K_1\sqrt{pd_{\mathrm{eff}}}.$$

Combining this with (78)–(79) and (28)–(29), and choosing $p = \log(ed_{\mathrm{eff}})$ and $q = 1 + 1/\log(d_{\mathrm{eff}})$, we obtain

$$L_{\mathcal{E}_0}(\widehat{\theta}_n) - L_{\mathcal{E}_0}(\theta_*) \lesssim K_1^2\sqrt{\frac{d_{\mathrm{eff}}\log(e/\delta)}{n}}\left(\sqrt{\frac{d_{\mathrm{eff}}\log(e/\delta)}{n}} + \delta^{\frac{\log(d_{\mathrm{eff}})}{\log(d_{\mathrm{eff}})+1}}\sqrt{d_{\mathrm{eff}}\log(ed_{\mathrm{eff}})}\right).$$

Finally, (31) implies that $\delta^{\frac{\log(d_{\mathrm{eff}})}{\log(d_{\mathrm{eff}})+1}}\sqrt{\log(d_{\mathrm{eff}})} \lesssim \sqrt{\log(e/\delta)/n}$, and (30) follows.

$\mathbf{5^o}$. It remains to verify (77), i.e., that the Hessians $\mathbf{H}$ and $\mathbf{H}_{\mathcal{E}_0}$ are close. First, the upper bound in (77) is trivial. Indeed defining the complementary $\mathbf{H}_{\mathcal{E}_0^c} := \nabla^2 L_{\mathcal{E}_0^c}(\theta_*)$, we see that $\mathbf{H}_{\mathcal{E}_0} = \mathbf{H} - \mathbf{H}_{\mathcal{E}_0^c} \preccurlyeq \mathbf{H}$ since $\mathbf{H}_{\mathcal{E}_0^c} \succcurlyeq 0$. On the other hand, the lower bound in (77) with $c \in (0,1)$ would follow from the bound

$$\|\mathbf{H}^{-1/2}\mathbf{H}_{\mathcal{E}_0^c}\mathbf{H}^{-1/2}\|_\infty \le c',$$

where $c' \in (0,1)$. Let us show that this bound is satisfied under the second bound in (31), using a technique similar to the one used to control $\nabla L_{\mathcal{E}_0}(\theta_*)$. For any $p, q \ge 1$ such that $1/p + 1/q = 1$, we have by Hölder's and Young's inequalities:

$$\begin{aligned}
\|\mathbf{H}^{-1/2}\mathbf{H}_{\mathcal{E}_0^c}\mathbf{H}^{-1/2}\|_\infty &\le \mathbb{E}[\|\mathbf{H}^{-1/2}\nabla^2\ell_Z(\theta_*)\mathbf{H}^{-1/2}\|_\infty^p]^{1/p}\delta^{1/q} \\
&= \mathbb{E}[\|\mathbf{H}^{-1/2}\widetilde{X}\widetilde{X}^\top\mathbf{H}^{-1/2}\|_\infty^p]^{1/p}\delta^{1/q} \\
&= \mathbb{E}[\|\mathbf{H}^{-1/2}\widetilde{X}\|_2^{2p}]^{1/p}\delta^{1/q} \lesssim K_2^2 pd\delta^{1/q},
\end{aligned}$$

where in the end we used that $\zeta = \|\mathbf{H}^{-1/2}\widetilde{X}\|_2$ satisfies $\|\zeta\|_{\psi_2} \le K_2\sqrt{d}$ by Corollary A.2. Choosing $p = \log(ed)$, we see that $K_2^2 pd\delta^{1/q} \lesssim 1$ under (31). ■

C.2 Proof of Theorem 3.2

Disclaimer. The key distinction from Theorem 3.1 is the absence of curvature parameter ρ in the derived critical sample size (cf. (32) viz. (27)). This improvement is achieved by carefully exploiting Assumption SCb. In particular, we invoke Case (b), instead of Case (a), in Propositions B.1 and B.3. Meanwhile, the role of the bounding vector W is now relegated from X to $\widetilde{X} = \ell''(Y, X^\top\theta_*)^{1/2}X$.

The proof of the theorem below recycles results from the proof of Theorem 3.1.

Proof. We repeat step $\mathbf{1^o}$ in the previous proof *verbatim*, arriving at (28) and (72).

$\mathbf{2^o}$. In order to prove (29), we use Case (b) of Proposition B.1. To this end, fix arbitrary $\theta \in \Theta$, let $\theta_t = \theta_* + t(\theta - \theta_*)$ for $t \in [0,1)$, and define $\phi_z(t) := \ell_z(\theta_t)$ for arbitrary $z \in \mathcal{Z}$. Due to Assumption SCb, for any $z \in \mathcal{Z} = \mathbb{R}^d \times \mathcal{Y}$, we have $|\phi_z'''(t)| \le 2[\phi_z''(t)]^{3/2}$. Hence, we can apply Proposition B.2 to $g(t) = \phi_z''(t)$ with $c = 1$. Thus, with $\widetilde{x} := [\ell''(y, x^\top\theta_*)]^{1/2}x$ for arbitrary $(x,y) \in \mathcal{Z}$, we have

$$\phi_z''(t) \le \frac{\phi_z''(0)}{(1 - t\sqrt{\phi_z''(0)})^2} = \frac{\langle\widetilde{x}, \theta - \theta_*\rangle^2}{(1 - t|\langle\widetilde{x}, \theta - \theta_*\rangle|)^2} \tag{80}$$

for any $t \ge 0$ such that the denominator is non-zero. Combining this with (56),

$$|\phi_n'''(t)| \le \phi_n''(t)\max_{i\in[n]}\frac{|\langle\widetilde{X}_i, \theta - \theta_*\rangle|}{1 - t|\langle\widetilde{X}_i, \theta - \theta_*\rangle|} = \phi_n''(t)\frac{|\langle\widetilde{X}_{j(\theta)}, \theta - \theta_*\rangle|}{1 - t|\langle\widetilde{X}_{j(\theta)}, \theta - \theta_*\rangle|}, \tag{81}$$

where $j(\theta) \in \operatorname{Argmax}_{i\in[n]}|\langle\widetilde{X}_i, \theta - \theta_*\rangle|$, and again we can take any $t \ge 0$ such that the denominator is positive. Thus, $L_n(\theta)$ falls into Case (b) of Proposition B.3 with $\theta_0 = \theta_*$, $\mathbf{H}_0 = \mathbf{H}_n$, and $W = W(\theta) = \widetilde{X}_{j(\theta)}$. On the other hand, repeating the analysis that led to (75), we obtain that, for any fixed θ,

$$\|\widetilde{X}_{j(\theta)}\|_{\mathbf{H}_n^{-1}}^2 \lesssim \mathfrak{B}_2^2 := K_2^2 d\log(en/\delta)$$

with probability $\ge 1 - \delta$. Combining this result with the second bound in (32),

$$\|\widetilde{X}_{j(\theta)}\|_{\mathbf{H}_n^{-1}}^2\|\nabla L_n(\theta_*)\|_{\mathbf{H}_n^{-1}}^2 \lesssim 1, \tag{82}$$

cf. (76). Hence, we can apply Proposition B.4 to $L_n(\theta)$ at $\theta_0 = \theta_*$, and repeating the final argument in step $\mathbf{2^o}$ of the proof of Theorem 3.1, we arrive at (29).

3°. We now prove (30) with $L_{\mathcal{E}_0}$ replaced by $L_{\mathcal{E}_2}$. Similarly to (81), from (80) and (57) we have

$$|\phi'''(t)| \le \phi''(t)\frac{|\langle W(\theta), \theta - \theta_*\rangle|}{1 - t|\langle W(\theta), \theta - \theta_*\rangle|}, \tag{83}$$

with probability $\ge 1-\delta$, where $W(\theta) \in \mathrm{Argmax}_{x\in\widetilde{\mathcal{X}}_{\mathcal{E}_2}} |\langle x, \theta - \theta_*\rangle|$ for the set

$$\widetilde{\mathcal{X}}_{\mathcal{E}_2} := \{\widetilde{x} = [\ell''(y, x^\top\theta_*)]^{1/2}x : \ \|\widetilde{x}\|^2_{\mathbf{H}^{-1}} \lesssim \mathfrak{B}_2^2\}.$$

(Clearly, $\widetilde{\mathcal{X}}_{\mathcal{E}_2}$ is the $(1-\delta)$-confidence set for the new observation $\widetilde{X}$.) Thus,

$$|\langle W(\theta), \widehat{\theta}_n - \theta_*\rangle| \le \mathfrak{B}_2 r, \quad r := \|\widehat{\theta}_n - \theta_*\|_{\mathbf{H}};$$

moreover, due to (28), (29), and the 2nd bound in (32) we have $\mathfrak{B}_2 r \lesssim 1$. As such, whenever $c\mathbf{H} \preccurlyeq \nabla^2 L_{\mathcal{E}_2}(\theta_*) \preccurlyeq C\mathbf{H}$, the restricted risk $L_{\mathcal{E}_2}(\cdot)$, cf. (26), falls under Case (**b**) of Proposition B.3 with $\theta_1 = \widehat{\theta}_n$ and $S < 1$; the upper bound in (63) then gives the analogue of (78):

$$L_{\mathcal{E}_2}(\widehat{\theta}_n) - L_{\mathcal{E}_2}(\theta_*) \lesssim \frac{r^2}{2 - \mathfrak{B}_2 r} + \nabla L_{\mathcal{E}_2}(\theta_*)^\top(\widehat{\theta}_n - \theta_*) \lesssim r^2 + r\|\nabla L_{\mathcal{E}_2}(\theta_*)\|_{\mathbf{H}^{-1}}.$$

It remains to estimate the right-hand side and to verify $c\mathbf{H} \preccurlyeq \nabla^2 L_{\mathcal{E}_2}(\theta_*) \preccurlyeq C\mathbf{H}$, using (31) in both cases. This repeats steps **4°**–**5°** in the proof of Theorem 3.1. ■

C.3 Proof of Theorem 4.1

1°. Without loss of generality, we assume that $\Theta = \mathbb{R}^d$; the argument can be extended to the general case simply by replacing all arising Dikin ellipsoids with their intersections with Θ. For simplicity, we also assume that Assumption D2* holds with $r = 1$, and denote $\bar{K}_2 := \bar{K}_2(1)$. First of all, for any $r \ge 0$ and $\theta \in \Theta_1(\theta_*)$, we define the Dikin ellipsoid with center θ and radius r:

$$\Theta_r(\theta) := \{\theta' \in \mathbb{R}^d : \|\theta' - \theta\|_{\mathbf{H}(\theta)} \le r\}.$$

We will prove that the Hessians $\mathbf{H}(\theta) := \nabla^2 L(\theta)$ are close to $\mathbf{H}(\theta_*)$ within the Dikin ellipsoid with radius $\Omega(1/\bar{K}_2^3)$. To this end, fix $\theta_0 = \theta_*$ and arbitrary $\theta_1 \in \mathbb{R}^d$, and let $\theta_t = \theta_0 + t(\theta_1 - \theta_0)$, $t \ge 0$. By using Assumptions SCb and D2*, we can prove that for the function $\phi(t) = L(\theta_t)$ it holds

$$\phi'''(t) \le 2\bar{c}[\phi''(t)]^{3/2}$$

for any $t \ge 0$ such that $\theta_t \in \Theta_{1/\bar{c}}(\theta_*)$ with $\bar{c} \gtrsim 1/\bar{K}_2^3$. Indeed, let $\Delta := \theta_1 - \theta_0$, and recall that

$$\phi^{(p)}(t) = \mathbb{E}[\ell^{(p)}(Y, \langle X, \theta_t\rangle)\langle X, \Delta\rangle^p], \quad p \in \{2, 3\},$$

cf. the proof of Proposition B.1. Putting $\widetilde{X}(\theta_t) := [\ell''(Y, \langle X, \theta_t\rangle)]^{1/2}X$, this gives

$$\phi''(t) = \mathbb{E}[\langle \widetilde{X}(\theta_t), \Delta\rangle^2] = \mathbb{E}[\langle \mathbf{H}(\theta_t)^{-1/2}\widetilde{X}(\theta_t), \mathbf{H}(\theta_t)^{1/2}\Delta\rangle^2] = \|\Delta\|^2_{\mathbf{H}(\theta_t)},$$

On the other hand, due to Assumption SCb,

$$\begin{aligned}
|\phi'''(t)| &\le \mathbb{E}[|\ell'''(Y, \langle X, \theta_t\rangle)| \cdot |\langle X, \Delta\rangle|^3] \\
&\le 2\mathbb{E}[|\langle [\ell''(Y, \langle X, \theta_t\rangle)]^{1/2}X, \Delta\rangle|^3] \\
&= 2\mathbb{E}[|\langle \mathbf{H}(\theta_t)^{-1/2}\widetilde{X}(\theta_t), \mathbf{H}(\theta_t)^{1/2}\Delta\rangle|^3].
\end{aligned}$$

Now, recall that whenever $\theta \in \Theta_c(\theta_*)$, one has $\|\mathbf{H}(\theta_t)^{-1/2}\widetilde{X}(\theta_t)\|_{\psi_2} \le \bar{K}_2$ due to Assumption D2*. Thus, for such θ_t we have

$$\|\langle \mathbf{H}(\theta_t)^{-1/2}\widetilde{X}(\theta_t), \mathbf{H}(\theta_t)^{1/2}\Delta\rangle\|_{\psi_2} \le \bar{K}_2\|\Delta\|_{\mathbf{H}(\theta_t)},$$

and by Lemma A.1,

$$\mathbb{E}[|\langle \mathbf{H}(\theta_t)^{-1/2}\widetilde{X}(\theta_t), \mathbf{H}(\theta_t)^{1/2}\Delta\rangle|^3] \le C\bar{K}_2^3\|\Delta\|^3_{\mathbf{H}(\theta_t)}$$

for some absolute constant $C > 0$. Without the loss of generality we can assume that $C \ge 1$. Combining the above inequalities, we observe that

$$|\phi'''(t)| \le 2C\bar{K}_2^3[\phi''(t)]^{3/2}, \quad 0 \le t[\phi''(0)]^{1/2} \le 1,$$

where we used that $\theta_t \in \Theta_1(\theta_*)$ is equivalent to $t^2\phi''(0) \le 1$. We can now apply Proposition B.2 to $g(t) = \phi''(t)$, putting

$$\bar{c} := C\bar{K}_2^3 \gtrsim 1,$$

and arriving at

$$\frac{\phi''(0)}{(1+\bar{c}t\sqrt{\phi''(0)})^2} \le \phi''(t) \le \frac{\phi''(0)}{(1-\bar{c}t\sqrt{\phi''(0)})^2}$$

whenever $0 \le \bar{c}t[\phi''(0)]^{1/2} \le 1$. Finally, since $\phi''(t) = \|\Delta\|^2_{\mathbf{H}(\theta_t)}$, this results in

$$\frac{\mathbf{H}(\theta_*)}{(1+\bar{c}\|\theta-\theta_*\|_{\mathbf{H}(\theta_*)})^2} \preccurlyeq \mathbf{H}(\theta) \preccurlyeq \frac{\mathbf{H}(\theta_*)}{(1-\bar{c}\|\theta-\theta_*\|_{\mathbf{H}(\theta_*)})^2},$$

whenever $\theta \in \Theta_{1/\bar{c}}(\theta_*)$. In particular, for any $\theta \in \Theta_{1/(2\bar{c})}(\theta_*)$ we have

$$\tfrac{4}{9}\mathbf{H}(\theta_*) \preccurlyeq \mathbf{H}(\theta) \preccurlyeq 4\mathbf{H}(\theta_*). \tag{84}$$

2°. Next, we derive a similar approximation result for the Hessian of *empirical* risk $\mathbf{H}_n(\theta) := \nabla^2 L_n(\theta)$. This can be done by constructing an epsilon-net on $\Theta_{1/(2\bar{c})}(\theta_*)$ with respect to the $\|\cdot\|_{\mathbf{H}(\theta_*)}$-norm. Then, one can control the uniform deviations of $\mathbf{H}_n(\theta)$ from $\mathbf{H}(\theta)$ for θ on the net, while approximating $\mathbf{H}_n(\theta)$ for θ outside the net, by exploiting the self-concordance of the *individual* losses, and appropriately choosing the net resolution. To this end, recall that $\mathbf{H}_n(\theta)$ writes

$$\mathbf{H}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell''(Y_i, X_i^\top\theta)X_iX_i^\top.$$

Hence, we can relate $\mathbf{H}_n(\theta)$ to $\mathbf{H}_n(\theta')$ at some other point θ' by relating $\ell''(Y_i, X_i^\top\theta)$ to $\ell''(Y_i, X_i^\top\theta')$. Namely, fix arbitrary $\theta_0 \in \Theta_{1/(2\bar{c})}(\theta_*)$ and $\theta_1 \in \Theta$, and observe that, by Assumption SCb, $\phi_Z(t) = \ell_Z(\theta_t)$ satisfies

$$|\phi_Z'''(t)| \le 2[\phi_Z''(t)]^{3/2},$$

hence we can apply Proposition B.2 to $\phi_Z''(t)$. For $0 \le t[\phi_Z''(0)]^{1/2} \le 1$ this gives

$$\frac{\phi_Z''(0)}{(1+t[\phi_Z''(0)]^{1/2})^2} \le \phi_Z''(t) \le \frac{\phi_Z''(0)}{(1-t[\phi_Z''(0)]^{1/2})^2}.$$

cf. (80). Recalling that $\phi_Z''(t) = \ell''(Y, X^\top\theta_t)\cdot\langle X,\Delta\rangle^2 = \langle\widetilde{X}(\theta_t),\Delta\rangle^2$, where again $\Delta = \theta_1 - \theta_0$ but now without the constraint that $\theta_0 = \theta_*$, we arrive at

$$\frac{\ell''(Y, X^\top\theta_0)}{(1+t|\langle\widetilde{X}(\theta_0),\Delta\rangle|)^2} \le \ell''(Y, X^\top\theta_t) \le \frac{\ell''(Y, X^\top\theta_0)}{(1-t|\langle\widetilde{X}(\theta_0),\Delta\rangle|)^2}$$

when $t|\langle \widetilde{X}(\theta_0), \Delta\rangle| \le 1$. By the Cauchy-Schwarz inequality and (84), this gives

$$\ell''(Y, X^\top \theta_t) \ge \frac{\ell''(Y, X^\top \theta_0)}{(1 + 2t\|\widetilde{X}(\theta_0)\|_{\mathbf{H}(\theta_0)^{-1}}\|\Delta\|_{\mathbf{H}(\theta_*)})^2}$$
$$\ell''(Y, X^\top \theta_t) \le \frac{\ell''(Y, X^\top \theta_0)}{(1 - 2t\|\widetilde{X}(\theta_0)\|_{\mathbf{H}(\theta_0)^{-1}}\|\Delta\|_{\mathbf{H}(\theta_*)})^2},$$

where $t \ge 0$ is such that the denominator is strictly positive. As a result, we have

$$\begin{aligned}\ell''(Y, X^\top \theta') &\ge \frac{\ell''(Y, X^\top \theta)}{(1 + 2\|\mathbf{H}(\theta)^{-1/2}\widetilde{X}(\theta)\|_2\|\theta' - \theta\|_{\mathbf{H}(\theta_*)})^2},\\ \ell''(Y, X^\top \theta') &\le \frac{\ell''(Y, X^\top \theta)}{(1 - 2\|\mathbf{H}(\theta)^{-1/2}\widetilde{X}(\theta)\|_2\|\theta' - \theta\|_{\mathbf{H}(\theta_*)})^2}\end{aligned} \tag{85}$$

for any $\theta \in \Theta_{1/(2\bar{c})}(\theta_*)$, and any θ' for which the denominator is strictly positive.

3°. Now, consider the smallest epsilon-net $\mathcal{N}_\varepsilon$ for $\Theta_{1/(2\bar{c})}(\theta_*)$ with respect to the norm $\|\cdot\|_{\mathbf{H}(\theta_*)}$, i.e., the smallest subset of $\Theta_{1/(2\bar{c})}(\theta_*)$ such that for any $\theta \in \Theta_{1/(2\bar{c})}(\theta_*)$ there exists a point $\theta' \in \mathcal{N}_\varepsilon$ such that $\|\theta' - \theta\|_{\mathbf{H}(\theta_*)} \le \varepsilon$. Note that such $\mathcal{N}_\varepsilon$ can be obtained as the affine image of the epsilon-net for the $\|\cdot\|_2$-ball with radius $1/(2\bar{c})$ with respect to the standard $\|\cdot\|_2$-norm. Hence, we can apply the bound for covering numbers of Euclidean balls: for any $\varepsilon \le 1$,

$$|\mathcal{N}_\varepsilon| \le \left(\frac{3}{2\bar{c}\varepsilon}\right)^d. \tag{86}$$

Consider random vectors $\mathbf{H}(\theta)^{-1/2}\widetilde{X}_i(\theta)$, where $i \in [n]$ and $\theta \in \mathcal{N}_\varepsilon$ for some ε to be defined later. Each of them has unit covariance matrix, and is subgaussian with ψ_2-norm at most $\bar{K}_2$ due to Assumption D2*. Repeating the argument from part **1°** of the proof of Theorem 3.1 (to account for the fact that the vectors are not centered), we can show that with probability at least $1 - \delta$,

$$\|\mathbf{H}(\theta)^{-1/2}\widetilde{X}_i(\theta)\|_2 \le C_2\bar{K}_2\sqrt{d\log(e/\delta)}$$

for some constant $C_2 \ge 1$. Here we used that $\mathcal{N}_\varepsilon \subset \Theta_{1/2\bar{c}}(\theta_*) \subseteq \Theta_1(\theta_*)$. Thus,

$$\sup_{i\in[n],\ \theta_0\in\mathcal{N}_\varepsilon}\|\mathbf{H}(\theta)^{-1/2}\widetilde{X}_i(\theta)\|_2 \le C_2\bar{K}_2\sqrt{d\log\left(\frac{en|\mathcal{N}_\varepsilon|}{\delta}\right)} \le C_2\bar{K}_2 d\sqrt{\log\left(\frac{3en}{\delta\varepsilon}\right)}, \tag{87}$$

with probability $\ge 1 - \delta$, where in the second step we used (86). Now, we choose

$$\varepsilon = \frac{1}{64C_2^2\bar{K}_2^2 d^2\log(en/\delta)}. \tag{88}$$

By some simple algebra, such choice of ε ensures that

$$\varepsilon\sqrt{\log\left(\frac{3en}{\delta\varepsilon}\right)} \le \frac{1}{4C_2\bar{K}_2 d}.$$

Combining this with (85) and (87), we see that the following is true with probability $\ge 1 - \delta$: for any $\theta' \in \Theta_{1/(2\bar{c})}(\theta_*)$, there exists $\theta \in \mathcal{N}_\varepsilon$ such that

$$\tfrac{4}{9}\ell''(Y_i, X_i^\top\theta) \le \ell''(Y_i, X_i^\top\theta') \le 4\ell''(Y_i, X_i^\top\theta), \quad i \in [n].$$

This implies that with probability $\geq 1-\delta$, it holds

$$\tfrac{4}{9}\mathbf{H}_n(\pi_*(\theta)) \preccurlyeq \mathbf{H}_n(\theta) \preccurlyeq 4\mathbf{H}_n(\pi_*(\theta)), \quad \forall \theta \in \Theta_{1/(2\bar{c})}(\theta_*), \tag{89}$$

where $\pi_*(\cdot)$ is the operation of $\|\cdot\|_{\mathbf{H}(\theta_*)}$-projection on the epsilon-net $\mathcal{N}_\varepsilon$. Finally, to establish the uniform approximation of $\mathbf{H}_n(\cdot)$ on $\Theta_{1/(2\bar{c})}(\theta_*)$, it remains to control $\mathbf{H}_n(\theta)$ on the net itself. This can be done by combining the deviation bounds for sample covariance matrices with the results of $\mathbf{1^o}$. First, by Theorem A.2, for any $\theta \in \mathcal{N}_\varepsilon$ we have that with probability at least $1-\delta$,

$$\tfrac{1}{2}\mathbf{H}(\theta) \preccurlyeq \mathbf{H}_n(\theta) \preccurlyeq 2\mathbf{H}(\theta),$$

provided that $n \gtrsim \bar{K}_2^4(d+\log(1/\delta))$. Taking the union bound over $\mathcal{N}_\varepsilon$, and using (86) and (88), we see that

$$\tfrac{1}{2}\mathbf{H}(\theta) \preccurlyeq \mathbf{H}_n(\theta) \preccurlyeq 2\mathbf{H}(\theta), \quad \forall \theta \in \mathcal{N}_\varepsilon \tag{90}$$

holds with probability $\geq 1-\delta$, provided that

$$n \gtrsim \bar{K}_2^4 d \log\left(\frac{e}{\bar{c}\delta\varepsilon}\right) \gtrsim \bar{K}_2^4 d\left[\log(ed/\delta) + \log\log(en/\delta)\right].$$

By simple algebra, it suffices that

$$n \gtrsim \bar{K}_2^4 d \log(e/\delta). \tag{91}$$

Combining (89), (90), and (84), we see that the sample size satisfying (91) guarantees uniform approximation of empirical Hessians on the Dikin ellipsoid $\Theta_{1/(2\bar{c})}(\theta_*)$: with probability $\geq 1-\delta$, for any $\theta \in \Theta_{1/(2\bar{c})}(\theta_*)$ it holds

$$0.09\,\mathbf{H}(\theta_*) \preccurlyeq \mathbf{H}_n(\theta) \preccurlyeq 32\,\mathbf{H}(\theta_*). \tag{92}$$

$\mathbf{4^o}$. With (92) at hand, we can localize the estimate through a similar argument to that in Proposition B.4, but with S replaced with a constant. Indeed, fixing $\theta_0 = \theta_*$ and taking arbitrary $\theta_1 \in \Theta_{1/(2\bar{c})}(\theta_*)$, we see that (92) reduces to

$$0.09\phi''(0) \leq \phi_n''(t) \leq 32\phi''(0), \quad 0 \leq t \leq 1.$$

Integrating this twice, we get $0.045\phi''(0)t^2 \leq \phi_n(t) - \phi_n(0) - \phi_n'(0)t \leq 16\phi''(0)t^2$. Putting $t=1$, and noting that $\phi''(0) = \|\theta_1 - \theta_*\|^2_{\mathbf{H}(\theta_*)}$, we obtain that for any $\theta \in \Theta_{1/(2\bar{c})}(\theta_*)$, with high probability it holds

$$0.045\|\theta-\theta_*\|^2_{\mathbf{H}(\theta_*)} \leq L_n(\theta) - L_n(\theta_*) - \langle \nabla L_n(\theta_*), \theta-\theta_*\rangle \leq 16\|\theta-\theta_*\|^2_{\mathbf{H}(\theta_*)}. \tag{93}$$

cf. (62). Now we can proceed as in the proof of Proposition B.4, Case **(b)**. Namely, consider the event $\widehat{\theta}_n \notin \Theta_{1/(2\bar{c})}(\theta_*)$. Under this event, there exists $\bar{\theta}_n \in [\theta_*, \widehat{\theta}_n]$ such that $\|\bar{\theta}_n - \theta_*\|_{\mathbf{H}(\theta_*)} = 1/2\bar{c}$. On the other hand, clearly, $L_n(\bar{\theta}_n) \leq L_n(\theta_*)$. Combining these facts with (93), we obtain that with probability at least $1-\delta$,

$$\|\nabla L_n(\theta_*)\|^2_{\mathbf{H}(\theta_*)^{-1}} \gtrsim 1/\bar{c}^2 \gtrsim 1/\bar{K}_2^6.$$

On the other hand, we know (see part $\mathbf{1^o}$ of the proof of Theorem 3.1) that

$$\|\nabla L_n(\theta_*)\|^2_{\mathbf{H}(\theta_*)^{-1}} \lesssim \frac{K_1^2 d_{\text{eff}} \log(e/\delta)}{n}$$

with probability $\geq 1-\delta$. Thus, whenever $n \gtrsim K_1^2 \bar{K}_2^6 d_{\text{eff}} \log(e/\delta)$, we have a contradiction, so $\widehat{\theta}_n$ must belong to $\Theta_{1/(2\bar{c})}(\theta_*)$. Then, (93) with $\theta = \widehat{\theta}_n$ yields

$$\|\widehat{\theta}_n - \theta_*\|^2_{\mathbf{H}(\theta_*)} \lesssim \|\nabla L_n(\theta_*)\|^2_{\mathbf{H}(\theta_*)^{-1}}.$$

It remains to bound the excess risk. To this end, recall (84) which translates to

$$\tfrac{4}{9}\phi''(0) \leq \phi''(t) \leq 4\phi''(0), \quad 0 \leq t \leq 1.$$

Integrating this twice on $[0,1]$, we obtain $\frac{4}{9}\phi''(0)t^2 \leq \phi(t) - \phi(0) \leq 4\phi''(0)t^2$, The upper bound translates to $L(\theta) - L(\theta_*) \leq \|\theta-\theta_*\|^2_{\mathbf{H}(\theta_*)}$ for any $\theta \in \Theta_{1/2\bar{c}}(\theta_*)$. But we have already proved that $\widehat{\theta}_n \in \Theta_{1/2\bar{c}}(\theta_*)$ with high probability. ■

C.4 Proof of Theorem 4.2

We use the same conventions as in the proof of Theorem 4.1. We assume w.l.o.g. that Assumption D2* holds with $r = 1/\sqrt{\rho}$, and use $\bar{K}_2 := \bar{K}_2(1/\sqrt{\rho})$ for brevity.

1°. Our first goal is to prove that the Hessians $\mathbf{H}(\theta) := \nabla^2 L(\theta)$ are close to $\mathbf{H}(\theta_*)$ within the Dikin ellipsoid with radius $1/(\bar{c}\sqrt{\rho})$ for some $\bar{c}$ depending on constants $K_0, \bar{K}_2$. Fix $\theta_0 = \theta_*$ and arbitrary $\theta_1 \in \mathbb{R}^d$, and let $\theta_t = \theta_0 + t(\theta_1 - \theta_0)$, $\Delta := \theta_1 - \theta_0$. Putting $\widetilde{X}(\theta_t) := [\ell''(Y, \langle X, \theta_t\rangle)]^{1/2} X$ as before, we have

$$\phi''(t) = \mathbb{E}[\ell''(Y, \langle X, \theta_t\rangle)\langle X, \Delta\rangle^2] = \mathbb{E}[\langle \mathbf{H}(\theta_t)^{-1/2}\widetilde{X}(\theta_t), \mathbf{H}(\theta_t)^{1/2}\Delta\rangle^2] = \|\Delta\|^2_{\mathbf{H}(\theta_t)}.$$

On the other hand, due to Assumption SCa,

$$\begin{aligned}
|\phi'''(t)| &\le \mathbb{E}[|\ell'''(Y, \langle X, \theta_t\rangle)| \cdot |\langle X, \Delta\rangle|^3] \\
&\le \mathbb{E}[\ell''(Y, \langle X, \theta_t\rangle) \cdot |\langle X, \Delta\rangle|^3] \\
&\le \mathbb{E}[\langle \widetilde{X}(\theta_t), \Delta\rangle^2 \cdot |\langle X, \Delta\rangle|] \\
&= \mathbb{E}[\langle \mathbf{H}(\theta_t)^{-1/2}\widetilde{X}(\theta_t), \mathbf{H}(\theta_t)^{1/2}\Delta\rangle^2 \cdot |\langle \mathbf{\Sigma}^{-1/2}X, \mathbf{\Sigma}^{1/2}\Delta\rangle|] \\
&\le \sqrt{\mathbb{E}[\langle \mathbf{H}(\theta_t)^{-1/2}\widetilde{X}(\theta_t), \mathbf{H}(\theta_t)^{1/2}\Delta\rangle^4]} \cdot \sqrt{\mathbb{E}[\langle \mathbf{\Sigma}^{-1/2}X, \mathbf{\Sigma}^{1/2}\Delta\rangle^2]},
\end{aligned}$$

where the last step is by Cauchy-Schwarz. Now, for $\theta_t \in \Theta_{1/\sqrt{\rho}}(\theta_*)$, one has $\|\mathbf{H}(\theta_t)^{-1/2}\widetilde{X}(\theta_t)\|_{\psi_2} \le \bar{K}_2$ due to Assumption D2*. On the other hand, $\|\mathbf{\Sigma}^{-1/2}X\|_{\psi_2} \le K_0$. Hence, by Lemma A.1 and Assumption C, we have

$$\begin{aligned}
\mathbb{E}[\langle \mathbf{H}(\theta_t)^{-1/2}\widetilde{X}(\theta_t), \mathbf{H}(\theta_t)^{1/2}\Delta\rangle^4] &\le C\bar{K}_2^4\|\Delta\|^4_{\mathbf{H}(\theta_t)}, \\
\mathbb{E}[\langle \mathbf{\Sigma}^{-1/2}X, \mathbf{\Sigma}^{1/2}\Delta\rangle^2] \le CK_0^2\|\Delta\|^2_{\mathbf{\Sigma}} &\le \rho C\bar{K}_0^2\|\Delta\|^2_{\mathbf{H}(\theta_*)},
\end{aligned}$$

for some constant $C > 0$; moreover, we can safely assume that $C > 1$ by weakening the bounds otherwise. Combining the above results, we arrive at

$$|\phi'''(t)| \le CK_0\bar{K}_2^2[\rho\phi''(0)]^{1/2}\phi''(t), \quad 0 \le t[\rho\phi''(0)]^{1/2} \le 1.$$

We now formulate a specification of Proposition B.2 for the present situation.

Proposition C.1. *Assume $g : \mathbb{R} \to \mathbb{R}$ is differentiable, non-negative, and*

$$|g'(t)| \le c\sqrt{g(0)}g(t), \quad |t| \le T$$

for $c \ge 0$. Then for $t : |t| \le T$ one has $g(0)e^{-c|t|\sqrt{g(0)}} \le g(t) \le g(0)e^{c|t|\sqrt{g(0)}}$.

Proof. We assume that $g(t) > 0$ for $t : |t| \le T$; the argument can be generalized in exactly the same way as in the proof of Proposition B.2. Denoting $\psi(t) = \log g(t)$, we obtain by integrating $\psi'(t)$ that $-c\sqrt{g(0)}t \le \log(g(t)) - \log(g(0)) \le c\sqrt{g(0)}t$, Rearranging this, we arrive at the claim. ■

Now, putting

$$\bar{c} := CK_0\bar{K}_2^2, \tag{94}$$

and applying Proposition C.1 to $g(t) = \phi''(t)$, under $\bar{c}|t|\sqrt{\bar{\rho}\phi''(0)} \le 1$ we get

$$\phi''(0)e^{-\bar{c}|t|\sqrt{\rho\phi''(0)}} \le \phi''(t) \le \phi''(0)e^{\bar{c}|t|\sqrt{\rho\phi''(0)}}.$$

Finally, since $\phi''(t) = \|\Delta\|^2_{\mathbf{H}(\theta_t)}$, this translates to the analogue of (84):

$$\frac{1}{e}\mathbf{H}(\theta_*) \preccurlyeq \mathbf{H}(\theta) \preccurlyeq e\mathbf{H}(\theta_*), \quad \theta \in \Theta_{\bar{r}}(\theta_*), \quad \bar{r} := \frac{1}{\bar{c}\sqrt{\bar{\rho}}}. \tag{95}$$

Here we used that $\Theta_{\bar{r}}(\theta_*) \subseteq \Theta_{1/\sqrt{\bar{\rho}}}(\theta_*)$ since $\bar{c} \geq 1$.

2°. We now provide a local approximation of $\mathbf{H}_n(\theta)$ using pseudo self-concordance of individual losses. Fix $\theta_0 \in \Theta_{\bar{r}}(\theta_*)$ and $\theta_1 \in \Theta$, and note that

$$\begin{aligned}|\phi_Z'''(t)| &= |\ell'''(Y, X^\top\theta_t) \cdot \langle X, \Delta\rangle|^3 \\ &\leq |\ell'''(Y, X^\top\theta_t) \cdot \langle X, \Delta\rangle|^3 = \langle \widetilde{X}(\theta_t), \Delta\rangle^2 \cdot |\langle X, \Delta\rangle| = \phi_Z''(t) \cdot |\langle X, \Delta\rangle|.\end{aligned}$$

By the argument analogous to those in Propositions B.2 and C.1, we obtain

$$\phi_Z''(0)e^{-t|\langle X,\Delta\rangle|} \leq \phi_Z''(t) \leq \phi_Z''(0)e^{t|\langle X,\Delta\rangle|},$$

which translates to $\ell''(Y, X^\top\theta_0)e^{-t|\langle X,\Delta\rangle|} \leq \ell''(Y, X^\top\theta_t) \leq \ell''(Y, X^\top\theta_0)e^{t|\langle X,\Delta\rangle|}$. Thus, denoting $\mathbf{H} := \mathbf{H}(\theta_*)$ for brevity, we have

$$\ell''(Y, X^\top\theta_0)e^{-t\|X\|_{\mathbf{H}^{-1}}\|\Delta\|_{\mathbf{H}}} \leq \ell''(Y, X^\top\theta_t) \leq \ell''(Y, X^\top\theta_0)e^{t\|X\|_{\mathbf{H}^{-1}}\|\Delta\|_{\mathbf{H}}}.$$

Equivalently, for any $\theta \in \Theta_{\bar{r}}(\theta_*)$ and $\theta' \in \Theta$,

$$\ell''(Y, X^\top\theta_0)e^{-\|X\|_{\mathbf{H}^{-1}}\|\theta'-\theta\|_{\mathbf{H}}} \leq \ell''(Y, X^\top\theta_t) \leq \ell''(Y, X^\top\theta_0)e^{\|X\|_{\mathbf{H}^{-1}}\|\theta'-\theta\|_{\mathbf{H}}}. \tag{96}$$

By Assumption D0, random vector $\mathbf{\Sigma}^{-1/2}X$ has ψ_2-norm at most $\bar{K}_0$. Hence, repeating the argument from **1°** in the proof of Theorem 3.1 we can show that, for some constant C_0, with probability at least $1-\delta$ one has

$$\max_{i\in[n]} \|X_i\|_{\mathbf{H}^{-1}} \leq C_0K_0\sqrt{\rho d \log\left(\frac{en}{\delta}\right)}. \tag{97}$$

3°. Let $\mathcal{N}_\varepsilon$ be the epsilon-net on $\Theta_{\bar{r}}(\theta_*)$, with respect to the norm $\|\cdot\|_{\mathbf{H}}$, with

$$\varepsilon = \frac{1}{C_0K_0\sqrt{\rho d \log(en/\delta)}}. \tag{98}$$

Combining this with (96) and (97), we obtain that with probability at most $1-\delta$,

$$\tfrac{1}{e}\mathbf{H}_n(\pi(\theta)) \preccurlyeq \mathbf{H}_n(\theta) \preccurlyeq e\mathbf{H}_n(\pi(\theta)), \quad \forall\theta \in \Theta_{\bar{r}}(\theta_*), \tag{99}$$

where $\pi(\cdot)$ is the projection operator on the net $\mathcal{N}_\varepsilon$. On the other hand, by Theorem A.2, it holds that

$$\tfrac{1}{2}\mathbf{H}(\theta) \leq \mathbf{H}_n(\theta) \leq 2\mathbf{H}(\theta), \quad \forall\theta \in \mathcal{N}_\varepsilon \tag{100}$$

with probability at least $1-\delta$, whenever $n \gtrsim d + \log(|\mathcal{N}_\varepsilon|/\delta)$. Recaling that $|\mathcal{N}_\varepsilon| \leq (3\bar{r}/\varepsilon)^d$, it is sufficient that

$$n \gtrsim d\log\left(\frac{e\bar{r}}{\varepsilon\delta}\right) \gtrsim d\log\left(\frac{eK_0\sqrt{d\log(en/\delta)}}{\bar{c}\delta}\right) \gtrsim d\log\left(\frac{e\sqrt{d\log(en/\delta)}}{\bar{K}_2^2\delta}\right),$$

where we used (94) and (98). Noting that $\bar{K}_2 \geq 1$, by simple algebra we have that (100) holds with probability at least $1-\delta$ whenever

$$n \gtrsim d\log(ed/\delta).$$

Finally, if this is the case, with probability at least $1-\delta$ it holds

$$\tfrac{e^2}{2}\mathbf{H}(\theta_*) \preccurlyeq \mathbf{H}_n(\theta) \preccurlyeq 2e^2\mathbf{H}(\theta_*), \quad \forall\theta \in \Theta_{\bar{r}}(\theta_*),$$

where we combined (100) with (99) and (95).

4°. As the empirical Hessians are uniformly approximated by $\mathbf{H}(\theta_*)$ in the Dikin ellipsoid with radius $\bar{r} = 1/(CK_0\bar{K}_2^2\sqrt{\rho})$, we can proceed in the same way as in step **4°** in the proof of Theorem 4.1, showing that (34) holds whenever $\|\nabla L_n(\theta_*)\|^2_{\mathbf{H}^{-1}} \lesssim 1/(\rho\bar{c}^2) \lesssim 1/(\rho K_0^2\bar{K}_2^4)$, cf. (94). This leads to the second bound on the critical sample size from the premise of the theorem. ■

C.5 Proof of Theorem 5.1

0^o. First, we follow the standard idea in the analysis of ℓ_1-penalized estimators (see, e.g., [BCW11]): using the convexity of $L_n(\theta)$, we show that whenever λ dominates $\nabla L_n(\theta)$ – which is in fact enforced by the lower bound in (40) – the essential part of the residual $\Delta := \widehat{\theta}_{\lambda,n} - \theta_*$ with high probability concentrates on the support $\mathcal{S}$. Indeed, due to the optimality of $\widehat{\theta} := \widehat{\theta}_{\lambda,n}$, we have

$$L_n(\widehat{\theta}) - L_n(\theta_*) \leq \lambda(\|\theta_*\|_1 - \|\widehat{\theta}\|_1). \tag{101}$$

Let $\Delta_{\mathcal{S}}$ be the orthogonal projection of Δ onto $\mathcal{S}$, and denote $\Delta_{\mathcal{S}_c} = \Delta - \Delta_{\mathcal{S}} = \widehat{\theta}_{\mathcal{S}}$ its projection onto $\mathcal{S}_c$, the orthogonal complement of $\mathcal{S}$. By the triangle inequality,

$$\|\theta_*\|_1 - \|\widehat{\theta}\|_1 \leq \|\Delta_{\mathcal{S}}\|_1 - \|\Delta_{\mathcal{S}_c}\|_1. \tag{102}$$

On the other hand, by convexity of $L_n(\theta)$, we have

$$L_n(\widehat{\theta}) - L_n(\theta_*) \geq -\|\nabla L_n(\theta_*)\|_\infty \|\widehat{\theta} - \theta_*\|_1 \geq -\|\nabla L_n(\theta_*)\|_\infty (\|\Delta_{\mathcal{S}}\|_1 + \|\Delta_{\mathcal{S}_c}\|_1). \tag{103}$$

Collecting (101)–(103), we get

$$(\lambda - \|\nabla L_n(\theta_*)\|_\infty)\, \|\Delta_{\mathcal{S}_c}\|_1 \leq (\lambda + \|\nabla L_n(\theta_*)\|_\infty)\, \|\Delta_{\mathcal{S}}\|_1.$$

Whence if

$$\lambda \geq 2\|\nabla L_n(\theta_*)\|_\infty, \tag{104}$$

we have that Δ satisfies the restricted subspace condition:

$$\|\Delta_{\mathcal{S}_c}\|_1 \leq 3\|\Delta_{\mathcal{S}}\|_1, \tag{105}$$

combining which with $\|\Delta_{\mathcal{S}}\|_1 \leq \sqrt{s}\|\Delta_{\mathcal{S}}\|_2 \leq \sqrt{s}\|\Delta\|_2$ results in

$$\|\Delta\|_1 \leq 4\sqrt{s}\|\Delta\|_2. \tag{106}$$

Later on, we will show that the lower bound in (40) implies (104) with probability at least $1 - \delta$. For now, let us assume that (104) holds.

1^o. To localize the estimate, we now use a similar technique to the one used in the proof of Proposition B.4, but replace the Cauchy-Schwarz inequality with Young's inequality. First, applying (61) to $L_n(\theta)$ with $\theta_0 = \theta_*$, $\theta_1 = \widehat{\theta}$, and $W = X_j$ for some (random) $j \in [n]$, we have

$$\frac{e^{-|\langle X_j, \Delta\rangle|} - 1 + |\langle X_j, \Delta\rangle|}{|\langle X_j, \Delta\rangle|^2} \|\Delta\|^2_{\mathbf{H}_n} \leq L_n(\widehat{\theta}) - L_n(\theta_*) - \langle \nabla L_n(\theta_*), \Delta\rangle,$$

Since function $u \mapsto (e^{-u} - 1 + u)/u^2$ is non-increasing, we can replace $|\langle X_j, \Delta\rangle|$ with $\|X_j\|_\infty \|\Delta\|_1$. Combining this with (101) and (102), bounding $-\langle \nabla L_n(\theta_*), \Delta\rangle$ via Young's inequality, and using (104), we get

$$\frac{e^{-\|X_j\|_\infty\|\Delta\|_1} - 1 + \|X_j\|_\infty\|\Delta\|_1}{\|X_j\|^2_\infty \|\Delta\|^2_1} \|\Delta\|^2_{\mathbf{H}_n} \leq \frac{3\lambda\|\Delta\|_1}{2}. \tag{107}$$

We now use the standard result from compressed sensing theory (see Theorem A.3 in Appendix) which states the following. Suppose that all $\mathtt{s}$-restricted eigenvalues of $\mathbf{H}$ belong to $[1/\rho, \varkappa_2]$, meaning that

$$\|\Delta\|^2/\rho \leq \|\Delta\|^2_{\mathbf{H}} \leq \varkappa \|\Delta\|^2$$

for any Δ satisfying the restricted subspace property (105) – which is clearly the case for $\mathbf{H}$ in question, due to Assumptions C and C*. Then, the corresponding sample covariance matrix $\mathbf{H}_n$ with probability at least $1-\delta$ satisfies

$$\tfrac{1}{2}\|\Delta\|_{\mathbf{H}}^2 \preccurlyeq \|\Delta\|_{\mathbf{H}_n}^2 \preccurlyeq 2\|\Delta\|_{\mathbf{H}}^2, \tag{108}$$

for any Δ satisfying (105), provided that

$$n \gtrsim \rho\varkappa_2 K_2^4 \mathbf{s}\log(ed/\delta),$$

cf. (39). Combining this result with

$$\|\Delta\|_{\mathbf{H}}^2 \geq \frac{\|\Delta\|_2^2}{\rho} \geq \frac{\|\Delta\|_1^2}{16\rho\mathbf{s}},$$

where we used (106), we have that under (39) with probability $1-\delta$ it holds

$$\|\Delta\|_{\mathbf{H}_n}^2 \geq \frac{\|\Delta\|_1^2}{32\rho\mathbf{s}}. \tag{109}$$

Combining this with (107), and denoting

$$\mathfrak{B}_{\sup} := \max_{i\in[n]}\|X_i\|_\infty, \quad u := \mathfrak{B}_{\sup}\|\Delta\|_1,$$

we obtain $e^{-u}-1+u \leq 48\rho\mathbf{s}\lambda\mathfrak{B}_{\sup}u$. From now on, we proceed as in the proof of Proposition B.4, cf. (70). That is, under

$$48\rho\mathbf{s}\lambda\mathfrak{B}_{\sup} \leq 1/2, \tag{110}$$

we sequentially obtain $u \leq 2$, $e^{-u}-1+u \geq \frac{u^2}{4}$, then $u \leq 192\rho\mathbf{s}\mathfrak{B}_{\sup}\lambda$, and

$$\|\Delta\|_1 \leq 192\rho\mathbf{s}\lambda.$$

This is the first inequality in (41), and the second one is obtained by combining it with (107)–(108). Thus, both inequalities in (41) are satisfied under the two assumed conditions (104) and (110). It remains to show that these conditions are indeed guatanteed to be satisfied with high probability under (40). For that, we have to bound the quantities $\|\nabla L_n(\theta_*)\|_\infty$ and $\mathfrak{B}_{\sup}$ from above. Indeed, due to Assumption D1, we have

$$\|\nabla\ell_Z(\theta_*)\|_{\psi_2} \leq K_1\sqrt{\varkappa_1}.$$

By Lemma A.4, this gives $\|\nabla L_n(\theta_*)\|_{\psi_2} \lesssim K_1\sqrt{\varkappa_1/n}$. Whence, $\|[\nabla L_n(\theta_*)]_i\|_{\psi_2} \lesssim K_1\sqrt{\varkappa_1/n}$ componentwise for any $i\in[n]$. Whence, by Lemma A.2, one has

$$\|\nabla L_n(\theta_*)\|_\infty \lesssim K_1\sqrt{\frac{\varkappa_1\log(ed/\delta)}{n}}$$

with probability at least $1-\delta$. This guarantees (104) under the lower bound in (40). Similarly, we can show that with probability at least $1-\delta$,

$$\mathfrak{B}_{\sup} \lesssim K_0\sqrt{\log(edn/\delta)},$$

which guarantees (110) under the upper bound in (40). The first claim of the theorem is proved; note that the upper bound in (39) is a corollary of (40).

2^o. To prove the second claim, we bound the excess risk using a similar technique as in the proof of Theorem 3.1. Note that $\mathbb{P}(\mathcal{E}) \geq 1-\delta$ by the results of **1^o**. As in the proof of Theorem 3.1,

let $\mathbf{H}_{\mathcal{E}} := \nabla^2 L_{\mathcal{E}}(\theta_*)$; recall that $\mathbf{H}_{\mathcal{E}} \preccurlyeq \mathbf{H}$. Applying (60) to $L_{\mathcal{E}}(\theta)$ with $S \leq \|X\|_\infty \|\Delta\|_1$ (recall that $X \in \mathcal{E}$), we have

$$\begin{aligned} L_{\mathcal{E}}(\widehat{\theta}) - L_{\mathcal{E}}(\theta_*) &\leq \|\nabla L_{\mathcal{E}}(\theta_*)\|_\infty \|\Delta\|_1 + \frac{e^{\|X\|_\infty \|\Delta\|_1} - 1 - \|X\|_\infty \|\Delta\|_1}{\|X\|_\infty^2 \|\Delta\|_1^2} \|\Delta\|_{\mathbf{H}}^2 \\ &\lesssim \|\nabla L_{\mathcal{E}}(\theta_*)\|_\infty \|\Delta\|_1 + \|\Delta\|_{\mathbf{H}}^2, \end{aligned}$$

where we bounded the factor ahead of $\|\Delta\|_{\mathbf{H}}^2$ by a constant using the results of $\mathbf{1}^o$. Now, define $L_{\mathcal{E}_c}(\theta) := \mathbb{E}[\ell_Z(\theta) \mathbb{1}_{\mathcal{E}_c}(X)]$ where $\mathcal{E}_0^c$ is the complimentary event to $\mathcal{E}$. Since $\nabla L(\theta_*) = 0$, we have $\nabla L_{\mathcal{E}}(\theta_*) = \nabla L_{\mathcal{E}_c}(\theta_*)$. On the other hand, for any $p, q \geq 1$ such that $1/p + 1/q = 1$, we have

$$\|\nabla L_{\mathcal{E}_c}(\theta_*)\|_\infty \leq \mathbb{E}[\|\nabla \ell_Z(\theta_*)\|_\infty \mathbb{1}_{\mathcal{E}_c}(X)] \leq \mathbb{E}[\|\nabla \ell_Z(\theta_*)\|_\infty^p]^{\frac{1}{p}} \delta^{\frac{1}{q}} \leq K_1 \sqrt{p \varkappa_1}\, d^{\frac{1}{p}} \delta^{\frac{1}{q}}.$$

where we applied Hölder's and Young's inequalities, and then Lemma A.3. Recall that in $\mathbf{1}^o$ we obtained that $\|\Delta\|_1 \lesssim \rho \mathtt{s} \lambda$ and $\|\Delta\|_{\mathbf{H}}^2 \lesssim \rho \mathtt{s} \lambda^2$ with probability at least $1 - \delta$. Combining these observations, we arrive at

$$L_{\mathcal{E}}(\widehat{\theta}) - L_{\mathcal{E}}(\theta_*) \leq (\lambda + K_1 \sqrt{p \varkappa_1}\, d^{1/p} \delta^{1/q}) \rho \mathtt{s} \lambda.$$

Choosing $p = \log(ed)$, so that $q = \log(ed)/\log(d)$, we arrive at the claim. ■

C.6 Proof of Theorem 5.2

$\mathbf{1}^o$. Let $\widehat{\theta} = \widehat{\theta}_{\lambda,n}$ for brevity. The step $\mathbf{0}^o$ of the proof of Theorem 5.1 can be repeated *verbatim*. As a result, whenever

$$\lambda \geq 2 \|\nabla L_n(\theta_*)\|_\infty, \tag{111}$$

we have

$$L_n(\widehat{\theta}) - L(\theta_*) \leq \lambda(\|\Delta_{\mathcal{S}}\|_1 - \|\Delta_{\mathcal{S}_c}\|_1) \leq \lambda \|\Delta\|_1, \tag{112}$$

$$\|\Delta_{\mathcal{S}_c}\|_1 \leq 3 \|\Delta_{\mathcal{S}}\|_1, \tag{113}$$

$$\|\Delta\|_1 \leq 4\sqrt{s} \|\Delta\|_2. \tag{114}$$

Moreover, we know (cf. the end of step $\mathbf{1}^o$ of the proof of Theorem 5.1) that (111) holds with probability at least $1 - \delta$ as long as

$$\|\nabla L_n(\theta_*)\|_\infty \lesssim K_1 \sqrt{\frac{\varkappa_1 \log(ed/\delta)}{n}}. \tag{115}$$

Hence, (111) and (115) are satisfied under the lower bound in (44). Finally, under (113) we have

$$\tfrac{1}{2} \|\Delta\|_{\mathbf{H}}^2 \preccurlyeq \|\Delta\|_{\mathbf{H}_n}^2 \preccurlyeq 2 \|\Delta\|_{\mathbf{H}}^2 \tag{116}$$

and

$$\|\Delta\|_{\mathbf{H}_n}^2 \geq \frac{\|\Delta\|_1^2}{32 \rho \mathtt{s}}, \tag{117}$$

both with probability at least $1 - \delta$, whenever $n \gtrsim \rho \varkappa_2 K_2^4 \mathtt{s} \log(ed/\delta)$.

$\mathbf{2}^o$. However, (107) does not hold since we cannot use (61). Instead, we prove

$$\frac{\|\Delta\|_{\mathbf{H}_n}^2}{1 + 3\|\widetilde{X}_j\|_\infty \|\Delta\|_1} \leq L_n(\widehat{\theta}) - L(\theta_*) - \langle \nabla L_n(\theta_*), \Delta \rangle, \tag{118}$$

where $j \in \operatorname{Argmax}_{i \in [n]} |\langle \widetilde{X}_i, \Delta \rangle|$. Indeed, to this end denote $S = |\langle \widetilde{X}_j, \Delta \rangle|$. Whenever $S < 1$, function $L_n(\theta)$ satisfies the second statement of Case **(b)** of Proposition B.3, and we obtain (118) from

the lower bound in (63). On the other hand, when $S \geq 1$ function $L_n(\theta)$ satisfies the basic statement of Case (**b**) of Proposition B.3, and we can use the lower bound in (62), i.e.,

$$\tfrac{1}{3S^2}\|\Delta\|^2_{\mathbf{H}_n} \leq L_n(\theta_{1/S}) - L(\theta_*) - \tfrac{1}{S}\langle \nabla L_n(\theta_*), \Delta\rangle, \tag{119}$$

where $\theta_{1/S}$ is the convex combination of θ_* and $\widehat{\theta}$ given by

$$\theta_{1/S} = (1 - 1/S)\cdot\theta_* + 1/S\cdot\widehat{\theta}.$$

By convexity, we have $L_n(\theta_{1/S}) \leq (1-\frac{1}{S})L_n(\theta_*) + \frac{1}{S}L_n(\widehat{\theta})$, whence $L_n(\widehat{\theta}) - L_n(\theta_*) \leq (L_n(\widehat{\theta}) - L_n(\theta_*))/S$. When combined with (119), this results in

$$\tfrac{1}{3S}\|\Delta\|^2_{\mathbf{H}_n} \leq L_n(\widehat{\theta}) - L(\theta_*) - \langle \nabla L_n(\theta_*), \Delta\rangle.$$

Whence (118) follows by Young's inequality. Now, (118), (112), and (111) imply

$$\frac{\|\Delta\|^2_{\mathbf{H}_n}}{1 + 3\|\widetilde{X}_j\|_\infty\|\Delta\|_1} \leq \frac{3\lambda\|\Delta\|_1}{2}, \tag{120}$$

which is an analogue of (107). Starting from this point, we can proceed in a similar way as in the proof of Theorem 5.2. Namely, let $\widetilde{\mathfrak{B}}_{\sup} := \|\widetilde{X}\|_\infty$ and $u := \widetilde{\mathfrak{B}}_{\sup}\|\Delta\|_1$, then (120) and (117) imply

$$\frac{u}{1+3u} \leq 48\rho\mathfrak{s}\lambda\widetilde{\mathfrak{B}}_{\sup}.$$

Hence, whenever

$$48\rho\mathfrak{s}\lambda\widetilde{\mathfrak{B}}_{\sup} \leq 1/4, \tag{121}$$

we have $u \leq 1$ and $u/(1+3u) \geq u/4$, which implies $u \leq 192\rho\mathfrak{s}\lambda\widetilde{\mathfrak{B}}_{\sup}$ and $\|\Delta\|_1 \leq 192\rho\mathfrak{s}\lambda$. This is the first inequality in (45). To obtain the second inequality, we combine (120) and (116). Thus, for (45) it remains to show that (121) holds under the upper bound in (44). We have $\|\widetilde{X}\|_{\psi_2} \leq \|\mathbf{H}^{1/2}\|_2\|\mathbf{H}^{-1/2}\widetilde{X}\|_{\psi_2} \leq K_2\sqrt{\varkappa_2}$, where we used Assumptions D2 and C*. This leads to $\widetilde{\mathfrak{B}}_{\sup} \lesssim K_2\sqrt{\varkappa_2\log(edn/\delta)}$ with probability $1-\delta$, which guarantees (121) under the upper bound in (44).

2^o. We now adapt the proof of the second claim of Theorem 5.1. Recall that in our case $\mathcal{E} := \{\|\widetilde{X}\|_\infty \lesssim K_2\sqrt{\varkappa_2\log(ed/\delta)}\}$, and $\mathbb{P}(\mathcal{E}) \geq 1-\delta$ by the results of **1^o**. As before, we put $\mathbf{H}_\mathcal{E} := \nabla^2 L_\mathcal{E}(\theta_*) \preccurlyeq \mathbf{H}$, but this time we note that $L_\mathcal{E}(\theta)$ satisfies Case (**b**) of Proposition B.3 with $S \leq \|\widetilde{X}\|_\infty\|\Delta\|_1 < 1$, cf. **1^o**. Thus, by the upper bound in (63) we have

$$L_\mathcal{E}(\widehat{\theta}) - L_\mathcal{E}(\theta_*) \lesssim \|\nabla L_\mathcal{E}(\theta_*)\|_\infty\|\Delta\|_1 + \|\Delta\|^2_{\mathbf{H}}.$$

Thence we proceed as in the proof of the second claim of Theorem 5.1. ■

D Logistic regression with Gaussian design

Change of variables. Consider a canonical GLM (17) with cumulant $a(\eta)$. Here, $\ell''(y,\eta) = a''(\eta)$ does not depend on y, hence $\widetilde{X}(\theta) = [a''(X^\top\theta)]^{1/2}X$ is fully defined by the distribution of X and the value of θ. Hence, the validity of Assumptions C, D2, D2* only depends on the distribution of X, the expression for $a''(\eta)$, and, possibly, the value of θ_* (or θ in the unit Dikin ellipsoid of θ_* in the case of Assumption D2*). Note, however, that the distribution of Y does influence Assumption D1 since the loss gradient $\ell'(Y, X^\top\theta)X = (a'(X^\top\theta) - Y)X$ contains Y. Now, consider the case of *zero-mean design*, which only makes sence when η is unrestricted, i.e., $\mathbb{R}^{(+)} = \mathbb{R}$ (note that this excludes the exponential

responce model). In this case, it is natural to pass from X and θ to the decorrelated design $Z := \mathbf{\Sigma}^{-1/2}X$ and parameter $\vartheta := \mathbf{\Sigma}^{-1/2}\theta$. Indeed, $X^\top\theta = Z^\top\vartheta$, and the corresponding vector $\widetilde{Z}(\vartheta)$,

$$\widetilde{Z}(\vartheta) := [a''(Z^\top\vartheta)]^{1/2}Z,$$

writes $\widetilde{Z}(\vartheta) = \mathbf{\Sigma}^{-1/2}\widetilde{X}(\theta)$, so that its 2nd-moment matrix $\Psi(\vartheta) := \mathbb{E}[\widetilde{Z}(\vartheta)\widetilde{Z}(\vartheta)^\top]$ is given by $\Psi(\vartheta) = \mathbf{\Sigma}^{-1/2}\mathbf{H}(\theta)\mathbf{\Sigma}^{-1/2}$. Verifying Assumption C thus reduces to bounding the lowest eigenvalue of $\Psi(\vartheta_*)$ at $\vartheta_* := \mathbf{\Sigma}^{1/2}\theta_*$, while Assumptions D2 and D2* reduce to checking $\|\Psi(\vartheta)^{-1/2}\widetilde{Z}(\vartheta)\|_{\psi_2} \lesssim K_2$ in the neighborhood of ϑ_*. Similarly, Assumption D1 can be reformulated in terms of the variables Z, ϑ, Y.

Here we consider the case of logistic regression with zero-mean Gaussian design (with arbitrary covariance), verifying the assumptions presented in Section 2.2.

Proposition D.1. *In logistic regression with $X \sim \mathcal{N}(0, \mathbf{\Sigma})$, the following holds:*

1. *Assumption C holds with*
$$\rho \lesssim 1 + \|\theta_*\|_{\mathbf{\Sigma}}^3.$$
2. *Assumption D2 holds with $K_2 \lesssim (1 + \log(1 + \|\theta_*\|_{\mathbf{\Sigma}}))\sqrt{1 + \|\theta_*\|_{\mathbf{\Sigma}}}$.*
Moreover, Assumption D2 with radius r of the Dikin ellipsoid holds with*
$$\bar{K}_2(r) \lesssim (1 + \log(1 + \|\theta_*\|_{\mathbf{\Sigma}} + r\sqrt{\rho}))\sqrt{1 + \|\theta_*\|_{\mathbf{\Sigma}} + r\sqrt{\rho}}.$$
That is, $\bar{K}_2(1/\sqrt{\rho})$ admits the same bound as K_2 up to a constant factor.
3. *If the model is well-specified, Assumption D1 holds with*
$$K_1 \lesssim \sqrt{\rho} \lesssim (1 + \|\theta_*\|_{\mathbf{\Sigma}})^{3/2}.$$
Moreover, for subexponential *norm $\|\cdot\|_{\psi_1}$, see [Ver12, Sec. 5.2.4], one has*
$$\|\mathbf{G}(\theta_*)^{-1/2}\ell'(Y, X^\top\theta_*)X\|_{\psi_1} \lesssim \log(1 + \|\theta_*\|_{\mathbf{\Sigma}})^2\sqrt{1 + \|\theta_*\|_{\mathbf{\Sigma}}};$$
equivalently, $\big(\mathbb{E}[\langle\mathbf{G}(\theta_*)^{-\frac{1}{2}}\ell'(Y, X^\top\theta_*)X, u\rangle^p\big)^{\frac{1}{p}} \lesssim Kp$ for all $u \in \mathcal{S}^{d-1}$ with
$$K = \log(1 + \|\theta_*\|_{\mathbf{\Sigma}})^2\sqrt{1 + \|\theta_*\|_{\mathbf{\Sigma}}}.$$

Proof. Note that $Z \sim \mathcal{N}(0, \mathbf{I}_d)$, and since this law is rotation-invariant, we can w.l.o.g. assume that the first coordinate vector is parallel to ϑ. Using the symmetries of $\mathcal{N}(0, 1)$, we can make sure that $\Psi(\vartheta) = \mathbf{\Sigma}^{-1/2}\mathbf{H}(\theta)\mathbf{\Sigma}^{-1/2}$ writes

$$\Psi(\vartheta) = \begin{bmatrix} \kappa & 0_{d-1}^\top \\ 0_{d-1} & \kappa_\perp \mathbf{I}_{d-1}, \end{bmatrix}, \tag{122}$$

where 0_{d-1} is the zero column, and $\kappa, \kappa_\perp$ are given in terms of the standard Gaussian density $\phi(\cdot)$ and

$$t := \|\vartheta_*\|_2 = \|\theta_*\|_{\mathbf{\Sigma}}$$

by

$$\kappa := \int_{-\infty}^{\infty} a''(tu)u^2\phi(u)\mathrm{d}u, \quad \kappa_\perp := \int_{\mathbb{R}} a''(tu)\phi(u)\mathrm{d}u.$$

In fact, the form (122) for $\Psi(\vartheta)$ will be preserved with any elliptical distribution of X, with somewhat more complicated expressions for κ and $\kappa_\perp$. Our next step is to lower-bound κ and $\kappa_\perp$, which automatically yields an upper bound for ρ in Assumption C:

$$\rho \le \frac{1}{\min(\kappa, \kappa_\perp)}. \tag{123}$$

1°. We bound κ and $\kappa_\perp$ for logistic regression. Here one has $a(\eta) = \log(1 + e^\eta)$,

$$a'(\eta) = \sigma(\eta), \quad a''(\eta) = \sigma(\eta)(1 - \sigma(\eta)),$$

where $\sigma(\eta) := 1/(1 + e^{-\eta})$ is the sigmoid. Clearly, we can bound $a''(\eta)$, $\forall \eta \in \mathbb{R}$:

$$\frac{1}{2(1 + e^{|\eta|})} \le a''(\eta) \le \frac{1}{1 + e^{|\eta|}},$$

which yields

$$\tfrac{1}{4} e^{-|\eta|} \le a''(\eta) \le e^{-|\eta|}. \tag{124}$$

Hence, letting $a \approx b$ denote the intersection of $a \lesssim b$ and $a \gtrsim b$, we have

$$\kappa_\perp \approx \int_0^\infty e^{-tu} \phi(u) \mathrm{d}u \approx \int_0^\infty e^{-tu - u^2/2} \mathrm{d}u = e^{t^2/2} G(t),$$

where

$$G(t) = \int_t^{+\infty} e^{-v^2/2} \mathrm{d}v$$

is the partial Gaussian integral. Now, [AS65, Eq. 7.1.13] gives sharp bounds for $G(t)$:

$$\frac{2e^{-t^2/2}}{t + \sqrt{t^2 + 4}} \le G(t) \le \frac{2e^{-t^2/2}}{t + \sqrt{t^2 + 8/\pi}}, \quad t \ge 0. \tag{125}$$

In particular, these bounds imply $G(t) \approx e^{-t^2/2}/(t + 1)$, whence,

$$\kappa_\perp \approx 1/(t + 1). \tag{126}$$

We can similarly bound κ:

$$\kappa \approx \int_0^\infty e^{-tu} u^2 \phi(u) \mathrm{d}u \approx e^{t^2/2} \int_0^\infty e^{-(u+t)^2/2} u^2 \mathrm{d}u = (t^2 + 1) G(t) - t e^{-t^2/2}.$$

Using the lower bound in (125), this gives

$$\kappa \ge \frac{4}{(t + \sqrt{t^2 + 4})(t^2 + 2 + \sqrt{t^4 + 4t^2})} \gtrsim \frac{1}{1 + t^3}. \tag{127}$$

Plugging (126) and (127) into (123), we arrive at $\rho \lesssim 1 + \|\theta_*\|_\Sigma^3$, as claimed. The dependency on t cannot be improved since the lower bound in (125) is sharp.

2°. On the other hand, we can estimate K_2 from Assumption D2 (and similarly $\bar{K}_2(r)$ from Assumption D2*). Indeed, note that

$$K_2 = \|\Psi(\vartheta_*)^{-1/2} \widetilde{Z}(\theta_*)\|_{\psi_2} = \sup_{u \in \mathcal{S}^{d-1}} \|\langle u, \Psi(\vartheta_*)^{-1/2} \widetilde{Z}(\theta_*)\rangle\|_{\psi_2}.$$

Let us consider separately the marginals for $u = \vartheta_*/t$ and for u from the othogonal complement of the span of ϑ. When $u = \vartheta_*/t$, we have

$$|\langle u, \Psi(\vartheta_*)^{-1/2} \widetilde{Z}(\theta_*)\rangle| = \sqrt{\frac{a''(tZ_1)}{\kappa}} |Z_1| \lesssim (1 + t^{3/2}) e^{-\frac{t|Z_1|}{2}} |Z_1|,$$

where $Z_1 \sim \mathcal{N}(0, 1)$, and we used (124) and (127). Thus, when $t \lesssim 1$, we have

$$\|\langle u, \Psi(\vartheta_*)^{-1/2} \widetilde{Z}(\theta_*)\rangle\|_{\psi_2} \lesssim \|Z_1\|_{\psi_2} \lesssim 1.$$

Let, on the contrary, $t \gtrsim 1$. Note that in the case where $|Z_1| \geq \frac{3\log(1+t)}{t}$, we have $(1+t^{3/2})e^{-t|Z_1|/2} \lesssim 1$, whence $|\langle u, \Psi(\vartheta_*)^{-1/2}\widetilde{Z}(\theta_*)\rangle| \lesssim |Z_1|$. On the other hand, when $|Z_1| \leq \frac{3\log(1+t)}{t}$, we have

$$(1+t^{3/2})e^{-t|Z_1|/2}|Z_1| \lesssim (1+t^{1/2})\log(1+t).$$

Hence, when u is parallel to ϑ_*, we have

$$\|\langle u, \Psi(\vartheta_*)^{-1/2}\widetilde{Z}(\theta_*)\rangle\|_{\psi_2} \lesssim (1+\log(1+t))\sqrt{1+t}.$$

Finally, when u is orthogonal to ϑ_*, we can use the trivial estimate

$$\|\langle u, \Psi(\vartheta_*)^{-\frac{1}{2}}\widetilde{Z}(\theta_*)\rangle\|_{\psi_2} = \left\|\sqrt{\tfrac{a''(tZ_1)}{\kappa_\perp}}\langle u, Z\rangle\right\|_{\psi_2} \lesssim \sqrt{1+t}\|\langle u, Z\rangle\|_{\psi_2} \lesssim \sqrt{1+t}.$$

In fact, this bound is tight, which can be verified by Item 2 of Lemma A.1 (note that Z_1 and $\langle Z, u\rangle$ are independent). Thus, overall we have

$$K_2 \lesssim (1+\log(1+\|\theta_*\|_{\mathbf{\Sigma}}))\sqrt{1+\|\theta_*\|_{\mathbf{\Sigma}}}. \tag{128}$$

Moreover, for $\bar{K}_2(r)$ from Assumption D2*, we clearly have

$$\begin{aligned}\bar{K}_2(r) &\lesssim \sup_{\theta\in\Theta_r(\theta_*)} (1+\log(1+\|\theta\|_{\mathbf{\Sigma}}))\sqrt{1+\|\theta\|_{\mathbf{\Sigma}}} \\ &\lesssim (1+\log(1+\|\theta_*\|_{\mathbf{\Sigma}}+r\sqrt{\rho}))\sqrt{1+\|\theta_*\|_{\mathbf{\Sigma}}+r\sqrt{\rho}}.\end{aligned}$$

This still gives (128) when $r \lesssim 1/\sqrt{\rho}$, motivating our condition in Theorem 4.2.

3°. Finally, let us verify Assumption D1, assuming well-specified model. In this case, $\mathbf{G}(\theta_*) = \mathbf{H}(\theta_*)$, and the trivial bound using $|Y - \sigma(X^\top\theta_*)| \leq 1$ is

$$K_1 \lesssim \sqrt{\rho} \lesssim 1 + t^{3/2}.$$

This is a rather discouraging result. However, we can show a weaker (subexponential) version of Assumption D1 with a milder dependency on t, replacing the $\|\cdot\|_{\psi_2}$ norm with the $\|\cdot\|_{\psi_1}$-norm as defined in [Ver12, Section 5.2.4]:

$$\|\ell'(Y, X^\top\theta_*)Z\|_{\psi_1} \lesssim \log(1+t)^2\sqrt{1+t}. \tag{129}$$

An equivalent definition of the subexponential norm is as follows: a random variable $\xi \in \mathbb{R}$ satisfies $\|\xi\|_{\psi_1} \leq K$ when its moments grow as $(\mathbb{E}[|\xi|^p])^{1/p} \lesssim Kp$, i.e., same as the moments of the exponential distribution; then, the ψ_1-norm of a random vector is defined as the maximum norm of its one-dimensional marginals. Recall that for subgaussian variables the scaling is $K\sqrt{p}$ (cf. Lemma A.1). For (129), note that in the well-specified case for $y \in \{0, 1\}$ we have

$$\mathbf{P}\{Y = y\} = \sigma(X^\top\theta_*)^y(1-\sigma(X^\top\theta_*))^{1-y},$$

thus we bound the moments of the marginals of $\ell'(Y, X^\top\theta_*)Z = (Y - \sigma(Z^\top\vartheta_*)Z$:

$$\begin{aligned}\mathbb{E}_{Z,Y}[(Y-\sigma(Z^\top\vartheta_*))\langle Z, u\rangle]^p &\leq 2\mathbb{E}_Z\left[\sigma(Z^\top\vartheta_*)(1-\sigma(Z^\top\vartheta_*))\langle Z, u\rangle^p\right] \\ &\lesssim 2\mathbb{E}_Z\left[e^{-|Z^\top\vartheta_*|}\langle Z, u\rangle^p\right], \quad p \geq 1,\end{aligned}$$

where we used (124). For u parallel to ϑ_*, we should prove that

$$(1+t)^{3/2}\left(\int_0^{+\infty} e^{-tu}u^p e^{-u^2/2}\mathrm{d}u\right)^{1/p} \lesssim p\log^2(1+t)\sqrt{1+t}. \tag{130}$$

We proceed similarly to $\mathbf{2^o}$, using that $(1+t)^{3p/2}e^{-tu} \leq 1$ for $u \geq 3p\log(1+t)/(2t)$. Thus, when $t \gtrsim 1$,

$$\begin{aligned}
&(1+t)^{3p/2}\int_0^{+\infty} e^{-tu}u^p e^{-u^2/2}\mathrm{d}u \\
\leq &(1+t)^{3p/2}\int_0^{\frac{3p\log(1+t)}{2t}} u^p \mathrm{d}u + \int_{\frac{3p\log(1+t)}{2t}}^{+\infty} u^p e^{-u^2/2}\mathrm{d}u \\
\lesssim &(1+t)^{3p/2}\frac{1}{p+1}\left(\frac{3p\log(1+t)}{2t}\right)^{p+1} + p^{p/2} \lesssim (2p)^p(1+t)^{p/2}\log(1+t)^{p+1},
\end{aligned}$$

which implies (130). The remaining cases (u parallel to ϑ_* with $t \lesssim 1$; $u \perp \vartheta_*$) are straightforward, by using that $\|\cdot\|_{\psi_1} \leq C\|\cdot\|_{\psi_2}$ for some constant C, see [Ver12]. ■

References

[AS65] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*, volume 55. Courier Corporation, 1965.

[Bac10] Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

[Bac14] Francis Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *The Journal of Machine Learning Research*, 15(1):595–627, 2014.

[Bar53] Maurice S. Bartlett. Approximate confidence intervals. II. More than one unknown parameter. *Biometrika*, 40(3/4):306–317, 1953.

[BCW11] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[BE15] Sébastien Bubeck and Ronen Eldan. The entropic barrier: a simple and optimal universal self-concordant barrier. In *Proceedings of The 28th Conference on Learning Theory*, volume 40, pages 279–279, 2015.

[BJM06] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[BKM+18] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Phase transitions, optimal errors and optimality of message-passing in generalized linear models. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pages 728–731, 2018.

[BM13] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, volume 1, pages 773–781, 2013.

[Bor98] Alexander A. Borovkov. *Mathematical Statistics*. Gordon and Breach Science Publishers, 1998.

[BRT09] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[CCK17] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45(4):2309–2352, 2017.

[CDV07] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[Chr06] Ronald Christensen. *Log-linear Models and Logistic Regression*. Springer Science & Business Media, 2006.

[CT07] Emmanuel Candes and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 36(5):2313–2351, 2007.

[DM16] David Donoho and Andrea Montanari. High-dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3-4):935–969, 2016.

[FKL$^+$18] Dylan J. Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic regression: the importance of being improper. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pages 167–208, 2018.

[HKL14] Elad Hazan, Tomer Koren, and Kfir Y. Levy. Logistic regression: tight bounds for stochastic and online optimization. In *Proceedings of The 27th Conference on Learning Theory*, volume 35, pages 197–209, 2014.

[HKZ12a] Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. *The Journal of Machine Learning Research*, 23(9):1–24, 2012.

[HKZ12b] Daniel Hsu, Sham M. Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17(52):1–6, 2012.

[HS16] Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, 17(1):543–582, 2016.

[Hub64] Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[Hub11] Peter J. Huber. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, 2011.

[IH13] Il'dar A. Ibragimov and Rafail Z. Hasminskii. *Statistical Estimation: Asymptotic Theory*. Springer Science & Business Media, 2013.

[JN11] Anatoli Juditsky and Arkadi S. Nemirovski. On verifiable sufficient conditions for sparse signal recovery via ℓ_1-minimization. *Mathematical Programming*, 127(1):57–88, 2011.

[KL17] Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 02 2017.

[Kle13] Achim Klenke. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013.

[LC06] Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer Science & Business Media, 2006.

[LM00] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

[Loh17] Po-Ling Loh. Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics*, 45(2):866–896, 2017.

[LSS14] Jason D. Lee, Yuekai Sun, and Michael A. Saunders. Proximal Newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014.

[LW11] Po-Ling Loh and Martin J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2011.

[LW15] Po-Ling Loh and Martin J. Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research*, 16(1):559–616, 2015.

[LW17] Po-Ling Loh and Martin J. Wainwright. Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 45(6):2455–2482, 2017.

[MBM18] Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

[Meh17] Nishant A. Mehta. Fast rates with high probability in exp-concave statistical learning. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 1085–1093, 2017.

[MFBR19] Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Globally convergent Newton methods for ill-conditioned generalized self-concordant losses. *arXiv:1907.01771*, 2019.

[MFOBR19] Ulysse Marteau-Ferey, Dmitrii Ostrovskii, Francis Bach, and Alessandro Rudi. *Beyond Least-Squares: Fast Rates for Regularized Empirical Risk Minimization through Self-Concordance*, volume 99. PMLR, Phoenix, USA, 25–28 Jun 2019.

[MN89] Peter McCullagh and John A. Nelder. *Generalized Linear Models, Second Edition*. Chapman & Hall, 1989.

[MZ18] Shahar Mendelson and Nikita Zhivotovskiy. Robust covariance estimation under $L_4 - L_2$ norm equivalence. *arXiv:1809.10462*, 2018.

[Nes13] Yurii Nesterov. *Introductory Lectures on Convex Optimization: a Basic Course*. Springer Science & Business Media, 2013.

[NN94] Yurii Nesterov and Arkadi S. Nemirovski. *Interior-point Polynomial Algorithms in Convex Programming*. Society of Industrial and Applied Mathematics, 1994.

[NRWY12] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[OB18] Dmitrii M. Ostrovskii and Francis Bach. Finite-sample analysis of m-estimators using self-concordance. *arXiv:1810.06838*, 2018.

[OR19] Dmitrii M. Ostrovskii and Alessandro Rudi. Affine invariant covariance estimation for heavy-tailed distributions. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99, pages 2531–2550, Phoenix, USA, 25–28 Jun 2019. PMLR.

[Pol90] David Pollard. *Empirical Processes: Theory and Applications*. NSF-CBMS regional conference series in probability and statistics. Institute of Mathematical Statistics and the American Statistical Association, 1990.

[Roc70] R. Tyrrell Rockafellar. *Convex analysis*. Princeton University Press, 1970.

[RV18] Cynthia Rush and Ramji Venkataramanan. Finite sample analysis of approximate message passing algorithms. *IEEE Transactions on Information Theory*, 2018.

[SC19] Pragya Sur and Emmanuel J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

[Spo12] Vladimir Spokoiny. Parametric estimation. Finite sample theory. *The Annals of Statistics*, 40(6):2877–2909, 2012.

[STD18] Tianxiao Sun and Quoc Tran-Dinh. Generalized self-concordant functions: a recipe for Newton-type methods. *Mathematical Programming*, 169(1):1–69, 2018.

[TAH18] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

[Tal06] Michel Talagrand. *The Generic Chaining: Upper and Lower Bounds of Stochastic Processes*. Springer Science & Business Media, 2006.

[TDKC15] Quoc Tran-Dinh, Anastasios Kyrillidis, and Volkan Cevher. Composite self-concordant minimization. *The Journal of Machine Learning Research*, 16(1):371–416, 2015.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, 58(1):267–288, 1996.

[vdGM12] Sara A. van de Geer and Patric Müller. Quasi-likelihood and/or robust estimation in high dimensions. *Statistical Science*, 27(4):469–480, 2012.

[vdV98] Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[Ver11] Roman Vershynin. Approximating the moments of marginals of high-dimensional distributions. *The Annals of Probability*, 39(4):1591–1606, 2011.

[Ver12] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge University Press, 2012.

[Vov98] Vladimir Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56(2):153–173, 1998.

[Whi82] Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, 50:1–25, 1982.

[WM17] Xiaohan Wei and Stanislav Minsker. Estimation of the covariance structure of heavy-tailed distributions. In *Advances in Neural Information Processing Systems*, pages 2859–2868, 2017.

[ZGG17] Chaoxu Zhou, Wenbo Gao, and Donald Goldfarb. Stochastic adaptive quasi-Newton methods for minimizing expected values. In *Proceedings of the 34th International Conference on Machine Learning*, pages 4150–4159, 2017.

[Zho09] Shuheng Zhou. Restricted eigenvalue conditions on subgaussian random matrices. *arXiv:0912.4045*, 2009.

[ZL15] Yuchen Zhang and Xiao Lin. DiSCO: distributed optimization for self-concordant empirical loss. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 362–370, 2015.

[ZWJ17] Yuchen Zhang, Martin J. Wainwright, and Michael I. Jordan. Optimal prediction for sparse linear models? Lower bounds for coordinate-separable M-estimators. *Electronic Journal of Statistics*, 11(1):752–799, 2017.