Finite-sample Analysis of M-estimators using Self-concordance

Dmitrii M. Ostrovskii, Francis Bach

Finite-sample Analysis of M-estimators using Self-concordance

Dmitrii M. Ostrovskii* Francis Bach*

October 16, 2018

Abstract

We demonstrate how *self-concordance* of the loss can be exploited to obtain asymptotically optimal rates for M-estimators in *finite-sample* regimes. We consider two classes of losses: (i) canonically self-concordant losses in the sense of Nesterov and Nemirovski (1994) [NN94], i.e., with the third derivative uniformly bounded with the $3/2$ power of the second; (ii) pseudo self-concordant losses, for which the power is removed, as introduced by Bach (2010) [Bac10]. These classes contain some losses arising in generalized linear models, including logistic regression; in addition, the second class includes some common pseudo-Huber losses. Our results consist in establishing the *critical sample size* sufficient to reach the asymptotically optimal excess risk, as characterized by the classical theory of local asymptotic normality, for both classes of losses. Denoting d the parameter dimension, and d_{eff} the effective dimension which takes into account possible misspecification of the parametric model, we find the critical sample size to be $O(d_{\text{eff}} \cdot d)$ for canonically self-concordant losses, and $O(\rho \cdot d_{\text{eff}} \cdot d)$ for pseudo self-concordant losses, where ρ is the problem-dependent parameter that characterizes the curvature of the risk at the best predictor θ_*. In contrast to the existing results, we only impose *local* assumptions on the data distribution, assuming that the calibrated design, i.e., the design scaled with the square root of the second derivative of the loss, is subgaussian at the best predictor. Moreover, we obtain the improved bounds on the critical sample size, scaling near-linearly in $\max(d_{\text{eff}}, d)$, under the extra assumption that the calibrated design is subgaussian in the Dikin ellipsoid of θ_*. Motivated by these findings, we construct canonically self-concordant analogues of the Huber and logistic losses with improved statistical properties. Finally, we extend some of these results to ℓ_1-regularized M-estimators in high dimensions.

1 Introduction and problem statement

Recall the standard setting of statistical learning: given a set $\Theta \subseteq \mathbb{R}^d$ that parameterizes the space of possible hypotheses, and a random observation $Z \in \mathcal{Z}$ with unknown distribution $\mathcal{P}$, one would like to minimize the *average risk*

$$L(\theta) := \mathbf{E}[\ell_Z(\theta)],$$

where 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 $\mathbf{E}[\cdot]$ is the expectation with respect to the distribution $\mathcal{P}$. This distribution is assumed unknown, so the average risk cannot be computed and minimized directly. Instead, we are granted access to the sample $(Z_1, ..., Z_n)$ of independent copies of Z, which is then used to construct an estimator $\widehat{\theta}$ of the true 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 $\mathcal{P}_n$ –

*SIERRA Project-Team, INRIA and École Normale Supérieure, PSL Research University, Paris, France. Correspondence to: `dmitrii.ostrovskii@inria.fr`

uniform probability measure supported on the sample – and the empirical risk $L_n(\theta)$, defined as the observable counterpart of $L(\theta)$:

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

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 *finite-sample* high-probability bounds on the excess risk of the ERM estimator. More precisely, our goal is to characterize the sample size sufficient for the asymptotically optimal excess risk bound to become available in finite sample.

Classical theory of local asymptotic normality (LAN). Our main focus in this paper is the setting where $L_n(\theta)$ is a negative log-likelihood, meaning that $\ell_z(\theta) = -\log p_\theta(z)$ 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 some *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*; ERM can then 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 $\mathcal{P}_n$. Now, this work has been motivated by the following question:

Can we extend the classical LAN theory for quasi-MLE to the non-asymptotic setting? (?)

To clarify and substantiate (?), let us briefly review the main results of the LAN theory. Most of them, see monographs [LC06, IH13, vdV98, Bor98], rely on the *local regularity* assumptions about the log-likelihood, allowing for a 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 internal 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., $\mathbf{H} \succ 0$. When combined together, these assumptions allow to derive the *local asymptotic normality* of quasi MLE: in the limit $n \to \infty$ with fixed d, we have

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

where $\rightsquigarrow$ denotes convergence in law, and $\mathbf{G}$ is the covariance matrix of the loss gradient at θ_*:

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

Matrices $\mathbf{G}$ and $\mathbf{H}$ remain fixed as n grows, and hence the above results imply, in particular, that the variance of $\widehat{\theta}_n$ decreases at the rate $1/n$. Moreover, in the well-specified case one has $\mathbf{G} = \mathbf{H}$, see, e.g., [Bar53], which leads to *Fisher's theorem*:

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

Thus, denoting $\|\cdot\|_{\mathbf{M}}$ the norm linked to positive semidefinite matrix $\mathbf{M}$ by $\|x\|_{\mathbf{M}} = \|\mathbf{M}^{1/2}x\|_2$,

$$n\|\widehat{\theta}_n - \theta_*\|_{\mathbf{H}}^2 \rightsquigarrow \chi_d^2, \tag{2}$$

where χ_d^2 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*. This implies, in particular, that

$$\mathbf{E}_n[L(\widehat{\theta}_n)] - L(\theta_*) = \frac{d}{2n} + o(n^{-1}), \tag{3}$$

where $\mathbf{E}_n$ is the expectation over the sample $(Z_1, ..., Z_n)$. Moreover, via the chi-square deviation bound (see [LM00, Lemma 1]), the limiting law (1) implies that with probability at least $1-\delta$,

$$L(\widehat{\theta}_n) - L(\theta_*) = \frac{(\sqrt{d} + \sqrt{2\log(1/\delta)})^2}{2n} + o(n^{-1}). \tag{4}$$

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

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

Note that in a well-specified model, $d_{\mathrm{eff}} = d$ since $\mathbf{G} = \mathbf{H}$; moreover, in the ill-specified case one can still have $d_{\mathrm{eff}} = O(d)$ "in favorable circumstances" – we will consider one such situation, that of misspecified linear regression, later on.[1] The expected excess risk bound (3) then changes to

$$\mathbf{E}_n[L(\widehat{\theta}_n)] - L(\theta_*) = \frac{d_{\mathrm{eff}}}{2n} + o(n^{-1}), \tag{5}$$

and the corresponding in-probability bound (see again [LM00, Lemma 1])[2] can be expressed as

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

In fact, the main term in the right-hand side of (5) is the minimum possible asymptotic variance of any unbiased estimator; this result is known as the Cramér-Rao bound. For what is to follow, it is important to note that the asymptotic approach can be summarized in the following steps:

- First, the estimate is localized: $\|\widehat{\theta}_n - \theta_*\|_{\mathbf{H}}^2$ is shown to behave as the squared "natural" norm of the score, $\|\nabla L_n(\theta_*)\|_{\mathbf{H}^{-1}}^2$, which is 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 LAN theory, one should keep in mind that this theory is limited to the asymptotic regime $n \to \infty$ with fixed parameter dimension, which hinders its use 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 the finite-sample regime is scarce: in fact, the only paper we are aware of is [RV18], which only considers linear regression with fixed design.

[1] Note, however, that it can also happen that $d_{\mathrm{eff}} < d$ if we get "extremely lucky". For example, consider the Gaussian shift model $y \sim \mathcal{N}(\theta, 1)$, and suppose that in reality $y \sim \mathcal{N}(0, \sigma)$. Then $d_{\mathrm{eff}} = \sigma^2$ which can be arbitrary.

[2] Notice that ($\star$) is weaker than (4) due to the replacement of d with d_{eff}, but also because of the main term scaling as $O(d_{\mathrm{eff}} \log(1/\delta))$. This is due to the fact that the chi-squared statistic in (1) is replaced with the "generalized chi-squared" statistic $\chi_{\mathbf{M}}^2 := \|\xi\|_{\mathbf{M}}^2$, where $\xi \in \mathcal{N}(0, \mathbf{I}_d)$, and $\mathbf{M} = \mathbf{H}^{-1/2}\mathbf{G}\mathbf{H}^{-1/2}$. Unless $\mathbf{M} = \mathbf{I}_d$, various matrix norms of $\mathbf{M}$ which control the deviations of $\|\xi\|_{\mathbf{M}}$, see [LM00, Lemma 1], might be as large as d_{eff}.

Finite-sample regime and empirical processes. One rather general approach for answering 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 *Dikin ellipsoid* of the true optimum, defined as

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

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 of 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 $n \ge c_\delta d_{\text{eff}}$, 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 their 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 process theory.

On the other hand, as we demonstrate next, in the case of linear regression with random design, finite-sample analysis is much simpler, and heavyweight techniques from empirical processes theory are not needed. In this case, the problem is reduced to controlling the sample covariance matrix of the design, which encapsulates the second-order information about the risk. Our primal goal in this paper is to extend these ideas to a wider class of models with non-quadratic losses, including conditional generalized linear models and regression models with robust losses. In these models, the second-order information about the risks is local, and covariance matrix estimation should be augmented with some local approximation techniques.

Simple case: linear regression. An original approach introduced in [HKZ12a] allows to answer (?) 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 *predictor* $\theta \in \mathbb{R}^d$, and one takes $\ell_Z(\theta)$ to be the normalized square loss (and ERM is the ordinary least-squares estimator):

$$\ell_Z(\theta) = \frac{1}{2\sigma^2}(Y - X^\top\theta)^2.$$

This corresponds 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 $\mathbf{E}[(Y - X^\top\theta)^2] - \sigma^2$. 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 $\nabla\ell_Z(\theta_*) = \varepsilon X$ and $\mathbf{H} = \mathbf{E}[XX^\top]$, we get $d_{\text{eff}} = \mathbf{E}[\varepsilon^2\|\mathbf{H}^{-1/2}X\|^2]$. Hence, $d_{\text{eff}} = d$ for any distribution of ε with $\mathbf{E}[\varepsilon^2] = 1$, provided that ε and X are independent. Moreover, assuming that Y and all one-dimensional marginals of X

have finite fourth moment, i.e.,

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

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 in this situation.

The approach of [HKZ12a] is based on the observation that since $L(\theta)$ is a quadratic form,

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

the empirical risk $L_n(\theta)$ is also a quadratic form with random matrix $\mathbf{H}_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$:

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

In other words, 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 has subgaussian marginals, and

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

where symbol $\gtrsim$ hides an absolute constant, and K is a constant which depends on the concentration properties of X, with high probability it holds

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

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 $\mathbf{H}_n \succ 0$,

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

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

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

Finally, a non-asymptotic version of ($\star$) is obtained by controlling $\|\nabla L_n(\theta_*)\|_{\mathbf{H}^{-1}}^2$, under subgaussian assumptions on $\nabla \ell_Z(\theta_*) = \varepsilon X$, through standard concentration inequalities. In fact, the subgaussian assumption can be relaxed to the fourth-moment assumption, see [HS16], using the generalized median-of-means estimator,[3] and the light-tailed assumptions on X can also be relaxed, through a rejection sampling argument similar to the one that we employ in Theorem 3.4.

The remarkable feature of the outlined argument 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 for the sample size is the lower bound (7), which allows to relate the norms $\|\cdot\|_{\mathbf{H}_n}$ and $\|\cdot\|_{\mathbf{H}}$. This is entirely due to the fact that the loss is quadratic, and the curvature information is *global*. However, in the general case the information about the curvature of the risk is not encoded in a single matrix, and there seems to be no direct way of extending the above argument. As discussed before, the known solution to the problem [Spo12] involved

[3] One divides the sample into $\log(1/\delta)$ non-overlapping subsamples, computes independent least-squares estimators over this subsamples, and take as the final estimator their multi-dimensional median in $\|\cdot\|_{\mathbf{H}_n}$-distance.

localization of the estimate, through the control of the *global* uniform deviations of $L_n(\theta)$, to the neighborhood of θ_* where the local quadratic approximations can be used, and 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 non-transparent localization step using generic chaining, as in [Spo12], can be circumvented. Namely, this is the case when the loss satisfies certain *self-concordance* assumptions, which allows to control the precision of local quadratic approximations of $L_n(\theta)$ and $L(\theta)$ more directly, using a simple integration technique.

1.1 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 must predict Y through a linear combination $\eta = X^\top \theta$ where $\theta \in \Theta \subseteq \mathbb{R}^d$. Accordingly, we consider M-estimators with losses of the form $\ell_Z(\theta) = \ell(Y, X^\top \theta)$ for some function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ which is 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 in [Bac10, BM13]. We consider both variants of self-concordance in our analysis, and show that the original self-concordance, as proposed in [NN94], leads to better statistical results than its modification suggested in [Bac10] (see Section 3), and is therefore more favorable from the statistical perspective. 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 true optimal point θ_*, which is needed to control the norm of $\nabla L_n(\theta)$ and the deviations of $\mathbf{H}_n$ from $\mathbf{H}$. We mention once again that the *global* assumptions in the vein of [Spo12] about the deviations of the empirical risk, its gradient, and Hessian can be avoided by making use of the self-concordance of the loss at hand.

Our framework includes random-design linear regression, which provides a "sanity check" for our results. However, as we show in Section 2, the framework 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 some natural assumptions about the moments of the response, 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 these losses to make them *canonically* self-concordant while preserving their first- and second-order structure. According to our theory, this leads to better tradeoff between the statistical performance of the M-estimator, as characterized by the sufficient sample size to reach the asymptotically optimal rate, and its robustness properties.

In our analysis, we execute the following plan. First, the local assumptions allow to make sure that starting from the certain sample size, the sample Hessian at the optimal point, that is,

$$\mathbf{H}_n := \frac{1}{n} \sum_{i=1}^{n} \ell''(Y_i, X_i^\top \theta_*) X_i X_i^\top,$$

approximates the true Hessian $\mathbf{H} = \mathbf{E}[\ell''(Y, X^\top \theta_*) X X^\top]$ up to a constant factor, completely analogous to the case of least squares. After that, self-concordance comes at play: using the properties of canonically and pseudo self-concordant functions presented in Section 3.1, we show that the ERM estimator $\widehat{\theta}_n$ gets localized around θ_* whenever the natural norm of the score $\|\nabla L_n(\theta_*)\|_{\mathbf{H}^{-1}}$ is small enough. Overall, we obtain the following results in Section 3:

- Our analysis in Sections 3.2–3.3 shows that for *pseudo* self-concordant losses as in [Bac10], $\|\nabla L_n(\theta_*)\|_{\mathbf{H}^{-1}}$ must be smaller, with high probability, than the quantity $\|X\|_{\mathbf{H}^{-1}}$. This is guaranteed to happen once the sample size achieves the threshold $O(\rho \cdot d \cdot d_{\text{eff}})$ up to distribution constants and logarithmic factors in $1/\delta$, where ρ is the local curvature parameter linking $\mathbf{H}$ and $\mathbf{\Sigma} := \mathbf{E}[XX^\top]$ as follows:

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

  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 extremely 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. For example, one can show (see Appendix C) that $\rho \lesssim 1 + \|\theta_*\|_{\mathbf{\Sigma}}^3$ in logistic regression with Gaussian design $X \sim \mathcal{N}(0, \mathbf{\Sigma})$. Moreover, for *canonically* self-concordant losses in the sense of [NN94], the dependency on ρ can be eliminated, and the critical sample size becomes $O(d \cdot d_{\text{eff}})$, see Theorems 3.2 and 3.4. Thus, canonically self-concordant losses have somewhat better statistical behavior according to our theory. Motivated by this result, we propose canonically self-concordant losses for robust estimation and classification in Section 2.1.

- In Section 3.4, we obtain improved bounds for the critical sample size, eliminating the extra d factor under a slightly stronger assumption on the data distribution. In particular,

$$n \gtrsim \max(d_{\text{eff}}, d \log d) \tag{10}$$

  turns out to be sufficient for canonically self-concordant losses, cf. Theorem 3.5, while

$$n \gtrsim \max(\rho d_{\text{eff}}, d \log d) \tag{11}$$

  sufficient in the case of pseudo self-concordant losses, cf. Theorem 3.6, up to the distribution constants and logarithmic in $1/\delta$ factors. The extra assumption is rather mild: we require that the *calibrated design* $\widetilde{X}(\theta) := [\ell''(Y, X^\top\theta)]^{1/2} X$ has subgaussian tails at any point θ within the *Dikin ellipsoid* of θ_* with some radius $r > 0$,

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

  rather than only at the true optimum θ_*. Specifically, we require $r \gtrsim 1$ for canonically self-concordant losses, and $r \gtrsim 1/\sqrt{\rho}$ for pseudo self-concordant losses (we also motivate the latter choice on the example of logistic regression with Gaussian design). This assumption allows to control the uniform deviations of the empirical Hessians from their means on $\Theta_r(\theta_*)$, leading to the reduced sample size. The analysis is complicated by the fact that self-concordance of the individual losses *does not* imply self-concordance of empirical risk. Instead, control of sample Hessians on $\Theta_r(\theta_*)$ is achieved by noting that self-concordance of the losses suffices to control Hessians in a small Dikin ellipsoid with radius $O(1/d^\kappa)$ for some $\kappa > 0$, and combining this observation with a simple covering argument.

  We hypothesize that the bounds (10)–(11) are optimal up to the $\log(d)$ factor, i.e., the empirical risk minimizer cannot provably achieve the nonasymptotic version of ($\star$) when the sample size is sublinear in d_{eff} or d. This hypothesis is based on the following observations:

  - the linear growth of n in d is necessary to estimate the local metric related to $\nabla L(\theta_*)$;
  - the linear growth of n in d_{eff} is necessary to have small score $\|\nabla L_n(\theta_*)\|_{\mathbf{H}} \le c$, which is the key property allowing to localize $\widehat{\theta}_n$ to the neighborhood of the true optimum θ_*.

- In Section 3.5, we extend some of the above results to the high-dimensional setup. Specifically, we obtain analogues of Theorems 3.3 and 3.4 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 3.7), 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 3.8), and the bounds essentially remain the same as in the case of pseudo self-concordance. This, however, is not surprising since both 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 the $\|\cdot\|_{\mathbf{H}}$-norm. Overall, our results are comparable to those in [Bac10] and [vdGB09]; however, there are some important differences as discussed in Section 3.5.

1.2 Related work

Our techniques are inspired from [Bac10], and we use and extend some of their technical results in our Propositions 3.4–3.5. 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 establish 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] study the ℓ_2-penalized estimator $\widehat{\theta}_{\lambda,n} = \arg\min_\theta L_n(\theta) + \lambda\|\theta\|_2^2$ with *strictly positive* regularization parameter λ, and moreover, impose some technical condition on the minimal magnitude of λ, see their equation (13). Close inspection of this condition shows that it implies $n \gtrsim \rho\, \mathtt{df}_1^2$, where

$$\mathtt{df}_1 := \operatorname{tr}(\mathbf{H}(\mathbf{H} + \lambda \mathbf{I})^{-1})$$

is *the ℓ_1-number of degrees of freedom*, a quantity replacing d in the ℓ_2-penalized setting. This, in turn, allows to carry out an argument analogous to ours, but applying Proposition 3.5 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 $\theta_\lambda = \arg\min_\theta L(\theta) + \lambda\|\theta\|_2^2$. 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 estimator, 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 large enough sample size.

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

$$\mathbf{E}_n[L(\widehat{\theta})] - L(\theta_*) \lesssim \frac{R^2(R^4 D_0^4 + 1)}{\mu n}, \tag{12}$$

where μ is the lowest eigenvalue of $\mathbf{H}$, R is an upper bound on $\|X\|_2$ and $\sup_{\theta\in\Theta} \|\nabla \ell_Z(\theta)\|_2$, and $D_0 := \|\theta_0 - \theta_*\|_2$ is the intial ℓ_2-distance from the optimum (in fact, if D_0 is known up to a constant factor, $R^4 D_0^4$ in (12) can be replaced with $R^2 D_0^2$). The bound (12) reflects the fact that gradient 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 (12), but R would be replaced with $\max(d_{\text{eff}}, \rho d)$, 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 in which the local quadratic subproblems are solved via stochastic approximation. This allows to conduct affine-invariant analysis of the outer loop, and results in

$$\mathbf{E}_n[L(\widehat{\theta})] - L(\theta_*) \lesssim \frac{\rho^2(R^4 D_0^4 + 1)\max(d_{\text{eff}}, \rho d)}{n} \tag{13}$$

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].

Finally, we should mention the recent surge of interest in stochastic quasi-Newton methods applied to the finite-sum setting with self-concordant losses, see, e.g., [ZGG17], [ZL15]. However, none of these works is concerned with establishing the asymptotically optimal rates.

Existing literature on ℓ_1-regularized M-estimation will be reviewed separately in Section 3.5.

Notation. We write $f \lesssim g$ or $f = O(g)$ to state that $f(\cdot) \le Cg(\cdot)$ for any possible arguments of $f(\cdot)$ and $g(\cdot)$ and some constant C; analogously for $f \gtrsim g$ or $f = \Omega(g)$. $[n]$ is the set of integers $\{1, 2, ..., n\}$. Throughout, θ_* is the unique minimizer of $L(\theta)$, Similarly, $\widehat{\theta}_n$ denotes the minimizer of $L_n(\theta)$, which is unique in all the cases of interest due to Assumptions SCa and SCb below. Similarly, $\widehat{\theta}_n$ is the minimizer of $L_n(\theta)$, which is proved to be unique with high probability in all cases of interest. Random vectors are denoted with capital letters (such as Z), and matrices with bold capital letters (such as $\mathbf{M}$). $\mathbf{I}_d$ denotes the $d \times d$ identity matrix. $\mathbf{M}^\top$ is the transpose of a matrix $\mathbf{M}$. For a pair of semidefinite matrices $\mathbf{M}_1, \mathbf{M}_2$ of the same size, we write $\mathbf{M}_1 \preccurlyeq \mathbf{M}_2$ whenever $\mathbf{M}_2 - \mathbf{M}_1$ is positive semidefinite. We denote with $\|\cdot\|_p$ both the ℓ_p-norm on $\mathbb{R}^d$ and the Schatten ℓ_p-norm of a matrix $\mathbf{M}$, in particular, $\|\mathbf{M}\|_2$ is the Frobenius norm, and $\|\mathbf{M}\|_\infty$ the operator norm. Given $\mathbf{M} \succ 0$, we define the norm $\|\theta\|_{\mathbf{M}} := \|\mathbf{M}^{1/2}\theta\|_2$.

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 always clear from the context. The derivatives of $\ell(y, \eta)$ are taken with respect to η.

2.1 Self-concordance assumptions

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

We first present the assumption of *pseudo self-concordance*, introduced in [Bac10] for the analysis of logistic regression. We refer to [STD18, TDKC15, BM13] for an overview of self-concordant-like functions in the general context of optimization and quasi-Newton algorithms.

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

On the other hand, we also consider the *canonical* self-concordance assumption which has been first introduced in [NN94] in the context of interior-point algorithms. The constant 2 allows to slightly simplify our results, and can be replaced with arbitrary constant by scaling the loss.

Assumption SCb. *For any $y \in \mathcal{Y}$ and $\eta \in \mathbb{R}^{(+)}$, $\ell(y,\eta)$ satisfies $|\ell'''(y,\eta)| \le 2[\ell''(y,\eta)]^{3/2}$.*

We now present some examples in which any of the two assumptions are satisfied.

Generalized linear models with canonical parametrization. In *generalized linear models* (GLM) with canonical parametrization, see [MN89], one has

$$\ell(y,\eta) = -y\eta + a(\eta) - b(y), \tag{14}$$

where the *cumulant generating function* $a(\eta) : \mathbb{R}^{(+)} \to \mathbb{R}$ normalizes $\ell(y,\eta)$ to be a valid negative 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 $\eta = X^\top\theta$. Some remarks are in order.

- 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 (note that $a(\cdot)$ is necessarily convex). Moreover, the derivatives of the cumulant are equal to the central moments of Y. In particular,

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

  where $\mathbf{E}_\eta[\cdot]$ is expectation with respect to the distribution with negative log-likelihood given by (14). 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].[4] 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 $\mathcal{Y} = \{0,1\}$:

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

[4] Note that this model is different from the *Poisson likelihood* model $Y \sim \mathrm{Poisson}(X^\top\theta)$, which is not a GLM.

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].[5] These considerations, along with certain minimax optimality results, lead to the well-known *Huber loss* as introduced in [Hub64]:

$$\varphi_\tau(t) = \begin{cases} \dfrac{t^2}{2}, & |t| \le \tau, \\ \tau t - \dfrac{\tau^2}{2}, & |t| > \tau. \end{cases} \tag{15}$$

The Huber loss is parametrized by $\tau > 0$, which allows to control the tradeoff of robustness and statistical performance. Indeed, on one hand, $|\varphi_\tau'(t)| \le \tau$ for any $t \in \mathbb{R}$, while on the other hand, the variance of the corresponding M-estimator usually decreases with 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. Examples include losses of the form $\varphi_\tau(t) = \tau^2\varphi(t/\tau)$, with the following contrasts:

$$\varphi(t) = \log\left(\frac{\exp(t) + \exp(-t)}{2}\right), \quad \varphi(t) = \sqrt{1+t^2} - 1. \tag{16}$$

In both cases, the resulting scaled function 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 $\varphi(t)$ satisfy Assumption SCa up to $c = 2$ in the first case, and $c = 3$ in the second case, whence,

$$|\varphi_\tau'''(t)| \le \frac{c\phi_\tau''(t)}{\tau}.$$

As such, our theory is applicable to both these losses once they are properly renormalized. Finally, note that we can obtain a pseudo-Huber loss with the above properties, for any $\tau > 0$, if we take $\varphi_\tau(t) = \tau^2\varphi(t/\tau)$ where function $\varphi(t)$ satisfies $\varphi''(0) = 1$, $|\varphi'(t)| \le 1$, and $|\varphi'''(t)| \le c\phi''(t)$.

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 is useful since, according to our theory, canonically self-concordant losses have provably better statistical properties than pseudo self-concordant ones. The key idea in this contruction is that self-concordance is preserved when passing to the convex conjugate, whose gradients we can easily control, see, e.g., [STD18, Proposition 6]. Namely, consider $\phi : (-1, 1) \to \mathbb{R}^+$ given by

$$\phi(u) = -\log(1 - u^2)/2, \tag{17}$$

that is, the negative log-barrier on $[-1, 1]$ normalized by $\phi''(0) = 1$. Its Fenchel dual is given by

$$\varphi(t) = \frac{1}{2}\left[\sqrt{1+4t^2} - 1 + \log\left(\frac{\sqrt{1+4t^2} - 1}{2t^2}\right)\right]. \tag{18}$$

[5] However, this condition is *not* necessary for the local asymptotic normality: the sample median in the shift model $y = \theta + \varepsilon \in \mathbb{R}$, corresponding to $\varphi(t) = |t|$, is asymptotically normal if the density of ε does not vanish at 0.

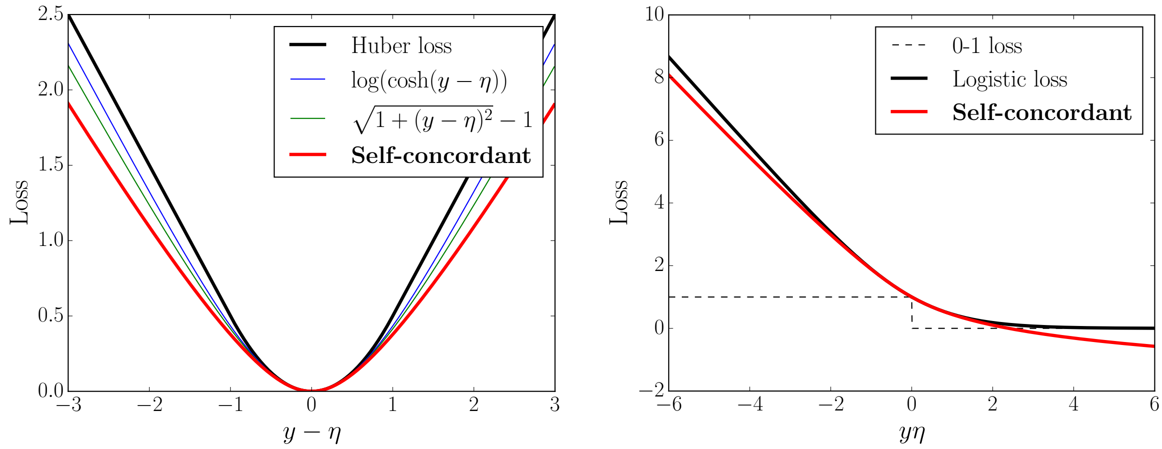

Figure 1: *Left:* self-concordant pseudo-Huber loss, cf. (18). *Right*: self-concordant analogue of the logistic loss suitable for classification, cf. (19). Although our classification loss does not upper-bound the 0-1 loss on $\mathbb{R}^+$, it can be lower-bounded with $\Omega(-\log(y\eta))$ for positive margins.

This function 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 some simple calculations detailed in Appendix B, $\varphi(t)$ defined in (18) 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 (16). It can also be generalized to have arbitrary slope, by considering $\varphi_\tau(t) = \tau^2\varphi(t/\tau)$ which satisfies $\varphi_\tau''(0) = 1$, $|\varphi_\tau'(t)| \le \tau$, 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 it is canonically self-concordant (up to a constant) by the same reasoning as before. By rescaling and shifting it, we obtain the loss

$$\ell(y,\eta) = 2 + \frac{1}{2\log 2}\left[-1 - y\eta + \sqrt{1+(y\eta)^2} + \log\left(\frac{\sqrt{1+(y\eta)^2}-1}{2(y\eta)^2}\right)\right] \tag{19}$$

which can be considered as a convex surrogate of the 0-1 loss similar to the logistic loss, see Figure 1. However, this loss is negative for $y\eta > 2.396...$, 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.[6] 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 $|([\varphi''(t)]^{-1/2})'| \le 1$.

[6]This effect can be quantified using the calibration theory developed in [BJM06]. We leave this for future work.

2.2 Distribution assumptions

We now introduce additional assumptions that are related to the distribution of the design and the derivatives of the loss at the true optimum θ_*. Note that all these assumptions are fully *local*, i.e., they only concern the true optimal point θ_*. Let us start with some preliminaries first.

- 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 some other pair of mutually dual convex cones in $\mathbb{R}^d$.

- We denote $\mathbf{\Sigma} := \mathbf{E}[XX^\top]$ the second-order moment matrix of X, and assume that it exists.

- Note that for any $\theta \in \Theta$ one has

$$\nabla \ell_Z(\theta) = \ell'(Y, X^\top \theta) X, \quad \nabla^2 \ell_Z(\theta) = \ell''(Y, X^\top \theta) XX^\top. \tag{20}$$

  Recall that $\mathbf{E}[\nabla \ell_Z(\theta_*)] = 0$, $\mathbf{E}[\nabla \ell_Z(\theta_*) \nabla \ell_Z(\theta_*)^\top] = \mathbf{G}$, and $\mathbf{E}[\nabla^2 \ell_Z(\theta_*)] = \mathbf{H}$. Generally, $\mathbf{\Sigma} \neq \mathbf{H}$ (unless the loss is quadratic), and $\mathbf{G} \neq \mathbf{H}$ (unless in a well-specified model).

- Following [Ver12], we use the standard definition of ψ_2-norms for subgaussian distributions. 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), for example, as follows:

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

  This definition is extended to random vectors $Z \in \mathbb{R}^d$ as

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

  In other words, $\|Z\|_{\psi_2}$ is the maximal $\|\cdot\|_{\psi_2}$-norm for all one-dimensional marginals of Z. We recount alternative definition of the ψ_2-norm and some useful results related to subgaussian distributions in Appendix A.

Assumption D0. *The decorrelated design is subgaussian:*

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

Assumption D0 is often satisfied with a constant K_0 not depending on n or d. For example, this the case for zero-mean Gaussian design $X \sim \mathcal{N}(0, \mathbf{\Sigma})$, 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 gradient of the loss at the optimal point is subgaussian:*

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

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

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

the design scaled by the curvature of the loss at the optimal point. Note that $\widetilde{X}$ is linked with the Hessian by $\mathbf{E}[\widetilde{X}\widetilde{X}^\top] = \mathbf{H}$, cf. (20). 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 decorrelated calibrated design* $\widetilde{X} := [\ell''(Y, X^\top \theta_*)]^{1/2} X$ *satisfies*

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

Assumption D2 can be reformulated in terms of the loss Hessian $\nabla^2 \ell_Z(\theta_*)$ due to (20). However, this formulation does not give new ideas, and we omit it. Some remarks are in order.

- 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 $(\mathbf{E}|\xi|^4)^{1/4} \ge (\mathbf{E}|\xi|^2)^{1/2}$ for any random variable $\xi \in \mathbb{R}$, and allows to simplify the formulation of the subsequent results.

- Assumptions D1–D2 are in fact quite restrictive, even when D0 is assumed. In particular, in the canonically-parametrized GLMs (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 as follows: even when Assumptions D1–D2 are satisfied, K_1 and K_2 can depend on the unknown solution θ_*. For example, in Appendix C, we show that for logistic regression with Gaussian design $X \sim \mathcal{N}(0, \mathbf{\Sigma})$ one has

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

  and, provided that the model is well-specified,

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

  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 C and Section 3.3 for details). Thus, in concrete applications one shoud carefully verify Assumptions D1–D2 and bound the quantities K_1 and K_2. Such verification can itself be a complicated task, especially without the exact knowledge of the distribution of X.

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

Assumption C (Compatibility of $\mathbf{\Sigma}$ and $\mathbf{H}$)**.** *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 only available generic upper bound for ρ is

$$\rho \le \frac{1}{\inf_{(y,\eta)\in\mathcal{Y}\times\mathbb{R}^{(+)}} \ell''(y,\eta)}, \tag{21}$$

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 (21) 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 $\rho = 1/\mathbf{E}[\varphi''(\varepsilon)]$. On the other hand, the worst-case bounds on ρ can be enforced if the data distribution is chosen *adversarially*. In particular, for the logistic regression [HKL14] construct an adversarial distribution that enforces $\rho = \Omega(e^{RD})$ in the regime $n = O(e^{RD})$.

Boundedness assumptions. To ease the presentation, we now give simplified versions of the results (Theorems 3.1–3.2) holding under the strengthened versions of Assumptions D0 and D2, in which the decorrelated design (or the calibrated design at θ_*) is almost surely bounded.

Assumption B0. *The decorrelated design is $\mathcal{P}$-a.s. bounded:*

$$\|\mathbf{\Sigma}^{-1/2}X\|_2 \le \mathfrak{B}_0.$$

Assumption B2. *The decorrelated calibrated design at θ_* is $\mathcal{P}$-a.s. bounded:*

$$\|\mathbf{H}^{-1/2}\widetilde{X}\|_2 \le \mathfrak{B}_2.$$

Some further remarks regarding these assumptions are summarized below.

- $\|X\|_2 \le R$ implies $\mathfrak{B}_0^2 \le R^2/\lambda_{\min}(\mathbf{H})$, and similarly for $\mathfrak{B}_2^2$ under the bound for $\|\widetilde{X}\|_2$.
- By Markov's inequality, both $\mathfrak{B}_0$ and $\mathfrak{B}_2$ must be lower bounded a constant times $\sqrt{d}$.
- On the other hand, Assumptions D0 and D2 "almost" imply Assumptions B0 and B2 with $O(\sqrt{d})$ radii. Specifically, by Corollary A.2 and Lemma A.5 in Appendix, with probability at least $1-\delta$ we have $\|\mathbf{\Sigma}^{-1/2}X\|_2 \lesssim K_0\sqrt{d\log(e/\delta)}$ and $\|\mathbf{H}^{-1/2}\widetilde{X}\|_2 \lesssim K_2\sqrt{d\log(e/\delta)}$.

3 Theoretical results

3.1 Useful technical propositions

Here we summarize technical results related to self-concordant-like functions. These results will be used later on to control the empirical and average risks. We defer the proofs to Appendix B.

Let us fix two arbitrary parameter values $\theta_0, \theta_1 \in \Theta$, and let $\theta_t := \theta_0 + t(\theta_1 - \theta_0)$ for $0 \le t \le 1$. Consider functions $\phi_Z(\cdot)$, $\phi(\cdot)$, and $\phi_n(\cdot)$ on $[0,1]$, defined by

$$\phi_Z(t) := \ell_Z(\theta_t), \quad \phi(t) := L(\theta_t) = \mathbf{E}[\phi_Z(t)], \quad \phi_n(t) := L_n(\theta_t) = \frac{1}{n}\sum_{i=1}^{n} \ell_{Z_i}(\theta_t). \tag{22}$$

We first state a direct consequence of the assumptions of Section 2.1. The proof of the next proposition is essentially just an application of the chain differentiation rule.

Proposition 3.1. *Suppose that the loss $\ell_z(\theta)$ is convex and three times differentiable on Θ.*

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

$$|\phi_n'''(t)| \le \phi_n''(t) \max_{i\in[n]} |\langle X_i, \theta_1 - \theta_0\rangle|, \tag{23}$$

$$|\phi'''(t)| \le \phi''(t) \sup_{x\in\mathcal{X}} |\langle x, \theta_1 - \theta_0\rangle|. \tag{24}$$

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

$$|\phi_z'''(t)| \le 2[\phi_z''(t)]^{3/2}, \quad \forall z \in \mathcal{Z}, \tag{25}$$

$$|\phi_n'''(t)| \le \phi_n''(t)\left[\max_{i\in[n]} \phi_{Z_i}''(t)\right]^{1/2}, \tag{26}$$

$$|\phi'''(t)| \le \phi''(t)\left[\sup_{z\in\mathcal{Z}} \phi_z''(t)\right]^{1/2}. \tag{27}$$

The next two propositions, whose proofs follow [Nes13], will allow us to quantify the change of the second derivative of the restriction of the loss to a straight line by using self-concordance.

Proposition 3.2. *Suppose $g : \mathbb{R} \to \mathbb{R}$ is differentiable, non-negative, and for some $c \ge 0$ satisfies*

$$|g'(t)| \le 2c[g(t)]^{3/2}, \quad \textit{for any} \quad t \in \mathbb{R}^{(+)} : c|t|\sqrt{g(0)} \le 1.$$

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

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

Proposition 3.3. *Suppose $g : \mathbb{R} \to \mathbb{R}$ is differentiable, non-negative, and for some $c \ge 0$ satisfies*

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

Then, for any $t \in T$ it holds

$$g(0)e^{-c|t|\sqrt{g(0)}} \le g(t) \le g(0)e^{c|t|\sqrt{g(0)}}.$$

The next proposition summarizes the local properties of multivariate functions whose restrictions to line segments behave essentially as pseudo self-concordant functions in Case **(a)**, or in a similar manner but with a weaker control of the third derivative in Cases **(b)** and **(c)**. Case **(a)** is a straightforward extension of [Bac10, Proposition 1], and is sufficient for pseudo self-concordant losses; the remaining cases are useful when dealing with canonically self-concordant losses.

Proposition 3.4. *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*

$$|\phi_F'''(t)| \le S\phi_F''(t), \quad 0 \le t \le 1.$$

Then,

$$F(\theta_1) - F(\theta_0) - \nabla F(\theta_0)^\top(\theta_1 - \theta_0) \le \frac{e^S - S - 1}{S^2}\|\theta_1 - \theta_0\|^2_{\mathbf{H}_0}, \tag{28}$$

$$F(\theta_1) - F(\theta_0) - \nabla F(\theta_0)^\top(\theta_1 - \theta_0) \ge \frac{e^{-S} + S - 1}{S^2}\|\theta_1 - \theta_0\|^2_{\mathbf{H}_0}. \tag{29}$$

(b) *Suppose that $\phi_F(t)$ satisfies, instead,*

$$|\phi_F'''(t)| \le \frac{S\phi_F''(t)}{1 - t}, \quad 0 \le t < 1.$$

Then,

$$F(\theta_1) - F(\theta_0) - \nabla F(\theta_0)^\top(\theta_1 - \theta_0) \le \frac{1}{2 - S}\|\theta_1 - \theta_0\|^2_{\mathbf{H}_0}, \tag{30}$$

$$F(\theta_1) - F(\theta_0) - \nabla F(\theta_0)^\top(\theta_1 - \theta_0) \ge \frac{1}{2 + S}\|\theta_1 - \theta_0\|^2_{\mathbf{H}_0}, \tag{31}$$

where (30) holds whenever $S < 2$.

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

$$|\phi_F'''(t)| \le \frac{S\phi_F''(t)}{1 - St}, \quad 0 \le t < 1/S,$$

Then,

$$F(\theta_{1/S}) - F(\theta_0) - \frac{1}{S}\nabla F(\theta_0)^\top(\theta_1 - \theta_0) \le \frac{\|\theta_1 - \theta_0\|^2_{\mathbf{H}_0}}{S^2}, \tag{32}$$

$$F(\theta_{1/S}) - F(\theta_0) - \frac{1}{S}\nabla F(\theta_0)^\top(\theta_1 - \theta_0) \ge \frac{\|\theta_1 - \theta_0\|^2_{\mathbf{H}_0}}{3S^2}. \tag{33}$$

The next proposition describes the behavior of such functions close to the optimum. In Case (a), it has already been proved in [Bac10, Proposition 2].

Proposition 3.5. *Suppose that the premise of Proposition 3.4 holds with fixed θ_0, all $\theta_1 \in \Theta$, and some $W \in \mathbb{R}^d$ which can depend on θ_1. If for any $W = W(\theta_1)$ it holds*

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

where one can take $c = 1/2$ in Cases (a)–(b) and $c = 1/4$ in Case (c), function $F(\theta)$ has a unique global 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 3.5 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 us to localize $\widehat\theta_n$ around θ_*, using that $\|\nabla L_n(\theta_*)\|^2_{\mathbf{H}^{-1}}$ decreases at the rate $O(d_{\text{eff}}/n)$ due to the i.i.d. assumption.

3.2 Preliminary results under boundedness assumptions

In this section, we present extensions of the asymptotic deviation bound ($\star$) to the finite-sample regime. In the proofs, we use some technical results related to subgaussian distributions, namely, deviation bounds for the quadratic forms of subgaussian random vectors, Theorem A.1, and for their sample covariance matrices, Theorem A.2. These technical results, as well as some useful corollaries from them, are collected in Appendix A. In what follows, we first prove simplified results, in which the design is assumed almost surely bounded, rather than just subgaussian. This simplifies the control of the average risk using Proposition 3.1, allowing to take supremum over z in (24) and (27). In Section 3.3, we will extend these results by dropping the boundedness assumptions. The bounds on the critical sample size will be improved later on in Section 3.4 under a strengthened version of Assumption D2. This will require a more subtle analysis relying upon some new ideas. Hence, the results presented next could be considered as preliminary ones.

Pseudo self-concordant losses. Let us first treat the case of pseudo self-concordant losses.

Theorem 3.1. *Let Assumptions SCa, B0, D1, D2, and C hold, and let for some $0 < \delta \le 1$,*

$$n \gtrsim \max\left\{K_2^4\left(d + \log(1/\delta)\right),\ \rho K_1^2 \mathfrak{B}_0^2 d_{\mathit{eff}} \log(e/\delta)\right\}. \tag{34}$$

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

$$\|\nabla L_n(\theta_*)\|^2_{\mathbf{H}^{-1}} \lesssim \frac{K_1^2 d_{\mathit{eff}} \log(e/\delta)}{n}, \tag{35}$$

and

$$L(\widehat\theta_n) - L(\theta_*) \lesssim \|\widehat\theta_n - \theta_*\|^2_{\mathbf{H}} \lesssim \|\nabla L_n(\theta_*)\|^2_{\mathbf{H}^{-1}}. \tag{36}$$

Proof. $\mathbf{0^o}$. Recall that $\mathbf{H} = \nabla^2 L(\theta_*)$, and let $\mathbf{H}_n := \nabla^2 L_n(\theta_*)$ be the empirical Hessian at the true optimum θ_*. 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 $1-\delta$,

$$\frac{1}{2}\mathbf{H} \preccurlyeq \mathbf{H}_n \preccurlyeq 2\mathbf{H}. \tag{37}$$

On the other hand, we can prove (35) using Assumption D1. Indeed, for $i \in [n]$, random vectors

$$\nabla \ell_{Z_i}(\theta_*) = \ell'(Y_i, X_i^\top \theta_*) X_i$$

are mutually independent, have zero mean and covariance $\mathbf{G}$. Hence, $\mathbf{G}^{-1/2}\nabla \ell_{Z_i}(\theta_*)$ are independent and isotropic (have zero mean and unit covariance). Moreover, by Assumption D1, we have $\|\mathbf{G}^{-1/2}\nabla \ell_{Z_i}(\theta_*)\|_{\psi_2} \le K_1$. Hence, by Lemma A.4 about the subgaussian norm of the sum of independent subgaussian random vectors, we have that the random vector V_n, given by

$$V_n := \sqrt{n}\mathbf{G}^{-1/2}\nabla L_n(\theta_*),$$

is an isotropic random vector satisfying $\|V_n\|_{\psi_2} \le CK_1$ for some constant C, and, moreover,

$$\|\nabla L_n(\theta_*)\|_{\mathbf{H}^{-1}}^2 = \frac{\|V_n\|_{\mathbf{M}}^2}{n}, \quad \mathbf{M} = \mathbf{G}^{1/2}\mathbf{H}^{-1}\mathbf{G}^{1/2}. \tag{38}$$

We further have $\|\mathbf{M}\|_\infty \le \|\mathbf{M}\|_2 \le \mathrm{tr}(\mathbf{M}) = d_{\mathrm{eff}}$. Applying Theorem A.1, we arrive at (35).

$\mathbf{1^o}$. Our next step is to obtain the right inequality in (36), by applying Proposition 3.5 to $L_n(\theta)$ at $\theta_0 = \theta_*$. Invoking the bound (23) of Proposition 3.1, and using the Cauchy-Schwarz inequality, we see that $L_n(\theta)$ falls into Case **(a)** of Proposition 3.4 at $\theta_0 = \theta_*$ with $\mathbf{H}_0 = \mathbf{H}_n$, $W(\theta) = X_j$ where $j \in \mathrm{Argmax}_{i\in[n]} |\langle X_i, \theta - \theta_*\rangle|$, and

$$\|W(\theta)\|_{\mathbf{H}_0^{-1}} \le \max_{i\in[n]} \|X_i\|_{\mathbf{H}_n^{-1}} \le \max_{x\in\mathcal{X}} \|x\|_{\mathbf{H}_n^{-1}}.$$

Hence, we can apply Proposition 3.5 with $\tilde{\theta} = \widehat{\theta}_n$. That is, whenever

$$\max_{x\in\mathcal{X}} \|x\|_{\mathbf{H}_n^{-1}}^2 \|\nabla L_n(\theta_*)\|_{\mathbf{H}_n^{-1}}^2 \le c, \tag{39}$$

for some absolute constant c (we can take, e.g. $c = 1/4$), we have $\|\widehat{\theta} - \theta_*\|_{\mathbf{H}_n}^2 \le C\|\nabla L_n(\theta_*)\|_{\mathbf{H}_n^{-1}}^2$, which, when combined with (37), implies the desired bound:

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

On the other hand, again using (37) and Assumptions B0 and C, we have that with probability at least $1-\delta$,

$$\max_{x\in\mathcal{X}} \|x\|_{\mathbf{H}_n^{-1}}^2 \le 2\rho\mathfrak{B}_0^2,$$

whence for (39) it suffices that

$$\rho\mathfrak{B}_0^2 \|\nabla L_n(\theta_*)\|_{\mathbf{H}^{-1}}^2 \le c. \tag{40}$$

But this holds due to (35) and the second bound in (34). The right inequality in (36) is proved.

$\mathbf{2^o}$. To prove the left inequality in (36), we apply case **(a)** of Proposition 3.4, namely (28), to $L(\theta)$ with $\theta_0 = \theta_*$, $\theta_1 = \widehat{\theta}_n$, $\mathbf{H}_0 = \mathbf{H}$, and $W = W(\theta) \in \mathrm{Argmax}_{x\in\mathcal{X}} |\langle x, \theta - \theta_*\rangle|$, so that $\|W(\theta)\|_{\mathbf{H}^{-1}} \le \sqrt{\rho}\mathfrak{B}_0$.[7] This is possible due to the bound (24) of Proposition 3.1, and using the Cauchy-Schwarz inequality as in $\mathbf{1^o}$. As such, noting that $\nabla L(\theta_*) = 0$, we arrive at

$$L(\widehat{\theta}_n) - L(\theta_*) \le \frac{e^{\sqrt{\rho}\mathfrak{B}_0 r} - 1 - \sqrt{\rho}\mathfrak{B}_0 r}{\rho \mathfrak{B}_0^2},$$

where $r = \|\widehat{\theta}_n - \theta_*\|_{\mathbf{H}}^2$. Finally, note that $\sqrt{\rho}\mathfrak{B}_0 r \lesssim 1$ as guaranteed by (34), (35), and the right inequality of (36). Since $f(u) = e^u - 1 - u \lesssim u^2$ whenever $u \lesssim 1$ (e.g., $f(u) \le u^2$ when $u \le 1$), we obtain the left inequality of (36). ■

Remark 3.1. *From the proof we can easily see that in the case of a well-specified MLE estimator, we can replace* $d_{\mathit{eff}} \log(e/\delta)$ *with* $(\sqrt{d} + \sqrt{\log(1/\delta)})^2$*, both in* (34) *and* (35).

Canonically self-concordant losses. We now present a counterpart of Theorem 3.1 for canonically self-concordant losses, i.e., those satisfying Assumption SCb. The crucial difference of the following result from Theorem 3.1 is the absence of the curvature parameter ρ. This improvement is achieved by carefully exploiting Assumption SCb while working with individual losses. This allows to reduce the situation to Case **(c)** of Propositions 3.1 and 3.4, where the role of W is now played by the *calibrated* design $\widetilde{X}$. However, the bound for the sufficient sample size is still inflated by the radius of the decorrelated design (now the calibrated one).

Theorem 3.2. *Let Assumptions SCb, B2, D1, D2 hold. Then,* (35)–(36) *are satisfied whenever*

$$n \gtrsim \max\left\{K_2^4\left(d + \log\left(1/\delta\right)\right),\ K_1^2\mathfrak{B}_2^2 d_{\mathit{eff}} \log\left(e/\delta\right)\right\}. \tag{41}$$

Proof. Note that step $\mathbf{0^o}$ completely repeats that of the previous proof, hence (35) and (37) remain valid. On the other hand, to prove (36) under (41), we will use Case **(b)** of Proposition 3.1.

$\mathbf{1^o}$. For any $\theta \in \Theta$ different from θ_*, let $\theta_t = \theta_* + t(\theta - \theta_*)$ for $0 \le t < 1$. Due to (25), we can apply Proposition 3.2 to $g(t) = \phi_z''(t)$ for any $z = (x, y) \in \mathcal{Z}$, taking $c = 1$. Thus, denoting $\widetilde{x} := [\ell''(y, x^\top\theta_*)]x$ for arbitrary $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we have

$$\phi_z''(t) \le \frac{\phi_z''(0)}{(1 - t\sqrt{\phi_z''(0)})^2} = \frac{\langle \widetilde{x}, \theta - \theta_*\rangle^2}{(1 - t|\langle \widetilde{x}, \theta - \theta_*\rangle|)^2}, \tag{42}$$

holding for any $t \ge 0$ for which the denominator is positive. Combining this with (26), we have

$$|\phi_n'''(t)| \le \phi_n''(t) \max_{i\in[n]} \frac{|\langle \widetilde{X}_i, \theta - \theta_*\rangle|}{1 - t|\langle \widetilde{X}_i, \theta - \theta_*\rangle|} = \phi_n''(t) \frac{|\langle \widetilde{X}_j, \theta - \theta_*\rangle|}{1 - t|\langle \widetilde{X}_j, \theta - \theta_*\rangle|}, \tag{43}$$

where $j = j(\theta) \in \mathrm{Argmax}_{i\in[n]} |\langle \widetilde{X}_i, \theta - \theta_*\rangle|$, and we can take any $t \ge 0$ for which the denominator in the right-hand side is positive. Thus, $L_n(\theta)$ falls into case **(c)** of Proposition 3.4 with $\theta_0 = \theta_*$, $\mathbf{H}_0 = \mathbf{H}_n$, and $W = \widetilde{X}_j$. Moreover, using (37), we have

$$\|\widetilde{X}_j\|_{\mathbf{H}_n^{-1}} \le \max_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \|\widetilde{x}\|_{\mathbf{H}_n^{-1}} \le \sqrt{2} \max_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \|\widetilde{x}\|_{\mathbf{H}^{-1}} = \sqrt{2}\mathfrak{B}_2,$$

where the second inequality holds with probability at least $1 - \delta$ due to (37). Hence, we can apply Proposition 3.5 to $L_n(\theta)$ at $\theta_0 = \theta_*$, which allows to obtain the right inequality in (36), proceeding in the same way as in part $\mathbf{1^o}$ of the proof of Theorem 3.1.

$\mathbf{2^o}$. We now prove the left inequality in (36). Similarly to (43), from (42) and (27) we obtain

[7] Hereinafter, we slightly abuse the notation by ignoring zero-probability subsets of $\mathcal{Z}$ when maximizing over it.

$$|\phi'''(t)| \le \phi''(t)\frac{|\langle W, \theta - \theta_*\rangle|}{1 - t|\langle W, \theta - \theta_*\rangle|}, \tag{44}$$

where $W = \widetilde{x}(x, y)$ for $(x, y) \in \operatorname{Argmax}_{\mathcal{X}\times\mathcal{Y}} |\langle \widetilde{x}, \theta - \theta_*\rangle|$. Thus, we have $\|W\|_{\mathbf{H}^{-1}} \le \mathfrak{B}_2$, and

$$|\langle W, \widehat{\theta}_n - \theta_*\rangle| \le \mathfrak{B}_2 \|\widehat{\theta}_n - \theta_*\|_{\mathbf{H}} \le 1$$

by (35), the right inequality in (36), and the second bound in (41). As such, the average risk $L(\widehat{\theta}_n)$ satisfies the premise of Case **(c)** of Proposition 3.4 with $\theta_1 = \widehat{\theta}_n$ and $S \le 1$. But this implies that Case **(b)** is also satisfied. Thus, denoting $r := \|\widehat{\theta}_n - \theta_*\|_{\mathbf{H}}$, and using (30), we obtain

$$L(\widehat{\theta}_n) - L(\theta_*) \le \frac{r^2}{2 - \mathfrak{B}_2 r} \le r^2.$$

Thus, the theorem is proved. ■

Discussion. Recall from the discussion in Section 2.2 that both $\mathfrak{B}_0$ and $\mathfrak{B}_2$ cannot be smaller than $O(\sqrt{d})$, hence, both bounds (34), (41) on the critical sample size must grow proportionally to the product of d_{eff} and d. As we will make sure in the next section, in the case of unbounded (but subgaussian) design vectors X and $\widetilde{X}$, parameters $\mathfrak{B}_0$ and $\mathfrak{B}_2$ indeed can be replaced with $O(\sqrt{d})$, so that the critical sample size can be bounded as $O(d \cdot d_{\mathrm{eff}})$ or $O(\rho \cdot d \cdot d_{\mathrm{eff}})$ in the case of canonical or pseudo self-concordant losses. As such, we see that the actual difference between the critical sizes in the case of pseudo/canonically self-concordant losses is in the curvature parameter ρ, which is absent in the latter case. Moreover, this difference is preserved in the improved bounds on the critical sample size, be presented in Section 3.4.

3.3 Extension to unbounded design

Next, we extend Theorems 3.1 and 3.2, dropping the almost-sure boundedness assumptions. While the main ideas are the same, the proof is more technical and is therefore placed into Appendix B. Essentially, the main difficulty is that we cannot apply Proposition 3.4 to $L(\theta)$ anymore, since the suprema in the right-hand sides of (24) and (27) can potentially be infinite. The issue can be circumvented by restricting the (calibrated) design vector is to its confidence set through rejection sampling. Note that such confidence sets are readily given under the subgaussian assumptions from Section 2.2. Namely, for any $0 < \delta \le 1$ consider the events

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

Using the tools from Appendix A, we can show that $\mathbb{P}(\mathcal{E}_0) \ge 1 - \delta$ under Assumption D0, and $\mathbb{P}(\mathcal{E}_2) \ge 1 - \delta$ under Assumption D2. Now, let us replace $L(\theta)$ with the *restricted risks*:

$$L_{\mathcal{E}_0}(\theta) := \mathbf{E}[\ell_Z(\theta)\mathbb{1}_{\mathcal{E}_0}(X)]; \quad L_{\mathcal{E}_2}(\theta) := \mathbf{E}[\ell_Z(\theta)\mathbb{1}_{\mathcal{E}_2}(\widetilde{X})]$$

where $\mathbb{1}_{\mathcal{E}_0}(X) := \mathbb{1}\{X \in \mathcal{E}\}$ and similarly for $\mathbb{1}_{\mathcal{E}_2}(\widetilde{X})$. This allows to exclude from averaging the low-probability outcomes in which the norms of X and $\widetilde{X}$ are too large, and work with X and $\widetilde{X}$ as if they were bounded. On the other hand, using Assumptions D1 and D2, we can demonstrate that $\nabla L_{\mathcal{E}_0}(\theta_*) \approx 0$ and $\nabla^2 L_{\mathcal{E}_0}(\theta_*) \approx \mathbf{H}$, and similarly for $L_{\mathcal{E}_2}(\cdot)$, provided that δ is small enough. Combining these ideas leads to the following result proved in Appendix B.

Theorem 3.3. *Let Assumptions SCa, D0, D1, D2, and C hold. Whenever*

$$n \gtrsim \max\left\{K_2^4\left(d + \log(1/\delta)\right),\ K_1^2 K_0^2 \rho d_{\mathit{eff}}\, d \log(ed/\delta)\right\}, \tag{45}$$

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

$$\|\nabla L_n(\theta_*)\|^2_{\mathbf{H}^{-1}} \lesssim \frac{K_1^2 d_{\mathit{eff}} \log\left(e/\delta\right)}{n}, \tag{46}$$

$$\|\widehat{\theta}_n - \theta_*\|^2_{\mathbf{H}} \lesssim \|\nabla L_n(\theta_*)\|^2_{\mathbf{H}^{-1}}. \tag{47}$$

Moreover, one has

$$L_{\mathcal{E}_0}(\widehat{\theta}_n) - L_{\mathcal{E}_0}(\theta_*) \lesssim \frac{K_1^2 d_{\mathit{eff}} \log\left(e/\delta\right)}{n} \tag{48}$$

whenever δ *satisfies*

$$\delta \lesssim \min\left\{\left(\frac{1}{\sqrt{n \log(e d_{\mathit{eff}})}}\right)^{q(d_{\mathit{eff}})}, \left(\frac{1}{K_2^2 d \log(ed)}\right)^{q(d)}\right\}, \tag{49}$$

where $q(t) = 1 + 1/\log(t)$.

As we see, Theorem 3.3 includes an extra technical condition (49) on the minimal violation probability δ. This condition is mild, as the admissible δ depends polynomially on n and d.

Theorem 3.2 can also be extended to the case of unbounded design, in a similar manner. Below we present such an extension. Its proof closely follows that of Theorem 3.3, and is omitted.

Theorem 3.4. *Let Assumptions SCb, D1, D2 hold together with* (49)*. Then,* (46)*–*(48) *are satisfied, with* $L_{\mathcal{E}_2}(\cdot)$ *instead of* $L_{\mathcal{E}_0}(\cdot)$*, whenever*

$$n \gtrsim \max\left\{K_2^4\left(d + \log\left(1/\delta\right)\right),\; K_1^2 K_2^2 d_{\mathit{eff}}\, d \log\left(ed/\delta\right)\right\}. \tag{50}$$

Discussion: extension to heavy-tailed observations. As mentioned in Section 2.2, Assumptions D0–D2 are quite restrictive. Let us point out some possibilities for relaxing them.

- One possibility is to reject X with large $\|\cdot\|_{\mathbf{\Sigma}}$-norm, namely with $\|X\|^2_{\mathbf{\Sigma}} \gtrsim C_\delta \rho d$ for some C_δ depending (polynomially) on $1/\delta$, when computing the estimator. This allows to replace Assumption D0 in Theorem 3.3 with the existence of some finite moments of the marginals of X_0, using the technique from steps $\mathbf{3^o}$–$\mathbf{4^o}$ of the proof (see also [Ver11, Section 1.3]), replacing Hölder's inequality with Cauchy-Schwarz. The price to pay is somewhat worse threshold for the admissible δ compared to (49). We can relax Assumptions D1 and D2 similarly, provided that $\ell'(y,\eta)$ and $\ell''(y,\eta)$ are uniformly bounded from above, which holds in logistic regression and in robust estimation. Such bounds are required since we cannot perform rejection sampling directly on $\nabla \ell_Z(\theta_*) = \ell'(Y, X^\top\theta_*)X$ and $\widetilde{X} = \ell''(Y, X^\top\theta_*)X$.

- Another possibility is to use the "confidence-boosting" technique based on a version of the multi-dimensional sample median as suggested 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 via Algorithm 3 of [HS16], 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 given by $\widehat{\theta}^{(\widehat{i})}$ 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 $\widehat{\theta}^{(i)}$ will admit a fixed-probability version of (46), say, with $\delta = 2/3$. On the other hand, for each $i \in [k]$ with fixed probability we will also have

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

This follows from our analysis in Theorems 3.1–3.2, which yields $\frac{1}{2}L(\theta_*) \preccurlyeq L_n(\theta_*) \preccurlyeq 2L(\theta_*)$, and an integration argument that allows to show that $\frac{1}{2}L_n(\theta_*) \preccurlyeq L_n(\widehat{\theta}^{(i)}) \preccurlyeq 2L_n(\theta_*)$, cf. Lemmas B.1–B.3. Finally, the estimators over different subsamples are mutually independent. Thus, we can apply Theorem 11 of [HS16], which yields the desired bound (48). This technique also allows to weaken Assumptions D0 and D2, replacing the subgaussian norm $\|\cdot\|_{\psi_2}$ with the subexponential norm $\|\cdot\|_{\psi_1}$ at the expense of a logarithmic factor.[8] This can be done by replacing Theorem A.2 with [Ver12, Theorem 5.48] and controlling the quantities $\mathbf{E}[\max_{i\in[n]} \|X_i\|^2_{\mathbf{H}}]$, $\mathbf{E}[\max_{i\in[n]} \|\widetilde{X}_i\|^2_{\mathbf{H}}]$ via Bernstein's inequality (Theorem A.1).

3.4 Improved bounds under stronger local assumptions

As we demonstrate next, the critical sample size obtained in Sections 3.2–3.3 can actually be improved: essentially, the product of d_{eff} and d can be replaced with $\max(d_{\text{eff}}, d\log d)$. To obtain these improvements, we have to introduce an extended version of Assumption D2.

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

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

where $\mathbf{H}(\theta)$ denotes its covariance matrix, for any θ in the Dikin ellipsoid $\Theta_r(\theta_*)$, as 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 extended assumption is still *local*, i.e., it only concerns the points r-close to θ_* 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 3.5. *Assume SCb, D1, and D2* with $r \gtrsim 1$. Then, (35)–(36) hold whenever*

$$n \gtrsim \max\left\{\bar{K}_2^4(r) d \log\left(\frac{ed}{\delta}\right),\ K_1^2\bar{K}_2^6(r) d_{\mathit{eff}} \log\left(\frac{e}{\delta}\right)\right\}. \tag{51}$$

Before presenting the full proof of this result, let us briefly explain the main ideas behind it. First of all, let us recall where the extra factor d in the bound of Theorem 3.4 comes from. This factor 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 3.4. 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 *average risk* following

[8] One of the equivalent definitions of $\|\cdot\|_{\psi_1}$-norm, see [Ver12, Section 5.2.4], is as follows: $X \in \mathbb{R}^d$ satisfies $\|X\|_{\psi_1} \le K$ if for any $\forall u \in \mathcal{S}^d$ one has $(\mathbf{E}[|\langle X, u\rangle|^p])^{1/p} \lesssim Kp$, compared to $K\sqrt{p}$ for $\|\cdot\|_{\psi_2}$, cf. Lemma A.1.

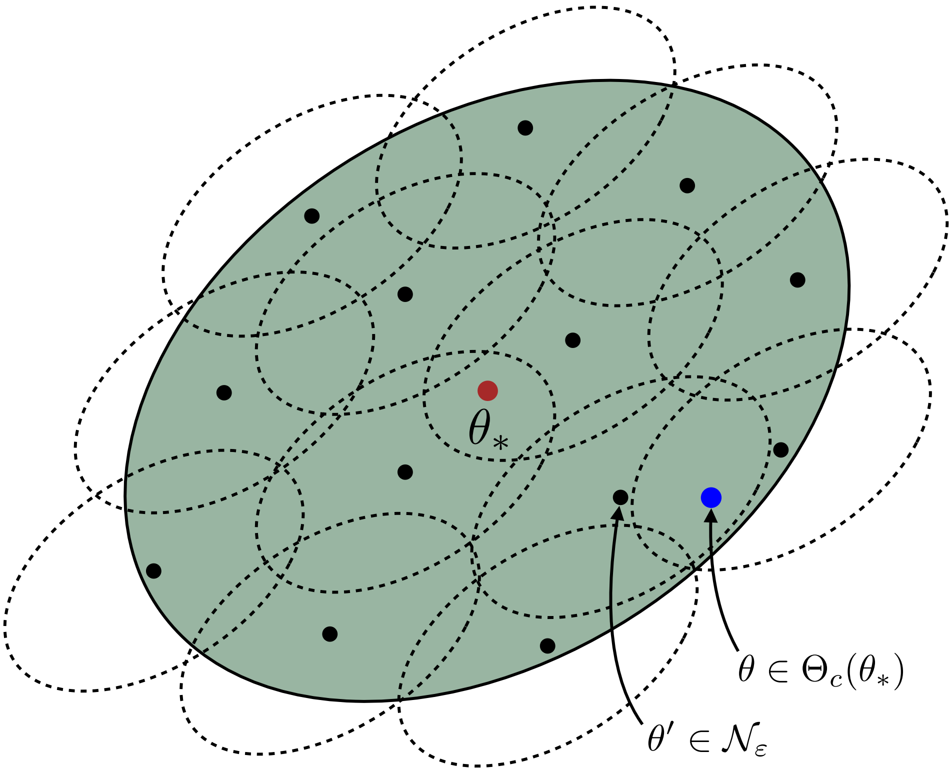

Figure 2: Crucial step in the proof of Theorem 3.5 – uniform approximation of the empirical Hessians $\mathbf{H}_n(\theta)$ in the constant-radius Dikin ellipsoid $\Theta_c(\theta_*)$ (in green). Using Assumption D2*, we first prove that $\mathbf{H}(\theta) \approx \mathbf{H}(\theta_*)$, up to a constant factor, for any $\theta \in \Theta_c(\theta_*)$. On the other hand, self-concordance of the individual losses can be used to obtain a constant-order approximation of $\mathbf{H}_n(\cdot)$ within a smaller ellipsoid with radius $O(1/d^\kappa)$, for some $\kappa \geq 1/2$, 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 $\varepsilon = O(1/d^\kappa)$. This is done using Theorem A.2. As a result, we obtain that $\mathbf{H}_n(\theta) \approx \mathbf{H}(\theta_*)$ uniformly over $\theta \in \Theta_c(\theta_*)$.

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 $O(1/d^\kappa)$ for some $\kappa \geq 1/2$.[9]

Next we state a counterpart of this result in the case of pseudo self-concordant losses. As we might expect, the critical sample size in this case increases by the factor of ρ.

Theorem 3.6. *Assume SCa, D0, D1, C, and D2* with $r \gtrsim 1/\sqrt{\rho}$. Then, (35)–(36) hold whenever*

$$n \gtrsim \max\left\{ \bar{K}_2^4(r) d \log\left(\frac{ed}{\delta}\right),\ \rho K_0^2 K_1^2 \bar{K}_2^4(r) d_{\mathit{eff}} \log\left(\frac{e}{\delta}\right) \right\}. \tag{52}$$

The proof of Theorem 3.6 is very similar to that for Theorem 3.5, and is given in Appendix B. Before proceeding with the proof of Theorem 3.5, let us discuss the results.

- First, note that in the case of pseudo self-concordance, the radius of the Dikin ellipsoid in which Assumption D2* is required to be satisfied is $\sqrt{\rho}$ times smaller than in the case of canonical self-concordance. As it will become clear from the proof of Theorem 3.6, this

[9] Namely, we choose $\kappa = 2$, rather than $\kappa = 1/2$. This simplifies probabilistic calculations in step $\mathbf{3^o}$ of the proof, and does not influence (51) since d^κ only enters underneath the logarithm, see also Remark 3.2 below.

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 $\bar{K}_2(r)$: as we show in Appendix C, in logistic regression with Gaussian design $X \sim \mathcal{N}(0, \mathbf{\Sigma})$ one has

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

so that with $r = 1/\sqrt{\rho}$ Assumption D2* is *equivalent* to Assumption D2 in this case.

- Note that the second threshold in (51) has the additional $\bar{K}_2^4(r)$ factor compared to that in (50) if we do not distinguish between $\bar{K}_2(r)$ and $K_2 = \bar{K}_2(0)$, and similarly when comparing (52) and (45). In fact, this can be a substantial difference since K_2 and $\bar{K}_2(r)$ can depend on the unknown θ_*. For example, in Appendix C (Proposition C.1), we show that in logistic regression with Gaussian design $X \sim \mathcal{N}(0, \mathbf{\Sigma})$, one has $\rho \lesssim (1 + \|\theta_*\|_{\mathbf{\Sigma}})^3$, this bound being tight, while the bound on $\bar{K}_2(1/\sqrt{\rho})$ is

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

  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 for 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 3.5–3.6:

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

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

Proof (of Theorem 3.5). $\mathbf{1^o}$. 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

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

cf. the proof of Proposition 3.1. In particular, putting $\widetilde{X}(\theta_t) := [\ell''(Y, \langle X, \theta_t\rangle)]^{1/2} X$, we have

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

where in the final step we used the definition of $\mathbf{H}(\theta_t)$. On the other hand, due to Assumption SCb,

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

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,

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

for some absolute constant $C>0$. Without the loss of generality we can assume that $C\ge 1$ by weakening the inequality in the opposite case. 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 3.2 to $g(t)=\phi''(t)$, putting

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

and arriving at

$$\frac{\phi''(0)}{(1+\bar{c}t\sqrt{\phi''(0)})^2} \le \phi''(t) \le \frac{\phi''(0)}{(1-\bar{c}t\sqrt{\phi''(0)})^2}, \quad 0\le \bar{c}t[\phi''(0)]^{1/2}\le 1.$$

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

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

In particular, we have

$$\frac{4}{9}\mathbf{H}(\theta_*) \preccurlyeq \mathbf{H}(\theta) \preccurlyeq 4\mathbf{H}(\theta_*), \quad \theta\in\Theta_{1/(2\bar{c})}(\theta_*). \tag{53}$$

2^o. 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

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

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 $\theta_1\in\Theta$. Due to (25), we can apply Proposition 3.2 to the second derivative $\phi_Z''(t)$ of the individual loss $\phi_Z(t) := \ell(Y, X^\top\theta_t)$. This results in

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

cf. (42). 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

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

Using the Cauchy-Schwarz inequality together with (53), this can be weakened to

$$\frac{\ell''(Y, X^\top \theta_0)}{(1 + 2t\|\mathbf{H}(\theta_0)^{-1/2}\widetilde{X}(\theta_0)\|_2\|\Delta\|_{\mathbf{H}(\theta_*)})^2} \le \ell''(Y, X^\top \theta_t) \le \frac{\ell''(Y, X^\top \theta_0)}{(1 - 2t\|\mathbf{H}(\theta_0)^{-1/2}\widetilde{X}(\theta_0)\|_2\|\Delta\|_{\mathbf{H}(\theta_*)})^2}.$$

where $t \ge 0$ is such that the denominator on the right-hand side is strictly positive. As a result,

$$\frac{\ell''(Y, X^\top \theta)}{(1 + 2\|\mathbf{H}(\theta)^{-1/2}\widetilde{X}(\theta)\|_2\|\theta' - \theta\|_{\mathbf{H}(\theta_*)})^2} \le \ell''(Y, X^\top \theta') \le \frac{\ell''(Y, X^\top \theta)}{(1 - 2\|\mathbf{H}(\theta)^{-1/2}\widetilde{X}(\theta)\|_2\|\theta' - \theta\|_{\mathbf{H}(\theta_*)})^2} \tag{54}$$

for any $\theta \in \Theta_{1/(2\bar{c})}(\theta_*)$, and any θ' such that the denominator in the right-hand side is positive.

3

$\mathbf{3^o}$. 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 Euclidean ball with radius $1/(2\bar{c})$ with respect to the standard Euclidean norm. Hence, we can apply the bound for covering numbers of Euclidean balls: for any $\varepsilon \le 1$,

$$|\mathcal{N}_\varepsilon| \le \left(\frac{3}{2\bar{c}\varepsilon}\right)^d. \tag{55}$$

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 $\mathbf{1^o}$ of the proof of Theorem 3.3 (to account for the fact that the vectors are not centered), we can show that with probability at least $1 - \delta$,

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

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

$$\sup_{i\in[n],\, \theta_0\in\mathcal{N}_\varepsilon} \|\mathbf{H}(\theta)^{-1/2}\widetilde{X}_i(\theta)\|_2 \le C_2\bar{K}_2\sqrt{d\log\left(\frac{en|\mathcal{N}_\varepsilon|}{\delta}\right)} \le C_2\bar{K}_2 d\sqrt{\log\left(\frac{3en}{\delta\varepsilon}\right)}, \tag{56}$$

with probability at least $1 - \delta$, where in the second transition we used (55). Now, let us choose

$$\varepsilon = \frac{1}{64C_2^2\bar{K}_2^2 d^2 \log(en/\delta)}. \tag{57}$$

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 (54) and (56), 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

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

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

$$\frac{4}{9}\mathbf{H}_n(\pi_*(\theta)) \preccurlyeq \mathbf{H}_n(\theta) \preccurlyeq 4\mathbf{H}_n(\pi_*(\theta)), \quad \forall \theta \in \Theta_{1/(2\bar{c})}(\theta_*), \tag{58}$$

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 $1-\delta$,

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

provided that $n \gtrsim \bar{K}_2^4(d + \log(1/\delta))$. Taking the union bound over $\mathcal{N}_\varepsilon$, and using (55) and (57), we see that

$$\frac{1}{2}\mathbf{H}(\theta) \preccurlyeq \mathbf{H}_n(\theta) \preccurlyeq 2\mathbf{H}(\theta), \quad \forall \theta \in \mathcal{N}_\varepsilon \tag{59}$$

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

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

By simple algebra, it suffices that

$$n \gtrsim \bar{K}_2^4 d \log\left(\frac{e}{\delta}\right). \tag{60}$$

Combining (58), (59), and (53), we see that the sample size satisfying (60) guarantees uniform approximation of empirical Hessians on the Dikin ellipsoid $\Theta_{1/(2\bar{c})}(\theta_*)$: with probability $\geq 1-\delta$,

$$0.09\,\mathbf{H}(\theta_*) \preccurlyeq \mathbf{H}_n(\theta) \preccurlyeq 32\,\mathbf{H}(\theta_*), \quad \forall \theta \in \Theta_{1/(2\bar{c})}(\theta_*). \tag{61}$$

$\mathbf{4^o}$. With (61) at hand, we can localize the estimate through a similar argument to that in Proposition 3.5, 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 (61) is equivalent to

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

Integrating this twice on $[0,1]$, we arrive at

$$\frac{0.09\phi''(0)t^2}{2} \leq \phi_n(t) - \phi_n(0) - \phi_n'(0)t \leq \frac{32\phi''(0)t^2}{2}.$$

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

$$0.045\|\theta - \theta_*\|^2_{\mathbf{H}(\theta_*)} \leq L_n(\theta) - L_n(\theta_*) - \langle \nabla L_n(\theta_*), \theta - \theta_* \rangle \leq 16\|\theta - \theta_*\|^2_{\mathbf{H}(\theta_*)}. \tag{62}$$

cf. (32)–(33). Now we can proceed as in the proof of Proposition 3.5, Case (**c**). Namely, consider the event $\widehat{\theta}_n \notin \Theta_{1/(2\bar{c})}(\theta_*)$. Under this event, there exists a point $\bar{\theta}_n \in [\theta_*, \widehat{\theta}_n]$ 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 (62), 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 $\mathbf{0^o}$ of the proof of Theorem 3.1) that with probability $\geq 1-\delta$.

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

Thus, whenever

$$n \gtrsim K_1^2 \bar{K}_2^6 d_{\mathrm{eff}} \log(e/\delta),$$

we have a contradiction, and $\widehat{\theta}_n$ must belong to $\Theta_{1/(2\bar{c})}(\theta_*)$. Then, (62) applied at $\theta = \widehat{\theta}_n$ yields

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

It only remains to control the excess average risk. This can be done by recalling the result (53), which translates to

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

Integrating this twice on $[0,1]$, we obtain

$$\frac{4\phi''(0)t^2}{9} \le \phi(t) - \phi(0) \le 4\phi''(0)t^2,$$

where we used that $\phi'(0) = \nabla L(\theta_*) = 0$. The upper bound leads to $L(\theta) - L(\theta_*) \le \|\theta - \theta_*\|^2_{\mathbf{H}(\theta_*)}$ for any $\theta \in \Theta_{1/2\bar{c}}(\theta_*)$. But we have already proved that this holds for $\widehat{\theta}_n$ with high probability. ■

Remark 3.2. *One fact that played a key role in the proof of the theorem is that the confidence interval enters* additively *into the bound for the sample size when estimating a covariance matrix:*

$$n \gtrsim K^2(d + \log(e/\delta)).$$

Hence, we can simultaneously estimate exponential *number of covariance matrices from $O(d)$ observations. In our case, this number is $d^{O(d)}$, which results in extra logarithmic factor in* (51).

3.5 Results in the 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 of the results of Section 3 in the sense that the ambient dimension d would be replaced with $\mathtt{s}$, and d_{eff} with the quantity $\text{tr}(\mathbf{H}_{\mathcal{S}}^{-1}\mathbf{G}_{\mathcal{S}})$ where $\mathbf{G}_{\mathcal{S}} = \mathbf{E}[\ell'(Y, X_{\mathcal{S}}^\top\theta_*)X_{\mathcal{S}}X_{\mathcal{S}}^\top]$ and $\mathbf{H}_{\mathcal{S}} = \mathbf{E}[\ell''(Y, X_{\mathcal{S}}^\top\theta_*)X_{\mathcal{S}}X_{\mathcal{S}}^\top]$. However, in reality $\mathcal{S}$ is unknown, and the common recommendation is to use the ℓ_1-penalized M-estimator, given by

$$\widehat{\theta}_{\lambda,n} \in \underset{\theta\in\mathbb{R}^d}{\text{Argmin}}\, L_n(\theta) + \lambda\|\theta\|_1. \tag{63}$$

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 3.1–3.4 in the high-dimensional setting, which give the quadratic dependence of the critical sample size from the sparsity index. 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 3.5–3.6, which would have resulted in a near-linear, rather than quadratic, dependence of the critical sample size from sparsity. We believe that such extension is possible, and leave it for future work.

Before stating the results, we introduce an extra assumption complimentary to Assumption C.

Assumption C*. *Design is uncorrelated: $\mathbf{\Sigma} = \mathbf{I}$. Moreover, for some positive $\varkappa_1$, $\varkappa_2$, it holds*

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

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

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

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 in order.

- Similarly to the case of ρ, we always have the following bounds on $\varkappa_1$ 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 similar bound (21) for ρ. For example, they allow to bound $\varkappa$ with a constant in robust estimation and logistic regression.

- Correlated designs can also be considered, but this would lead to the inflation of the bounds by the condition number of $\mathbf{\Sigma}$.

The result presented next characterizes the statistical properties of the ℓ_1-penalized M-estimator defined in (63) in the case of canonically self-concordant losses, extending Theorem 3.3.

Theorem 3.7. *Assume SCa, D0, D1, D2, C, C*, and* $|\theta_*|_0 \le \mathtt{s}$.

1. *Whenever*

   $$n \gtrsim \max\left\{\rho\varkappa_2 K_2^4 \mathtt{s}\log\left(\frac{ed}{\delta}\right),\ \rho^2\varkappa_1 K_0^2 K_1^2 \mathtt{s}^2 \log\left(\frac{edn}{\delta}\right)\right\}, \tag{64}$$

   and the regularization parameter satisfies

   $$K_1\sqrt{\frac{\varkappa_1\log(ed/\delta)}{n}} \lesssim \lambda \lesssim \frac{1}{\rho K_0 \mathtt{s}\sqrt{\log(edn/\delta)}}, \tag{65}$$

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

   $$\|\widehat{\theta}_{\lambda,n} - \theta_*\|_1 \lesssim \rho\mathtt{s}\lambda, \quad \|\widehat{\theta}_{\lambda,n} - \theta_*\|_{\mathbf{H}}^2 \lesssim \rho\mathtt{s}\lambda^2. \tag{66}$$

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

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

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

   $$L_{\mathcal{E}}(\widehat{\theta}_{\lambda,n}) - L_{\mathcal{E}}(\theta_*) \lesssim \rho\mathtt{s}\lambda^2. \tag{67}$$

Discussion. Clearly, the right choice of λ is the one attaining the lower bound in (65), that is,

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

up to a constant factor (note that this choice is always possible since the left-hand side in (65) is always upper-bounded with the right-hand side due to the second bound in (64). With such λ, both the prediction error and the (restricted) excess risk $L_{\mathcal{E}}(\widehat{\theta}_{\lambda,n}) - L_{\mathcal{E}}(\theta_*)$ are bounded with

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

whenever $n \gtrsim \max(Q\mathtt{s}, \rho Q \mathtt{s}^2)\log(ed/\delta)$, where we ignored the dependence on the subgaussian constants. Thus, we conclude that in the case of pseudo self-concordant losses, ℓ_1-penalization allows to replace d and d_{eff} with $\mathtt{s}$, at the expense of inflating all bounds by $O(Q\log d)$.

We now state an analogue of Theorem 3.7 in the case of canonically self-concordant losses. Its proof is similar, and is outlined in Appendix B.

Theorem 3.8. *Assume SCb, D1, D2, C, C*, and* $|\theta_*|_0 \le \mathtt{s}$.

1. *Whenever*

$$n \gtrsim \max\left\{\rho\varkappa_2 K_2^4 \mathtt{s}\log\left(\frac{ed}{\delta}\right),\ \rho^2\varkappa_1\varkappa_2 K_1^2 K_2^2 \mathtt{s}^2 \log\left(\frac{edn}{\delta}\right)\right\}, \tag{68}$$

and the regularization parameter satisfies

$$K_1\sqrt{\frac{\varkappa_1\log(ed/\delta)}{n}} \lesssim \lambda \lesssim \frac{1}{\rho K_2 \mathtt{s}\sqrt{\varkappa_2 \log(edn/\delta)}}, \tag{69}$$

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

$$\|\widehat{\theta}_{\lambda,n} - \theta_*\|_1 \lesssim \rho\mathtt{s}\lambda, \quad \|\widehat{\theta}_{\lambda,n} - \theta_*\|_{\mathbf{H}}^2 \lesssim \rho\mathtt{s}\lambda^2. \tag{70}$$

2. *Define the event* $\mathcal{E} := \{\|\widetilde{X}\|_\infty \lesssim K_2\sqrt{\varkappa_2\log(ed/\delta)}\}$. *Then,* $\mathbb{P}(\mathcal{E}) \ge 1-\delta$, *and whenever*

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

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

$$L_{\mathcal{E}}(\widehat{\theta}_{\lambda,n}) - L_{\mathcal{E}}(\theta_*) \lesssim \rho\mathtt{s}\lambda^2. \tag{71}$$

Comparison of Theorems 3.7 and 3.8. We see that the usual gain of ρ that we observed so far when replacing pseudo self-concordance with canonical self-concordance is not preserved in the case of ℓ_1-regularized estimators. Instead, the second bound in (64) and the upper bound in (65) get inflated with $\varkappa_2$, and the critical sample size, given the "ideal" choice of the regularization parameter corresponding to the lower bound in (69), 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 case without further assumptions.

Comparison with the prior work. Our result in Theorem 3.7 extends the result of [Bac10, Theorem 5] for logistic regression with fixed design, obtained using 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 $\mathbf{s}$ 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 requires a sample of size $\Omega(\mathbf{s}d)$.

On the other hand, our results can be compared to those in [vdGM12] who establish the rate $O(\lambda\mathbf{s})$ for the ℓ_1-error and $O(\lambda^2 s)$ 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, [vdGM12] assumes 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, the results of [vdGM12] do not address the case of unbounded parameter. Remarkably, these results have a similar requirement for the sample size to scale as $\Omega(\mathbf{s}^2\log d)$.

Proof (of Theorem 3.7). $\mathbf{0^o}$. 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 (65) – 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

$$L_n(\widehat{\theta}) - L_n(\theta_*) \le \lambda(\|\theta_*\|_1 - \|\widehat{\theta}\|_1). \tag{72}$$

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,

$$\|\theta_*\|_1 - \|\widehat{\theta}\|_1 \le \|\Delta_{\mathcal{S}}\|_1 - \|\Delta_{\mathcal{S}_c}\|_1. \tag{73}$$

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

$$L_n(\widehat{\theta}) - L_n(\theta_*) \ge -\|\nabla L_n(\theta_*)\|_\infty \|\widehat{\theta} - \theta_*\|_1 \ge -\|\nabla L_n(\theta_*)\|_\infty(\|\Delta_{\mathcal{S}}\|_1 + \|\Delta_{\mathcal{S}_c}\|_1). \tag{74}$$

Collecting (72)–(74), we get $(\lambda - \|\nabla L_n(\theta_*)\|_\infty)\,\|\Delta_{\mathcal{S}_c}\|_1 \le (\lambda + \|\nabla L_n(\theta_*)\|_\infty)\,\|\Delta_{\mathcal{S}}\|_1$. Whence if

$$\lambda \ge 2\|\nabla L_n(\theta_*)\|_\infty, \tag{75}$$

we have that Δ satisfies the restricted subspace condition:

$$\|\Delta_{\mathcal{S}_c}\|_1 \le 3\|\Delta_{\mathcal{S}}\|_1, \tag{76}$$

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

$$\|\Delta\|_1 \le 4\sqrt{s}\|\Delta\|_2. \tag{77}$$

Later on, we will show that the lower bound in (65) implies (75) with probability at least $1-\delta$. For now, let us assume that (75) holds.

$\mathbf{1^o}$. To localize the estimate, we now use a similar technique to the one used in the proof of Proposition 3.5, but replace the Cauchy-Schwarz inequality with Young's inequality. First, applying (29) to $L_n(\theta)$ with $\theta_0 = \theta_*$, $\theta_1 = \widehat{\theta}$, and $W = X_j$ for some (random) $j \in [n]$, we have

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

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 (72) and (73), bounding $-\langle \nabla L_n(\theta_*), \Delta\rangle$ via Young's inequality, and using (75), we get

$$\frac{e^{-\|X_j\|\|\Delta\|_1} - 1 + \|X_j\|_\infty \|\Delta\|_1}{\|X_j\|_\infty^2 \|\Delta\|_1^2} \|\Delta\|_{\mathbf{H}_n}^2 \le \frac{3\lambda\|\Delta\|_1}{2}. \tag{78}$$

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

$$\frac{\|\Delta\|^2}{\rho} \le \|\Delta\|_{\mathbf{H}}^2 \le \varkappa\|\Delta\|^2$$

for any Δ satisfying the restricted subspace property (76) – 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

$$\frac{1}{2}\|\Delta\|_{\mathbf{H}}^2 \preccurlyeq \|\Delta\|_{\mathbf{H}_n}^2 \preccurlyeq 2\|\Delta\|_{\mathbf{H}}^2, \tag{79}$$

for any Δ satisfying (76), provided that

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

cf. (64). Combining this result with

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

where we used (77), we obtain that under (64) with probability $1-\delta$ it holds

$$\|\Delta\|_{\mathbf{H}_n}^2 \ge \frac{\|\Delta\|_1^2}{32\rho\mathbf{s}}. \tag{80}$$

Combining this with (78), and denoting $\mathfrak{B}_{\sup} := \max_{i\in[n]} \|X_i\|_\infty$ and $u := \mathfrak{B}_{\sup}\|\Delta\|_1$, we obtain

$$e^{-u} - 1 + u \le 48\rho\mathbf{s}\lambda\mathfrak{B}_{\sup} u.$$

From now on, we can proceed as in the proof of Proposition 3.5, cf. (95). That is, whenever

$$48\rho\mathbf{s}\lambda\mathfrak{B}_{\sup} \le 1/2, \tag{81}$$

we sequentially obtain $u \le 2$, $e^{-u} - 1 + u \ge \frac{u^2}{4}$, and $u \le 192\rho\mathbf{s}\mathfrak{B}_{\sup}\lambda$, and finally arrive at

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

This is the first inequality in (66), and the second one is obtained by combining it with (78)–(79). Thus, both inequalities in (66) are satisfied under the two assumed conditions (75) and (81). It remains to show that these conditions are indeed guatanteed to be satisfied with high probability under (65). 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

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

Via Lemma A.4, this gives $\|\nabla L_n(\theta_*)\|_{\psi_2} \lesssim K_1\sqrt{\varkappa_1/n}$; in particular, $\|[\nabla L_n(\theta_*)]_i\|_{\psi_2} \lesssim K_1\sqrt{\varkappa_1/n}$ componentwise for any $i \in [n]$. Whence, by Lemma A.2, with probability at least $1-\delta$ it holds

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

This guarantees (75) under the lower bound in (65). Similarly, we can show that with probability at least $1-\delta$,

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

which guarantees (81) under the upper bound in (65). The first claim of the theorem is proved; note that the upper bound in (64) is simply a corrolary of (65).

2^o. To prove the second claim, we bound the excess average risk using a similar technique as when proving Theorem 3.3. To simplify the presentation, we defer the proof to Appendix B. ■

Remark 3.3. *In the above analysis (as well as in the proof of Theorem 3.8 in Appendix), the matrices* $\mathbf{H}$ *and* $\mathbf{H}_n$ *only interact with the residual* Δ, *which with high probability satisfies the restricted subspace condition* (76). *Hence, we can strengthen the result, replacing Assumption C and the second inequality of Assumption C* with the requirement that*

$$\frac{1}{\rho}\|\Delta\|_2^2 \le \|\Delta\|_{\mathbf{H}}^2 \le \varkappa_2\|\Delta\|_2^2$$

in the case where $\Delta \in \mathbb{R}^d$ *is* approximately sparse, *i.e., satisfies* $\|\Delta - [\Delta]_{\mathsf{s}}\|_1 \le 3\|[\Delta]_{\mathsf{s}}\|_1$, *where* $[\Delta]_{\mathsf{s}}$ *is the projection of* Δ *to the span of its* s *largest coordinates. This observation can be exploited to accelerate computation of the estimator* (63) *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 with size* $m \ll n$.

4 Conclusion and perspectives

We have demonstrated how to obtain the asymptotically optimal rates $O(d\log(1/\delta)/n)$ for the excess risk of M-estimators in finite-sample regimes. Our analysis encompasses M-estimators with losses satisfying certain self-concordance-type assumptions; these include some generalized linear models (notably, logistic regression) as well as some robust estimation models. Such assumptions allow to control the precision of the local quadratic approximations of empirical risk through a simple integration technique. However, self-concordance alone only allows to address the large-sample regime $n = \Omega(d^2)$. In order to handle the moderate sample size regime $n = \Omega(d)$, we conduct a somewhat non-standard analysis in which self-concordance is combined with a covering argument, allowing to control the uniform deviations of the Hessians of empirical risk in a Dikin ellipsoid of the best predictor. We also study ℓ_1-regularized M-estimators in high-dimensional regimes, showing that in this case d essentially gets replaced with the number of non-zero components s of the best predictor whenever $n = \Omega(\mathsf{s}^2)$ up to a logarithmic factor.

One question left untouched is the practical performance of M-estimators with canonically self-concordant losses. In principle, it is possible that the comparative advantage for such losses predicted by our theory is hardly observed empirically, e.g., due to the curvature parameter ρ being moderate in practice. Other possible directions for future research are discussed further.

Improved results in the high-dimensional setup. For ℓ_1-regularized estimators, we have only showed a suboptimal result that requires $n = \Omega(\mathsf{s}^2)$ up to a logarithmic factor. Improving this sample size bound to a near-linear one is a long-standing open problem, see, e.g, [vdGM12]. We believe that with some technical work, our covering argument in Section 3.4 could be extended to ℓ_1-regularized estimators, resulting in a near-linear boundary for the admissible sample size.

Nonparametric regression. One possible direction for extending our results is the nonparametric setting where d and d_{eff} can be infinite, and θ lives in a Hilbert space $\mathcal{H}$ with norm $\|\cdot\|_{\mathcal{H}}$. Naturally, M-estimator should be replaced with its regularized counterpart, a unique solution to

$$\min_{\theta \in \mathcal{H}} L_n(\theta) + \lambda \|\theta\|_{\mathcal{H}}^2$$

for some $\lambda > 0$. For pseudo self-concordant losses, [Bac10] has already obtained analogues of our basic results presented in Section 3.2, showing the optimal bound for the bias and variance of the regularized estimator in the regime $n \gtrsim \rho \mathtt{df}_1^2$, where $\mathtt{df}_p := \operatorname{tr}[\mathbf{H}^p(\mathbf{H} + \lambda \mathbf{I})^{-p}]$ is the ℓ_p-number of degrees of freedom that replaces d in the nonparametric setting. However, without additional conditions on the eigenvalues of $\mathbf{H}$, these results lead to a suboptimal bias-variance balance because of the restriction $n \gtrsim \rho \mathtt{df}_1^2$. As such, the optimal bias-variance tradeoff is likely to be implied by an extension of our refined results (cf. Section 3.4) to the nonparametric setting.

Efficient algorithms. It would also be interesting to revisit algorithmically efficient procedures such as stochastic approximation, varianced reduction techniques, and quasi-Newton methods. In particular, one could be interested in extending Hessian-sketching procedures from least-squares linear regression to general M-estimators with (pseudo) self-concordant losses. On the other hand, the stand-of-the-art stochastic-approximation-type algorithm for logistic regression in [BM13] as well relies on Hessian approximation, and is also worth revisiting in this connection.

Matrix-parametrized models. In this work, we did not investigate M-estimators with matrix-valued design and predictors, arising, for example, in covariance matrix estimation and independent component analysis. In some of them, one commonly uses the log-determinant loss which is self-concordant in the sense of [NN94]. Our techniques could shed light on the statistical performance of such estimators in finite-sample regimes with random measurements.

Acknowledgements

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

A Probabilistic tools

A.1 Subgaussian distributions

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

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

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 $t \ge 0$,*

$$\mathbb{P}\left\{ |\xi| > t \right\} \le \exp\left(1 - \frac{ct^2}{\sigma^2} \right).$$

2. Subgaussian moments: for any $p \ge 1$,

$$\mathbf{E}[|\xi|^p]^{1/p} \le c\sigma\sqrt{p}.$$

Moreover, if $\mathbf{E}[\xi] = 0$, either of these properties is equivalent to the exponential moment bound:

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

Subgaussian norm can also be extended to $Z \in \mathbb{R}^d$ as the maximal $\|\cdot\|_{\psi_2}$-norm of all one-dimensional marginals of Z:

$$\|Z\|_{\psi_2} := \sup_{\theta \in \mathcal{S}^{d-1}} \|\langle Z, \theta\rangle\|_{\psi_2}, \tag{82}$$

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 $\|\mathbf{M}Z\|_{\psi_2} \le \|\mathbf{M}\|_\infty \|Z\|_{\psi_2}$ for arbitrary matrix $\mathbf{M}$. Some well-known properties of subgaussian random vectors are summarized in the following lemmata.

Lemma A.2. *Let the components 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 statement of the lemma follows from item 1 of Lemma A.1 by the union bound. ∎

The next lemma provides a simple bound for the p-th moment of $\|Z\|_\infty$. Although this bound is loose for any fixed p, we use it only in the regime $p \approx \log d$ in which it is tight.[10]

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

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

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

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

We obtain the claimed bound by extracting the p-th root and doing some simple estimates. ∎

Lemma A.4 (Hoeffding-type inequality, follows from [Ver12, Lemma 5.9] through (82)). *Let $Z_1, ..., Z_n \in \mathbb{R}^d$ be independent and zero-mean. Then,*

$$\left\|\sum_{i=1}^n Z_i\right\|_{\psi_2}^2 \lesssim \sum_{i=1}^n \|Z_i\|_{\psi_2}^2.$$

The following simple result allows to pass from a subgaussian vector to its centered version.

Lemma A.5 (Subgaussian norm after affine transform and decorrelation). *Suppose that $X \in \mathbb{R}^d$ satisfies $\mathbf{E}[X] = 0$, $\mathbf{\Sigma} := \mathbf{E}[XX^\top]$, 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*

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

[10] 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 [Pol90]. This requires some technical work, and is beyond the scope of this paper.

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

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

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, by the Sherman-Morrison formula,

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

Finally, note that $K \gtrsim 1$, as follows from the inequality $\mathbf{E}[\xi^4] \ge (\mathbf{E}[\xi^2])^2$ applied to $\xi = \langle u, X\rangle$, together with Item 2 of Lemma A.1. ■

A.2 Deviation bounds for quadratic forms of subgaussian random vectors

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

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

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

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

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

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*

$$|\langle u, Z\rangle| \lesssim \|u\|_2 K\sqrt{\log(e/\delta)}. \tag{83}$$

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

The next result immediately follows from Theorem A.1 using that $\|\mathbf{M}\|_\infty \le \|\mathbf{M}\|_2 \le \mathrm{tr}(\mathbf{M})$.

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

$$\mathbb{P}\left\{\frac{\zeta}{\sqrt{\mathrm{tr}(\mathbf{M})}} \ge cK(1+t)\right\} \le \exp(-t^2),$$

and, as a consequence,

$$\mathbb{P}\left\{\frac{\zeta}{cK\sqrt{\mathrm{tr}(\mathbf{M})}} \ge u\right\} \le \exp\left(c_1 - \frac{u^2}{2c_2}\right),$$

so that

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

A.3 Sample covariance matrices

Next we describe the behavior of the sample second-moment matrix of a subgaussian vector.

Theorem A.2 ([Ver12, Theorem 5.39]). *Assume that* $\widetilde{X} \in \mathbb{R}^d$ *satisfies* $\mathbf{E}[\widetilde{X}\widetilde{X}^\top] = \mathbf{H}$ *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*

$$\tfrac{1}{2}\|\Delta\|_{\mathbf{H}}^2 \le \|\Delta\|_{\mathbf{H}_n}^2 \le 2\|\Delta\|_{\mathbf{H}}^2, \quad \forall \Delta \in \mathbb{R}^d. \tag{84}$$

What follows next is an extension of this result to the high-dimensional and sparse 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 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*

$$(1/\rho)\|\Delta\|_2^2 \le \|\Delta\|_{\mathbf{H}}^2 \le \varkappa\|\Delta\|_2^2.$$

Then, whenever

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

it holds that with probability at least $1 - \delta$, *for any* $\Delta \in \mathbb{R}^d$ *satisfying the same condition one has*

$$\tfrac{1}{2}\|\Delta\|_{\mathbf{H}}^2 \le \|\Delta\|_{\mathbf{H}_n}^2 \le 2\|\Delta\|_{\mathbf{H}}^2.$$

B Deferred proofs

B.1 Properties of pseudo-Huber loss (18)

We can check that the Fenchel dual of $\phi : (-1, 1) \to \mathbb{R}$ defined in (17) is indeed $\varphi(t)$, cf. (18), 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, by the known property of Fenchel-dual functions,

$$\phi'(\varphi'(t)) = t, \quad \forall t \in \mathbb{R}.$$

Differentiating this, we obtain

$$\phi''(\varphi'(t)) \cdot \varphi''(t) = 1. \tag{85}$$

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 $\phi(u)$ satisfies $|\phi'''(u)| \le c[\phi''(u)]^{3/2}$, we get that $|\phi'''(u)| \le c[\phi''(u)]^{3/2}$ using (85). ■

B.2 Proof of Proposition 3.1

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 p-th derivative of $\phi(t) = L(\theta_t)$ and $\phi_n(t) = L_n(\theta_t)$, are given by

$$\phi_Z^{(p)}(t) = \ell^{(p)}(Y, \langle X, \theta_t\rangle) \cdot \langle X, \Delta\rangle^p, \tag{86}$$
$$\phi^{(p)}(t) = \mathbf{E}[\ell^{(p)}(Y, \langle X, \theta_t\rangle) \cdot \langle X, \Delta\rangle^p], \tag{87}$$
$$\phi_n^{(p)}(t) = \frac{1}{n}\sum_{i\in[n]} \ell^{(p)}(Y_i, \langle X_i, \theta_t\rangle) \cdot \langle X_i, \Delta\rangle^p. \tag{88}$$

This holds for $p \le 3$ due to the basic smoothness assumption. Now, let Assumption SCa be satisfied. Using (87) with $p \in \{2,3\}$, we get

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

arriving at (24). Analogously, we obtain (23) from (88), replacing $\mathcal{X}$ with the set $\{X_1, ..., X_n\}$. Let Assumption SCb be satisfied instead, then (25) is obvious from (86). On the other hand,

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

which amounts to (27). We obtain (26) analogously, replacing the expectation with the sum. ■

B.3 Proof of Proposition 3.2

We first treat the case $g(0) > 0$. Denote

$$T_0 = \left[-\frac{1}{c\sqrt{g(0)}}, \frac{1}{c\sqrt{g(0)}}\right],$$

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 $|\psi'(t)| \le c$. Now, let $t \in T_0$ be positive without loss of generality. Integrating from 0 to t, we get

$$-ct \le \frac{1}{\sqrt{g(t)}} - \frac{1}{\sqrt{g(0)}} \le ct.$$

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, $g(0) > 0$ actually 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

$$-c(t_2 - t_1) \le \frac{1}{\sqrt{g(t_2)}} - \frac{1}{\sqrt{g(t_1)}} \le c(t_2 - t_1),$$

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

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

Taking the limit $t_1 \to t_0$, by the continuity of $g(t)$ we obtain a contradiction with $g(t_0) = 0$. ■

B.4 Proof of Proposition 3.3

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 3.2. 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 which, we arrive at the claim. ■

B.5 Proof of Proposition 3.4

We first treat the one-dimensional situation, proving analogues of Proposition 3.2 in all cases.

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 $S \ge 0$,*

$$|g'''(t)| \le Sg''(t), \quad 0 \le t \le 1.$$

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

$$\frac{e^{-St} + St - 1}{S^2} g''(0) \le g(t) - g(0) - g'(0)t \le \frac{e^{St} - St - 1}{S^2} g''(0), \quad 0 \le t \le 1. \tag{89}$$

Proof. First assume that $g''(t) > 0$ on $[0,1]$. Then, the premise of the lemma implies that for any $t \in [0,1]$,

$$-S\mathrm{d}t \le \mathrm{d} \log g''(t) \le S\mathrm{d}t.$$

Integrating this, we obtain:

$$g''(0)e^{-St} \le g''(t) \le g''(0)e^{St}. \tag{90}$$

Two more integrations successively give

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

and then (89). 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 (90). ■

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 $S \ge 0$,*

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

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

$$\frac{(1-t)^{2+S} + (2+S)t - 1}{(1+S)(2+S)} g''(0) \le g(t) - g(0) - g'(0)t \le \frac{(1-t)^{2-S} + (2-S)t - 1}{(1-S)(2-S)} g''(0), \quad (91)$$

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

$$\frac{g''(0)}{2+S} \le g(1) - g(0) - g'(0) \le \frac{g''(0)}{2-S}.$$

Proof. Without loss of generality, we assume that $g''(t) > 0$; the general case can be treated as in Lemma B.1. The proof follows the same steps as in Lemma B.1. The first integration gives

$$(1-t)^S g''(0) \le g''(t) \le (1-t)^{-S} g''(0), \quad (92)$$

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

$$\frac{1-(1-t)^{1+S}}{1+S} g''(0) \le g'(t) - g'(0) \le \frac{1-(1-t)^{1-S}}{1-S} g''(0),$$

and then (91). Finally, when $t = 1$, the term $(1-S)$ vanishes from the denominator of the right-hand side of (91), 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 $S \ge 0$,*

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

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

$$\left(\frac{t^2}{2} - \frac{St^3}{6}\right) g''(0) \le g(t) - g(0) - g'(0)t \le \frac{St + (1-St)\log(1-St)}{S^2} g''(0). \quad (93)$$

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

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

Proof. We again assume w.l.o.g. that $g''(t) > 0$, and integrate three times. obtaining

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

then

$$\left(t - \frac{St^2}{2}\right) g''(0) \le g'(t) - g'(0) \le \left(-\frac{\log(1-St)}{S}\right) g''(0), \quad (94)$$

and finally (93). Note that we can take $t = 1/S$ in (93) by continuity of $f(u) = u \log u$ at 0. ■

Proof of the proposition. Cases (a)–(c) follow from Lemmata B.1–B.3 applied to $g(t) = \phi_F(t)$, 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 will be exploited in Proposition 3.5. ■

B.6 Proof of Proposition 3.5

Note that from (90), (92), or (94), 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 θ_1 be arbitrary point from $\Theta_F(F(\theta_0))$, and let $r = \|\theta_1 - \theta_0\|_{\mathbf{H}_0}$. Denote $\nu := \|\nabla F(\theta_0)\|_{\mathbf{H}_0^{-1}}$ and $R := \|W\|_{\mathbf{H}_0^{-1}}$; note that $S \le Rr$. We now separately consider all cases of Proposition 3.4.

Case (a). By (29), we have

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

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

$$e^{-u} - 1 + u \le \nu R u. \tag{95}$$

By the premise, we know that $\nu R \le 1/2$, hence

$$e^{-u} - 1 + \frac{u}{2} \le 0.$$

We can check numerically that this implies $u \le 2$. We can also check that for such u, it holds

$$e^{-u} - 1 + u \ge \frac{u^2}{4}.$$

Plugging this back into (95), 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). By (31), we have

$$\begin{aligned} F(\theta_1) &\ge F(\theta_0) + \langle \nabla F(\theta_0), \theta_1 - \theta_0 \rangle + \frac{1}{2 + Rr} r^2 \\ &\ge F(\theta_0) - \nu r + \frac{1}{2 + Rr} r^2, \end{aligned}$$

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

$$\frac{u}{u + 2} \le \nu R.$$

Since $\nu R \le 1/2$, we have $u \le 2$. Then, we arrive at $u \le 4\nu R$, that is, $r \le 4\nu$ as required.

Case (c). First assume that $Rr \ge S \ge 1$. Then, $\theta_{1/S}$ belongs to the segment $[\theta_0, \theta_1]$ and to Θ. Whence $F(\theta_{1/S}) \le F(\theta_0)$ by the convexity of $\Theta_F(F(\theta_0))$. On the other hand, from (33) we have

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

Whence

$$\nu \ge \frac{r}{3S} \ge \frac{1}{3R},$$

and we arrive at the contradiction. Thus, $S < 1$, but for such S, the premise of Case (c) implies that of Case (b). ■

B.7 Proof of Theorem 3.3

$\mathbf{1^o}$. Denote $\mu = \mathbf{E}[X]$, and let $\boldsymbol{\Sigma}_o := \mathbf{E}[(X-\mu)(X-\mu)^\top]$. Note that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_o + \mu\mu^\top$. Denoting $\mathbf{Q} = \boldsymbol{\Sigma}_o^{1/2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_o^{1/2}$, we have

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

By construction, $\boldsymbol{\Sigma}_o^{-1/2}(X_i-\mu)$ is an isotropic random vector; moreover, $\|\boldsymbol{\Sigma}_o^{-1/2}(X_i-\mu)\|_{\psi_2} \lesssim K_0$ due to Assumption D0 combined with Lemma A.5. Note that $\|\mathbf{Q}\|_2 \le \operatorname{tr}(\mathbf{Q}) \le d$, and $\|\mathbf{Q}\|_\infty \le 1$. Hence, by Theorem A.1, with probability at least $1-\delta$ it holds

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

The second term can be controlled as follows:

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

where the last inequality holds with probability at least $1-\delta$ by Corollary A.1 Finally, we have

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

Combining these bounds, and taking the union bound, we get that with probability at least $1-\delta$, for any $i \in [n]$ it holds

$$\max_{i\in[n]}\|X_i\|^2_{\mathbf{H}_n^{-1}} \lesssim \rho K_0^2 d\log\left(\frac{en}{\delta}\right),$$

where we also used (37). Recalling the bound (35), and again making use of (37), we conclude that (39) holds under (45). This implies (46) and (47) by the same argument as in Theorem 3.1.

$\mathbf{2^o}$. Let us now prove (48). 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 (24) can now be replaced with

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

where $\mathcal{X}_{\mathcal{E}}$ is the confidence set for X under the event $\mathcal{E}_0$, namely,

$$\mathcal{X}_{\mathcal{E}} := \{x\in\mathcal{X} : \|x\|_{\mathbf{H}^{-1}} \le \sqrt{\rho}\mathfrak{B}\},$$

where $\mathfrak{B} := K_0\sqrt{d\log(e/\delta)}$, and we used Assumption C. Besides, let us assume, for time being, that the new Hessian $\mathbf{H}_{\mathcal{E}_0} := \nabla^2 L_{\mathcal{E}_0}(\theta_*)$ is invertible, and approximates the non-perturbed one:

$$c\mathbf{H} \preccurlyeq \mathbf{H}_{\mathcal{E}_0} \preccurlyeq C\mathbf{H} \tag{96}$$

for some pair of absolute constants $c, C > 0$. Later on, we will make sure that this is indeed true under the condition (49) on δ. Under this assumption, applying Proposition 3.4, case (a), to $L_{\mathcal{E}}(\theta)$ with $\theta_0 = \theta_*$, $\theta_1 = \widehat{\theta}_n$, $\mathbf{H}_0 = \mathbf{H}_{\mathcal{E}_0}$, and $W \in \mathcal{X}_{\mathcal{E}_0}$, we get by (28):

$$\begin{aligned}L_{\mathcal{E}_0}(\widehat{\theta}_n) - L_{\mathcal{E}_0}(\theta_*) &\lesssim \left(\frac{e^{\sqrt{\rho}\mathfrak{B}r} - 1 - \sqrt{\rho}\mathfrak{B}r}{\rho\mathfrak{B}^2r^2}\right)r^2 + \nabla L_{\mathcal{E}_0}(\theta_*)^\top(\widehat{\theta}_n - \theta_*) \\ &\lesssim r^2 + r\|\nabla L_{\mathcal{E}_0}(\theta_*)\|_{\mathbf{H}^{-1}},\end{aligned} \tag{97}$$

where in the last transition we acted as in the proof of Theorem 3.1, and used the results of $\mathbf{1^o}$.

$\mathbf{3^o}$. Comparing with the proof of Theorem 3.1, the novelty in (97) is the additional term depending on $\|\nabla L_{\mathcal{E}_0}(\theta_*)\|_{\mathbf{H}^{-1}}$. To control this term, let us introduce the *complementary* risk:

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

where $\mathcal{E}_0^c$ is the complement of $\mathcal{E}_0$, so that $\mathbb{P}(\mathcal{E}_0^c) \le \delta$. 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 similar to the one used in [Ver11, Section 1.3]. For any p, q such that $1/p + 1/q = 1$, we have by Hölder's inequality:

$$\|\nabla L_{\mathcal{E}_0^c}(\theta_*)\|_{\mathbf{H}^{-1}} \le \mathbf{E}[\|\nabla \ell_Z(\theta_*)\|_{\mathbf{H}^{-1}}\mathbb{1}_{\mathcal{E}_0^c}] \le \mathbf{E}[\|\nabla \ell_Z(\theta_*)\|_{\mathbf{H}^{-1}}^p]^{1/p}\delta^{1/q}, \tag{98}$$

Note that

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

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

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

Combining this with (97), (98), (46), (47), and choosing $p = \log(e d_{\mathrm{eff}})$, $q = 1 + 1/\log(d_{\mathrm{eff}})$, we obtain

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

Hence, for (48) it suffices that $\delta^{\frac{\log(d_{\mathrm{eff}})}{\log(d_{\mathrm{eff}})+1}}\sqrt{\log(d_{\mathrm{eff}})} \lesssim \sqrt{\log(e/\delta)/n}$, which follows from the first bound in (49).

$\mathbf{4^o}$. It remains to make sure that the Hessians $\mathbf{H}$ and $\mathbf{H}_{\mathcal{E}_0}$ are indeed close in the sense of (96). First of all, the upper bound in (96) is trivial. Indeed defining the complementary Hessian $\mathbf{H}_{\mathcal{E}_0^c} := \nabla^2 L_{\mathcal{E}_0^c}(\theta_*)$, we see that

$$\mathbf{H}_{\mathcal{E}_0} = \mathbf{H} - \mathbf{H}_{\mathcal{E}_0^c} \preccurlyeq \mathbf{H},$$

simply because $\mathbf{H}_{\mathcal{E}_0^c} \succcurlyeq 0$. On the other hand, the lower bound in (96) 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 (49), 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:

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

where in the last line 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 $K_2^2 p d\delta^{1/q} \lesssim 1$ under the second bound in (49). ■

B.8 Proof of Theorem 3.6

We use the same conventions as in the proof of Theorem 3.5. Besides, we assume w.l.o.g. that Assumption D2* holds with $r = 1/\sqrt{\rho}$, and let $\bar{K}_2 := \bar{K}_2(1/\sqrt{\rho})$ for brevity.

1^o. 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 the subgaussian constants $K_0, \bar{K}_2$. Fix $\theta_0 = \theta_*$ and arbitrary $\theta_1 \in \mathbb{R}^d$, let $\theta_t = \theta_0 + t(\theta_1 - \theta_0)$, $t \geq 0$, and let $\Delta := \theta_1 - \theta_0$. Putting $\widetilde{X}(\theta_t) := [\ell''(Y, \langle X, \theta_t \rangle)]^{1/2} X$ as before, we have

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

On the other hand, due to Assumption SCa,

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

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

$$\begin{aligned}
\mathbf{E}[\langle \mathbf{H}(\theta_t)^{-1/2} \widetilde{X}(\theta_t), \mathbf{H}(\theta_t)^{1/2} \Delta \rangle^4] &\leq C \bar{K}_2^4 \|\Delta\|^4_{\mathbf{H}(\theta_t)}, \\
\mathbf{E}[\langle \mathbf{\Sigma}^{-1/2} X, \mathbf{\Sigma}^{1/2} \Delta \rangle^2] &\leq C K_0^2 \|\Delta\|^2_{\mathbf{\Sigma}} \leq \rho C \bar{K}_0^2 \|\Delta\|^2_{\mathbf{H}(\theta_*)},
\end{aligned}$$

for some constant $C > 0$; moreover, we can enforce that $C > 1$ by weaking the bounds by a constant factor if it is not the case. Combining the above inequalities, we arrive at

$$|\phi'''(t)| \leq C^2 K_0 \bar{K}_2^2 [\rho \phi''(0)]^{1/2} \phi''(t), \quad 0 \leq t [\rho \phi''(0)]^{1/2} \leq 1.$$

Putting

$$\bar{c} := C K_0 \bar{K}_2^2, \tag{99}$$

and applying Proposition 3.3 to $g(t) = \phi''(t)$, we obtain that whenever $\bar{c}|t|\sqrt{\bar{\rho}\phi''(0)} \leq 1$, it holds

$$\phi''(0) \exp(-\bar{c}|t|\sqrt{\rho \phi''(0)}) \leq \phi''(t) \leq \phi''(0) \exp(\bar{c}|t|\sqrt{\rho \phi''(0)}).$$

Finally, since $\phi''(t) = \|\Delta\|^2_{\mathbf{H}(\theta_t)}$, this translates to the analogue of (53):

$$\frac{1}{e} \mathbf{H}(\theta_*) \preccurlyeq \mathbf{H}(\theta) \preccurlyeq e \mathbf{H}(\theta_*), \quad \theta \in \Theta_{\bar{r}}(\theta_*), \quad \bar{r} := \frac{1}{\bar{c}\sqrt{\bar{\rho}}}. \tag{100}$$

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

2^o. Let us now provide the local approximation of $\mathbf{H}_n(\theta)$ using pseudo self-concordance of individual losses. To this end, 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 3.2–3.3, we arrive at

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

which translates to $\ell''(Y, X^\top\theta_0)e^{-t|\langle X,\Delta\rangle|} \le \ell''(Y, X^\top\theta_t) \le \ell''(Y, X^\top\theta_0)e^{t|\langle X,\Delta\rangle|}$. Thus,

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

where we recycled the simplified notation $\mathbf{H} := \mathbf{H}(\theta_*)$. Equivalently, for any $\theta \in \Theta_{\bar r}(\theta_*)$ and $\theta' \in \Theta$,

$$\ell''(Y, X^\top\theta_0)\exp(-\|X\|_{\mathbf{H}^{-1}}\|\theta'-\theta\|_{\mathbf{H}}) \le \ell''(Y, X^\top\theta_t) \le \ell''(Y, X^\top\theta_0)\exp(\|X\|_{\mathbf{H}^{-1}}\|\theta'-\theta\|_{\mathbf{H}}). \tag{101}$$

By Assumption D0, random vector $\mathbf{\Sigma}^{-1/2}X$ has ψ_2-norm at most $\bar K_0$. Hence, repeating the argument from $\mathbf{1^o}$ in the proof of Theorem 3.3 we can show that with probability at least $1-\delta$,

$$\max_{i\in[n]} \|X_i\|_{\mathbf{H}^{-1}} \le C_0 K_0 \sqrt{\rho d \log\left(\frac{en}{\delta}\right)} \tag{102}$$

for some constant C_0.

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

$$\varepsilon = \frac{1}{C_0K_0\sqrt{\rho d\log(en/\delta)}}. \tag{103}$$

Combining this with (101) and (102), we obtain that with probability at most $1-\delta$,

$$\frac{1}{e}\mathbf{H}_n(\pi(\theta)) \preccurlyeq \mathbf{H}_n(\theta) \preccurlyeq e\mathbf{H}_n(\pi(\theta)), \quad \forall \theta \in \Theta_{\bar r}(\theta_*), \tag{104}$$

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

$$\frac{1}{2}\mathbf{H}(\theta) \le \mathbf{H}_n(\theta) \le 2\mathbf{H}(\theta), \quad \forall \theta \in \mathcal{N}_\varepsilon \tag{105}$$

with probability at least $1-\delta$, whenever $n \gtrsim d + \log(|\mathcal{N}_\varepsilon|/\delta)$. Recaling that $|\mathcal{N}_\varepsilon| \le (3\bar r/\varepsilon)^d$, 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 (99) and (103). Noting that $\bar K_2 \ge 1$, by simple algebra we have that (105) holds with probability at least $1-\delta$ under

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

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

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

where we combined (105) with (104) and (100).

$\mathbf{4^o}$. 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 $\mathbf{4^o}$ in the proof of Theorem 3.5, showing that (36) 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. (99). This leads to the second bound on the critical sample size from the premise of the theorem. ■

B.9 Proof of the second claim of Theorem 3.7

Note that $\mathbb{P}(\mathcal{E}) \geq 1-\delta$ by the results of $\mathbf{1^o}$. As in the proof of Theorem 3.3, let $\mathbf{H}_{\mathcal{E}} := \nabla^2 L_{\mathcal{E}}(\theta_*)$; recall that $\mathbf{H}_{\mathcal{E}} \preccurlyeq \mathbf{H}$. Applying (28) 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 $L_{\mathcal{E}_c}(\theta) := \mathbf{E}[\ell_Z(\theta) \mathbb{1}_{\mathcal{E}_c}(X)]$ 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

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

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

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

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

B.10 Proof of Theorem 3.8

$\mathbf{1^o}$. Let $\widehat{\theta} = \widehat{\theta}_{\lambda,n}$. Note that step $\mathbf{0^o}$ of the proof of Theorem 3.7 can be repeated *verbatim*. Thus, whenever

$$
\lambda \geq 2\|\nabla L_n(\theta_*)\|_\infty, \tag{106}
$$

we have

$$
L_n(\widehat{\theta}) - L(\theta_*) \leq \lambda(\|\Delta_{\mathcal{S}}\|_1 - \|\Delta_{\mathcal{S}_c}\|_1) \leq \lambda \|\Delta\|_1, \tag{107}
$$

$$
\|\Delta_{\mathcal{S}_c}\|_1 \leq 3\|\Delta_{\mathcal{S}}\|_1, \tag{108}
$$

$$
\|\Delta\|_1 \leq 4\sqrt{s}\|\Delta\|_2. \tag{109}
$$

Moreover, we know (cf. the end of step $\mathbf{1^o}$ of the proof of Theorem 3.7) that (106) holds with probability at least $1-\delta$ as long as

$$
\|\nabla L_n(\theta_*)\|_\infty \lesssim K_1 \sqrt{\frac{\varkappa_1 \log(ed/\delta)}{n}}. \tag{110}
$$

Hence, (106) and (110) are satisfied under the lower bound in (69). Finally, under (108) we have

$$
\frac{1}{2}\|\Delta\|_{\mathbf{H}}^2 \preccurlyeq \|\Delta\|_{\mathbf{H}_n}^2 \preccurlyeq 2\|\Delta\|_{\mathbf{H}}^2 \tag{111}
$$

and

$$
\|\Delta\|_{\mathbf{H}_n}^2 \geq \frac{\|\Delta\|_1^2}{32\rho\mathsf{s}}, \tag{112}
$$

both with probability at least $1-\delta$, whenever

$$
n \gtrsim \rho \varkappa_2 K_2^4 \mathsf{s} \log\left(\frac{ed}{\delta}\right).
$$

On the other hand, (78) does not hold since we cannot use (29). Instead, let us prove that

$$\frac{\|\Delta\|_{\mathbf{H}_n}^2}{1+3\|\widetilde{X}_j\|_\infty\|\Delta\|_1} \le L_n(\widehat{\theta}) - L(\theta_*) - \langle \nabla L_n(\theta_*), \Delta\rangle, \tag{113}$$

where $j \in \mathrm{Argmax}_{i\in[n]} |\langle \widetilde{X}_i, \Delta\rangle|$. Indeed, denote $S = |\langle \widetilde{X}_j, \Delta\rangle|$. Whenever $S \le 1$, function $L_n(\theta)$ satisfies Case (b) of Proposition 3.4, and we obtain (113) from (31). On the other hand, when $S \ge 1$ function $L_n(\theta)$ always satisfies Case (c) of Proposition 3.4, and we can use (33), i.e.,

$$\frac{\|\Delta\|_{\mathbf{H}_n}^2}{3S^2} \le L_n(\theta_{1/S}) - L(\theta_*) - \frac{1}{S}\langle \nabla L_n(\theta_*), \Delta\rangle, \tag{114}$$

where

$$\theta_{1/S} = \left(1 - \frac{1}{S}\right)\theta_* + \frac{1}{S}\widehat{\theta}$$

is a convex combination of θ_* and $\widehat{\theta}$. By convexity, we have $L_n(\theta_{1/S}) \le (1-\frac{1}{S})L_n(\theta_*) + \frac{1}{S}L_n(\widehat{\theta})$, whence $L_n(\widehat{\theta}) - L_n(\theta_*) \le (L_n(\widehat{\theta}) - L_n(\theta_*))/S$. When combined with (114), this results in

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

Whence (113) follows in this case as well by Young's inequality. Now, (113), (107), and (106) imply

$$\frac{\|\Delta\|_{\mathbf{H}_n}^2}{1+3\|\widetilde{X}_j\|_\infty\|\Delta\|_1} \le \frac{3\lambda\|\Delta\|_1}{2}, \tag{115}$$

which is an analogue of (78). Starting from this point, we can proceed in a similar way as in the proof of Theorem 3.8. Namely, let $\widetilde{\mathfrak{B}}_{\sup} := \|\widetilde{X}\|_\infty$ and $u := \widetilde{\mathfrak{B}}_{\sup}\|\Delta\|_1$, then (115) and (112) imply

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

Hence, whenever

$$48\rho\mathbf{s}\lambda\widetilde{\mathfrak{B}}_{\sup} \le 1/4, \tag{116}$$

we have $u \le 1$ and $u/(1+3u) \ge u/4$, which implies $u \le 192\rho\mathbf{s}\lambda\widetilde{\mathfrak{B}}_{\sup}$ and $\|\Delta\|_1 \le 192\rho\mathbf{s}\lambda$. This is the first inequality in (70). To obtain the second inequality, we combine (115) and (111). Thus, for (70) it remains to show that (116) holds under the upper bound in (69). We have

$$\|\widetilde{X}\|_{\psi_2} \le \|\mathbf{H}^{1/2}\|_2\|\mathbf{H}^{-1/2}\widetilde{X}\|_{\psi_2} \le K_2\sqrt{\varkappa_2},$$

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, in turn, guarantees (116) under the upper bound in (69).

$\mathbf{2^o}$. We now adapt the proof of the second claim of Theorem 3.7. Recall that in our case $\mathcal{E} := \{\|\widetilde{X}\|_\infty \lesssim K_2\sqrt{\varkappa_2 \log{(ed/\delta)}}\}$, and $\mathbb{P}(\mathcal{E}) \ge 1-\delta$ by the results of $\mathbf{1^o}$. 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 3.4 with $S \le \|\widetilde{X}\|_\infty\|\Delta\|_1 \le 1$, cf. $\mathbf{1^o}$. Thus, by (30) we have

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

Thence we proceed exactly in the same way as when proving the second claim of Theorem 3.7. ■

C Case study of the distribution assumptions in Section 2.2

Change of variables. Consider a canonical GLM (14) with cumulant $a(\eta)$. In such a model, $\ell''(y,\eta) = a''(\eta)$ 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 the parameter $\vartheta := \mathbf{\Sigma}^{-1/2}\theta$. Indeed, $X^\top\theta = Z^\top\vartheta$, and the corresponding calibrated vector $\widetilde{Z}(\vartheta)$,

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

satisfies $\widetilde{Z}(\vartheta) = \mathbf{\Sigma}^{-1/2}\widetilde{X}(\theta)$, so that its second-moment matrix $\Psi(\vartheta) := \mathbf{E}[\widetilde{Z}(\vartheta)\widetilde{Z}(\vartheta)^\top]$ is given by $\Psi(\vartheta) = \mathbf{\Sigma}^{-1/2}\mathbf{H}(\theta)\mathbf{\Sigma}^{-1/2}$. Verifying Assumptions 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 $\|\Psi(\vartheta)^{-1/2}\widetilde{Z}(\vartheta)\|_{\psi_2} \lesssim K_2$ at ϑ_* and closeby points. Similarly, Assumption D1 can be reformulated in terms of the new variables Z, ϑ, and Y.

Logistic regression. In what follows, we consider the particular example of logistic regression with Gaussian design $X \sim \mathcal{N}(0, \mathbf{\Sigma})$, verifying the assumptions introduced in Section 2.2.

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

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 the radius r of the Dikin ellipsoid holds with*
$$\bar{K}_2(r) \lesssim (1 + \log(1 + \|\theta_*\|_{\mathbf{\Sigma}} + r\sqrt{\rho}))\sqrt{1 + \|\theta_*\|_{\mathbf{\Sigma}} + r\sqrt{\rho}},$$
In particular, $\bar{K}_2(1/\sqrt{\rho})$ admits the same bound as K_2.
3. *If the model is well-specified, Assumption D1 holds with*
$$K_1 \lesssim \sqrt{\rho} \lesssim (1 + \|\theta_*\|_{\mathbf{\Sigma}})^{3/2}.$$
Moreover, a better bound holds for the subexponential *norm (see [Ver12, Section 5.2.4]):*
$$\|\mathbf{G}(\theta_*)^{-1/2}\ell'(Y, X^\top\theta_*)X\|_{\psi_1} \lesssim \log(1 + \|\theta_*\|_{\mathbf{\Sigma}})^2\sqrt{1 + \|\theta_*\|_{\mathbf{\Sigma}}};$$
equivalently, for any $u \in \mathcal{S}^{d-1}$ one has $\left(\mathbf{E}[\langle \mathbf{G}(\theta_*)^{-1/2}\ell'(Y, X^\top\theta_*)X, u\rangle^p)\right)^{1/p} \lesssim Kp$ 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

$$\Psi(\vartheta) = \begin{bmatrix} \kappa & 0_{d-1}^\top \\ 0_{d-1} & \kappa_\perp \mathbf{I}_{d-1}, \end{bmatrix}, \tag{117}$$

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

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

as

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

In fact, the form (117) for $\Psi(\vartheta)$ will be preserved with any elliptical distribution of X, with somewhat more complicated expressions for κ and $\kappa_\perp$ in the non-Gaussian case. Our next step is to lower-bound κ and $\kappa_\perp$, which automatically yields an upper bound for ρ in Assumption C:

$$\rho \le \frac{1}{\min(\kappa, \kappa_\perp)}. \tag{118}$$

1^o. Let us bound κ and $\kappa_\perp$ for logistic regression, i.e., when $a(\eta) = \log(1 + e^\eta)$. In this case,

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

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

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

which yields

$$\frac{e^{-|\eta|}}{4} \le a''(\eta) \le e^{-|\eta|}. \tag{119}$$

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. From [AS65, Eq. 7.1.13], we know the following bounds for $G(t)$:

$$\frac{2e^{-t^2/2}}{t + \sqrt{t^2 + 4}} \le G(t) \le \frac{2e^{-t^2/2}}{t + \sqrt{t^2 + 8/\pi}}, \quad t \ge 0. \tag{120}$$

These bounds are *sharp* in the constant terms under the square root; in particular, they imply

$$G(t) \approx \frac{e^{-t^2/2}}{t+1},$$

whence,

$$\kappa_\perp \approx \frac{1}{t+1}. \tag{121}$$

We can similarly bound κ:

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

Using the lower bound in (120), this gives

$$\kappa \geq \frac{4}{(t + \sqrt{t^2+4})(t^2 + 2 + \sqrt{t^4 + 4t^2})} \gtrsim \frac{1}{1+t^3}. \tag{122}$$

Plugging (121) and (122) into (118), we arrive at $\rho \lesssim 1 + \|\theta_*\|_{\mathbf{\Sigma}}^3$, as claimed The dependency on $\|\theta_*\|_{\mathbf{\Sigma}}$ cannot be improved since the lower bound in (120) is sharp.

$\mathbf{2^o}$. 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

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

where $Z_1 \sim \mathcal{N}(0,1)$, and we used (119) and (122). Thus, whenever $t \lesssim 1$, for such u we have

$$\|\langle u, \Psi(\vartheta_*)^{-1/2} \widetilde{Z}(\theta_*) \rangle\|_{\psi_2} \lesssim \|Z_1\|_{\psi_2} \lesssim 1. \tag{123}$$

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

$$|\langle u, \Psi(\vartheta_*)^{-1/2} \widetilde{Z}(\theta_*) \rangle| \lesssim |Z_1|. \tag{124}$$

On the other hand, in the case

$$|Z_1| \leq \frac{3 \log(1+t)}{t},$$

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

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

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

$$K_2 \lesssim (1 + \log(1 + \|\theta_*\|_{\mathbf{\Sigma}})) \sqrt{1 + \|\theta_*\|_{\mathbf{\Sigma}}}. \tag{125}$$

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 results in (125) whenever $r \lesssim 1/\sqrt{\rho}$, motivating our condition in Theorem 3.6.

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

$$\|\ell'(Y, X^\top \theta_*) Z\|_{\psi_1} \lesssim \log(1+t)^2 \sqrt{1+t}. \tag{126}$$

One of the equivalent definitions of the subexponential norm is as follows: a random variable $\xi \in \mathbb{R}$ satisfies $\|\xi\|_{\psi_1} \le K$ when its moments scale as $(\mathbf{E}[|\xi|^p])^{1/p} \lesssim Kp$, 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 random variables the scaling is $K\sqrt{p}$ (cf. Lemma A.1). To see (126), 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 can estimate the moments of the marginals of $\ell'(Y, X^\top \theta_*) Z = (Y - \sigma(Z^\top \vartheta_*) Z$ for $p \ge 1$:

$$\mathbf{E}_{Z,Y}[(Y - \sigma(Z^\top \vartheta_*))\langle Z, u\rangle]^p \le 2\mathbf{E}_Z\left[\sigma(Z^\top \vartheta_*)(1 - \sigma(Z^\top \vartheta_*))\langle Z, u\rangle^p\right] \lesssim 2\mathbf{E}_Z\left[e^{-|Z^\top \vartheta_*|}\langle Z, u\rangle^p\right],$$

where we used (119). For u parallel to ϑ_*, we should prove that

$$(1+t)^{3/2}\left(\int_0^{+\infty} e^{-tu} u^p e^{-u^2/2} \mathrm{d}u\right)^{1/p} \lesssim p \log^2(1+t)\sqrt{1+t}. \tag{127}$$

We proceed similarly to **2^o**, using that $(1+t)^{3p/2} e^{-tu} \le 1$ when $u \ge \frac{3p\log(1+t)}{2t}$. Thus, when $t \gtrsim 1$,

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

which implies (127). The remaining cases (u parallel to ϑ_* with $t \lesssim 1$, u orthogonal to ϑ_*) are straightforward, noting that a K-subgaussian random variable is also $O(K)$-subexponential. ■

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.

[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.

[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.

[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.

[MN89] Peter McCullagh and John A. Nelder. *Generalized Linear Models, Second Edition.* Chapman & Hall, 1989.

[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.

[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.

[RV18] Cynthia Rush and Ramji Venkataramanan. Finite sample analysis of approximate message passing algorithms. *IEEE Transactions on Information Theory*, 2018.

[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.

[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.

[vdGB09] Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[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.

[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.