Influence Diagnostics under Self-concordance

Jillian Fisher[1] **Lang Liu**[1] **Krishna Pillutla**[1] **Yejin Choi**[1,2] **Zaid Harchaoui**[1]
University of Washington[1] Allen Institute for Artificial Intelligence[2]

Abstract

Influence diagnostics such as influence functions and approximate maximum influence perturbations are popular in machine learning and in AI domain applications. Influence diagnostics are powerful statistical tools to identify influential datapoints or subsets of datapoints. We establish finite-sample statistical bounds, as well as computational complexity bounds, for influence functions and approximate maximum influence perturbations using efficient inverse-Hessian-vector product implementations. We illustrate our results with generalized linear models and large attention based models on synthetic and real data.

1 INTRODUCTION

Statistical machine learning models have been increasingly used in fully or partially automatized data analysis processes and artificial intelligence applications (Rudin, 2019). The automatizing of decisions impacting the society inspire a parallel effort to develop methods to identify the factors impacting specific decisions. The heightened scrutiny on the way statistical models now operate at a large scale and at a fast pace has led to a renewed interest in statistical diagnostics such as the influence function (Cook and Weisberg, 1982; Koh and Liang, 2017; Schioppa et al., 2022; Louvet et al., 2022).

The influence function or curve of a statistical estimator has been proposed to measure the sensitivity of the estimator to individual datapoints. Computing the influence of a particular datapoint boils down to computing an inverse-Hessian-vector product. Due to a greater focus on least-squares-type estimator with small samples, the computational aspects have received relatively little attention until recently (Koh and Liang, 2017; Schioppa et al., 2022), while the statistical aspects have mainly focused on large sample classical asymptotics (Rousseeuw et al., 2011; Avella-Medina, 2017).

The statistical analysis of influence functions for generalized linear models presents several challenges. For non-squared loss functions, the curvature captured by the Hessian varies away from the true parameter $\theta_\star$, a property that can be modelled using self-concordance. Moreover, non-asymptotic analyses for misspecified generalized linear models require recently developed tools such as matrix concentration inequalities (Mackey et al., 2014). We present non-asymptotic statistical bounds for influence functions of generalized linear models under pseudo self-concordance assumptions. Thanks to a novel interpretation of Broderick et al. (2020)'s maximum subset influence using superquantiles, we also obtain non-asymptotic guarantees for this diagnostic tool as well.

The computational analysis of influence is equally interesting. The statistical and computational trade-offs have not received attention to the best of our knowledge. We review classical algorithms such as the conjugate gradient method (Saad, 2003; Bai and Pan, 2021) and an approach using the Arnoldi iteration (Schioppa et al., 2022), and we develop approaches using variance reduced stochastic optimization algorithms (Bertsekas, 2015; Bach, 2021). Our analysis reveals interesting trade-offs depending on the near low-rank structure that is the eigendecay of the Hessian for small to moderate sample sizes relative to the dimension, as well as the potential benefits of using linearly convergent stochastic algorithms.

Outline. In Section 2, we introduce influence diagnostics and the computational challenges they present in high dimensional settings. In Section 3, we obtain finite-sample bounds on empirical influence functions for generalized linear models. We also achieve computational accuracy bounds on empirical influence functions computed using deterministic Krylov-based methods and stochastic optimization based methods. In Section 4, we provide similar guarantees for maximum subset influence owing to a novel superquantile interpretation. Lastly, in Section 5, we provide numerical illustrations of our theoretical bounds on synthetic data and real data, with generalized linear models and large attention based models.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

2 INFLUENCE FUNCTIONS

We are interested in the parameter $\theta_\star \in \Theta = \mathbb{R}^p$ defined as

$$\theta_\star := \arg\min_{\theta\in\Theta} \Big[F(\theta) := \mathbb{E}_{Z\sim P}\left[\ell(Z,\theta)\right] \Big], \tag{1}$$

where P is an unknown probability distribution over a data space $\mathcal{Z}$ and $\ell : \mathcal{Z} \times \Theta \to \mathbb{R}_+$ is a loss function that is closed, convex, and thrice continuously differentiable in the second argument. We assume this argmin is unique.

For instance, binary logistic regression corresponds to $\mathcal{Z} = \mathbb{R}^p \times \{\pm 1\}$ and a loss $\ell\big((x,y),\theta\big) = \log\big(1 + \exp(-y\langle\theta, x\rangle)\big)$. Here, problem (1) is equivalent to finding parameters $\theta_\star \in \Theta$ that minimize the Kullback-Leiblier divergence between the unknown data distribution P and the parametric model $P_\theta(Y|X=x) = 1/\big(1+\exp(-y\langle\theta,x\rangle)\big)$.

Since the data distribution P is unknown, we estimate $\theta_\star$ using an i.i.d. sample $Z_{1:n} := (Z_1, \cdots, Z_n) \sim P^n$. This leads to the M-estimation problem,

$$\theta_n := \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \ell(Z_i,\theta)\,, \tag{2}$$

where we assume the argmin to be unique. For the logistic regression example, θ_n is also the maximum likelihood estimator of $\theta_\star$.

Influence Functions. We quantify the influence of a fixed data point z on the estimator θ_n using the perturbation

$$\theta_{n,\varepsilon,z} := \arg\min_{\theta\in\Theta} \left\{ \frac{1-\varepsilon}{n}\sum_{i=1}^n \ell(Z_i,\theta) + \varepsilon\,\ell(z,\theta) \right\}$$

for some $\varepsilon > 0$. The difference $(\theta_{n,\varepsilon,z} - \theta_n)/\varepsilon$ is a measure of the local effect that the datapoint z has on the estimator θ_n, as illustrated in Figure 1. Influence functions provide a way to avoid recomputing this estimator for each $z \in \mathcal{Z}$ of interest by using a linear approximation of the map $\varepsilon \mapsto \theta_{n,\varepsilon,z}$ (Hampel, 1974). Concretely, we approximate

$$\frac{\theta_{n,\varepsilon,z} - \theta_n}{\varepsilon} \approx \frac{\mathrm{d}\theta_{n,\varepsilon,z}}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0} =: I_n(z)\,. \tag{3}$$

This quantity is well-defined when the Hessian $H_n(\theta) := (1/n)\sum_{i=1}^n \nabla^2\ell(Z_i,\theta)$ is invertible at $\theta = \theta_n$. We bound this approximation error in Theorem 2.

This idea of taking infinitesimal perturbations to approximate the effect of modifying data in statistics dates back to the Ph.D. dissertation of Hampel (1968) and subsequently, the infinitesimal jackknife (Jaeckel, 1972). A celebrated result of Cook and Weisberg (1982), obtained from invoking the implicit function theorem to differentiate through the first order optimality conditions of θ_n, gives the closed-form

$$I_n(z) = -H_n(\theta_n)^{-1}\nabla\ell(z,\theta_n)\,. \tag{4}$$

Since $I_n(z)$ does not depend on $\theta_{n,\varepsilon,z}$, there is no need to re-solve the M-estimation problem for each z. Instead, we solve a single linear system involving $H_n(\theta_n)$; we return to the computational aspects later.

In this work, we are interested in the non-asymptotic statistical behavior of the influence function $I_n(z)$. To define the population limit, we denote the perturbed population minimizer with an ε-fraction of the mass moved to z as,

$$\theta_{\star,\varepsilon,z} := \arg\min_{\theta\in\Theta} \left\{ \mathbb{E}_{Z\sim(1-\varepsilon)P+\varepsilon\delta_z}\left[\ell(Z,\theta)\right] \right\},$$

where δ_z denotes the point mass at z. The population influence function is defined similar to (3) as the derivative

$$I(z) := \frac{\mathrm{d}\theta_{\star,\varepsilon,z}}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0} = \lim_{\varepsilon\to 0}\frac{\theta_{\star,\varepsilon,z} - \theta_\star}{\varepsilon}\,. \tag{5}$$

If the Hessian $H_\star = \nabla^2 F(\theta_\star)$ of the population objective (1) is strictly positive definite at $\theta_\star$, we get a closed form expression similar to (4) due to Cook and Weisberg (1982):

$$I(z) = -H_\star^{-1}\nabla\ell(z,\theta_\star)\,. \tag{6}$$

As $n \to \infty$, uniform convergence arguments would give $\theta_n \to \theta_\star$ in probability under appropriate assumptions. From the continuous mapping theorem, we would expect that the sample influence function $I_n(z) = -H_n(\theta_n)^{-1}\nabla\ell(z,\theta_n)$ converges to the population influence $I(z) = -H_\star^{-1}\nabla\ell(z,\theta_\star)$. We establish finite-sample bounds in Section 3 to formalize this convergence.

Most Influential Subset. Similar to measuring the influence of a fixed point z, we also consider the influence of subsets of the sample $Z_{1:n}$. Given a scalar $\alpha \in (0,1)$, the most influential subset method of Broderick et al. (2020) aims to find the subset of the data of size at most αn that, when removed, leads to the largest increase of a continuously differentiable test function $h : \mathbb{R}^p \to \mathbb{R}$. A typical example of h is the loss $h(\theta) = \ell(z_{\text{test}},\theta)$ of a fixed test point z_{test}.

This approach relies on perturbing the weights of a weighted M-estimation problem around the nominal weights (Giordano et al., 2019). Given weights w in the probability simplex Δ^{n-1}, define $\theta_{n,w} := \arg\min_{\theta\in\Theta}\sum_{i=1}^n w_i\ell(Z_i,\theta)$, so that $\theta_n = \theta_{n,\mathbf{1}_n/n}$. Finding the maximum influence of any subset of data of size at most αn for a test function h amounts to solving $\max_{w\in W_\alpha} h\big(\theta_{n,w}\big)$ where

$$W_\alpha := \left\{ w \in \Delta^{n-1} \;:\; \begin{array}{l}\text{at most } \alpha n \text{ elements of } w \text{ are}\\ \text{zero and the rest are equal}\end{array} \right\}.$$

The most influential subset corresponds to the zero entries of the maximizing w. Unfortunately, this expression cannot be computed tractably as $|W_\alpha|$ grows exponentially in n. Instead, Broderick et al. (2020) use a linear approximation

$$h(\theta_{n,w}) \approx h(\theta_n) + \left\langle w - \frac{\mathbf{1}_n}{n}, \nabla_w h(\theta_{n,w})\Big|_{w=\mathbf{1}_n/n} \right\rangle.$$

Finding the most influential subset according to this linear approximation leads to the maximum subset influence

$$I_{\alpha,n}(h) := \max_{w \in W_\alpha} \left\langle w, \nabla_w h(\theta_{n,w}) \Big|_{w=\mathbf{1}_n/n} \right\rangle . \quad (7)$$

Similar to (4), the implicit function theorem together with the chain rule gives the closed form

$$I_{\alpha,n}(h) = \max_{w \in W_\alpha} \sum_{i=1}^n w_i v_i, \quad \text{where} \quad v_i = -\left\langle \nabla h(\theta_n), H_n(\theta_n)^{-1} \nabla \ell(Z_i, \theta_n) \right\rangle . \quad (8)$$

While the maximization over W_α in (8) is an instance of the NP-hard knapsack problem, its solution coincides with that of its continuous relaxation over $\operatorname{conv} W_\alpha$ when αn is an integer and the v_i's are unique. This continuous knapsack problem is solved by a greedy algorithm that zeros out the smallest αn entries of v_i's (Dantzig, 1957).

In this work, we also study the non-asymptotic statistical behavior of the subset influence $I_{\alpha,n}$. The population limit in this case is more subtle than for I_n of (4). Using similar arguments, we would expect the vector v to be related to the random variable $\phi(Z)$ where $\phi : \mathcal{Z} \to \mathbb{R}$ maps $z \mapsto -\langle \nabla h(\theta_\star), H_\star^{-1} \nabla \ell(z, \theta_\star) \rangle$, but the maximum over W_α is tricky. In Section 4, we rigorously define this population limit and establish convergence guarantees.

Computational Aspects. While linearization methods based on the infinitesimal jackknife avoid recomputing the M-estimator for each z, a naïve implementation of $I_n(z)$ (and similarly, $I_{\alpha,n}$) requires materializing and inverting the Hessian matrix $H_n(\theta_n) \in \mathbb{R}^{p \times p}$ in $O(np^2 + p^3)$ time with $O(p^2)$ storage. This approach does not scale to modern applications in deep learning with dense Hessians and large n, p. Instead, we rely on iterative algorithms to approximately minimize the convex quadratic

$$g_n(u) := \frac{1}{2} \langle u, H_n(\theta_n) u \rangle + \langle \nabla \ell(z, \theta_n), u \rangle . \quad (9)$$

Indeed, the unique minimizer $u_\star$ of g_n satisfies $0 = \nabla g_n(u_\star) = H_n(\theta_n) u_\star + \nabla \ell(z, \theta_n)$ so that $u_\star = I_n(z)$ in (4) as desired. Modern automatic differentiation software supports the efficient computation of the Hessian-vector product $u \mapsto \nabla^2 \ell(z, \theta) u$ without materializing the Hessian. We review some iterative algorithms that can achieve this.

The conjugate gradient method is a classical algorithm to solve linear systems defined by a positive definite matrix. It converges linearly, but each iteration requires a full batch Hessian-vector product $u \mapsto H_n(\theta_n) u$. We postpone precise rates to Section 3.

Alternatively, one might optimize the quadratic $g_n(u)$ with stochastic gradient descent (SGD). Here, each iteration requires a Hessian-vector product at only one sample Z_i, but the convergence rate is sublinear. We can get a linear rate at the same $O(1)$ per-iteration complexity through the use of variance reduction with the stochastic variance reduced gradient (SVRG; Johnson and Zhang, 2013) or its accelerated counterpart (Lin et al., 2018).

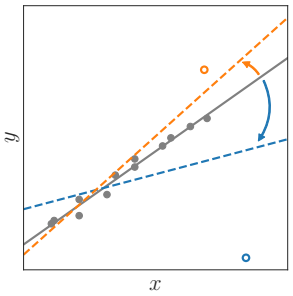

Figure 1: Illustration of the influence of a point z on the model parameters. The base model (gray) line is drastically affected when the blue point is included (blue dotted line) but less affected when the orange point is included (orange dotted line).

The LiSSA algorithm (Agarwal et al., 2017) solves this linear system by approximating the matrix inverse with its Neumann series $M^{-1} = \sum_{k=0}^\infty (\mathbf{I} - M)^k$ for positive definite M with $\|M\|_2 < 1$. By using an unbiased stochastic estimator $\nabla^2 \ell(Z_I, \theta_n)$ to $M = H_n(\theta_n)$, where I is a random index, this reduces exactly to the SGD baseline. See Appendix B for details.

Schioppa et al. (2022) propose to solve the linear system with a low-rank approximation of the Hessian. Concretely, let $H_n(\theta_n) = Q \Lambda Q^\top$ denote its eigenvalue decomposition with $\Lambda = \operatorname{Diag}(\lambda_1, \ldots, \lambda_d)$ arranged in non-increasing order. The rank-k approximation of $v = H_n(\theta_n)^{-1} u$ is given by $v_k = Q \operatorname{Diag}(\lambda_1^{-1}, \ldots, \lambda_k^{-1}, 0, \ldots, 0) Q^\top u$. The k-largest eigenvalues and their eigenvectors are approximated using the Lanczos/Arnoldi iterations (Lanczos, 1950; Arnoldi, 1951). This algorithm requires computations of a full batch Hessian-vector product.

For a full error characterization of the influence estimate $\hat{I}_n(z)$ returned by an iterative algorithm, we must take into account both the statistical error $I_n(z) - I(z)$ and the computational error $\hat{I}_n(z) - I_n(z)$. This will be our goal for the next section.

3 ERROR ANALYSIS OF INFLUENCE ESTIMATION

We start by establishing a bound on the statistical error of the influence $I_n(z) = -H_n(\theta_n)^{-1} \nabla \ell(z, \theta_n)$ of a data point z to the population limit $I(z) = -H(\theta_\star)^{-1} \nabla \ell(z, \theta_*)$.

We give an error bound $\|I_n(z) - I(z)\|_{H_\star}$ in the natural geometry implied by the population Hessian $H_\star := H(\theta_\star)$ at the true parameter $\theta_\star$; here we use the notation $\|u\|_A^2 = \langle u, Au \rangle$ for a positive definite matrix A. The $H_\star$-norm captures the behavior of $I(z)$ and $I_n(z)$ in an *affine-invariant manner*. That is, if we parameterize the problem in terms of $\theta' = A\theta$ for an invertible matrix A so that the loss is $\ell'(z, \theta') = \ell(z, A^{-1}\theta')$, the influ-

ence functions I' in this new parameterization satisfies $I'(z) = A\, I(z)$ and similarly for its sample version. Letting $H'_\star := \mathbb{E}_{Z\sim P}[\nabla^2 \ell'(z, \theta'_\star)]$ be the (reparameterized) Hessian at the minimizer $\theta'_\star = A\theta_\star$, we can verify that $\|I'_n(z) - I'(z)\|_{H'_\star} = \|I_n(z) - I(z)\|_{H_\star}$, i.e., the error criterion is affine-invariant.

3.1 Statistical Error Bound

Our statistical error bound depends on a notion of effective dimension of the statistical model. Define the covariance matrix of the gradient as

$$G(\theta) = \mathrm{Cov}_{Z\sim P}\big(\nabla \ell(Z, \theta)\big)\,, \tag{10}$$

where $\mathrm{Cov}(\xi) = \mathbb{E}[\xi\xi^\top] - \mathbb{E}[\xi]\,\mathbb{E}[\xi]^\top$ is the covariance matrix of a random vector ξ. We define the **effective dimension** of this problem as

$$p_\star = \mathbf{Tr}\left[H_\star^{-1/2} G_\star H_\star^{-1/2}\right]\,, \tag{11}$$

where $G_\star := G(\theta_\star)$ is the gradient covariance at $\theta_\star$.

The covariance $G_\star$ has a special meaning for maximum likelihood estimation. Concretely, if the loss $\ell(z, \theta) = -\log P_\theta(z)$ is the negative log likelihood and the statistical model $P_{\theta_\star}$ is well-specified, then $G_\star$ is the information matrix at $\theta_\star$. In this case, we have $G_\star = H_\star$ so that the effective dimension $p_\star$ equals the ambient dimension p.

For misspecified models or for general M-estimation problems beyond maximum likelihood, $G_\star$ and $H_\star$ are distinct in general. The effective dimension $p_\star$ captures the mismatch between the two; it can be much smaller or much larger than p. We can have $p_\star \ll p$ when the eigenvalues of $G_\star$ decay faster than those of $H_\star$. Conversely, we get that $p_\star > p$ when the eigenvalues of $G_\star$ decay slower than those of $H_\star$. We refer to Appendix C for precise calculations. Note that regardless of whether $p_\star > p$ or $p_\star < p$, a dependence on $p_\star$ is unavoidable since $p_\star/n$ is a lower bound on the estimation error (Fortunati et al., 2016).

Assumptions. We make the following assumptions.

(a) For any $z \in \mathcal{Z}$, the loss function $\ell(z, \cdot)$ is pseudo self-concordant for some $R \geq 1$:

$$|D^3_\theta \ell(z, \theta)[u, u, v]| \leq R\|u\|^2_{\nabla^2 \ell(z,\theta)}\, \|v\|_2\,,$$

where $D^3_x f(x)[u, u, v] := \frac{\mathrm{d}}{\mathrm{d}t}\langle u, \nabla^2 f(x + tv)\, u\rangle|_{t=0}$ for f thrice continuously differentiable and where $\|\cdot\|_2$ denotes the spectral norm for matrices.

(b) There exists a constant $K_1 \geq 1$ such that the normalized gradient $H_\star^{-1/2}\nabla \ell(Z, \theta_\star)$ at $\theta_\star$ is sub-Gaussian with parameter K_1.

(c) There exists $K_2 \geq 1$ such that the standardized Hessian $H_\star^{-1/2}\, \nabla^2 \ell(Z, \theta_\star)\, H_\star^{-1/2} - \mathbf{I}_p$ at $\theta_\star$ satisfies a Bernstein condition with parameter K_2 (Definition 30 in Appendix I). Moreover,

$$\sigma_H^2 := \Big\|\mathbb{V}\big(H(\theta_\star)^{-1/2}\, \nabla^2 \ell(Z, \theta_\star)\, H(\theta_\star)^{-1/2}\big)\Big\|_2$$

is finite, where we denote $\mathbb{V}(H) = \mathbb{E}[HH^\top] - \mathbb{E}[H]\,\mathbb{E}[H]^\top$ for a random matrix H.

Self-concordance was introduced by Nesterov and Nemirovskii (1994) to give an affine-invariant analysis of Newton's method and was adapted by Bach (2010) to apply to logistic regression; we use the latter assumption. This assumption prevents $\nabla^2 \ell(z, \theta)$ from changing too quickly with θ. The most useful consequence of this assumption is a spectral approximation of the Hessian $(1/2)H(\theta') \preceq H(\theta) \preceq 2\, H(\theta')$ for θ and θ' close enough in terms of the Euclidean distance.

We make the last two assumptions to argue about the concentration of $\nabla \ell(Z, \theta_\star)$ and $\nabla^2 \ell(Z, \theta_\star)$ respectively to their expected values for $Z \sim P$. We make appropriate normalizations so that the assumptions are affine invariant, similar to the error criterion. Since $\mathbb{E}[\nabla \ell(Z, \theta_\star)] = 0$, Assumption (b) gives a high-probability bound on $\|\nabla \ell(Z, \theta_\star)\|_{H_\star^{-1}}$ in the natural $H_\star^{-1}$ norm of the gradient. Assumption (c) gives the spectral concentration $(1/2)H(\theta) \preceq H_n(\theta) \preceq 2\, H(\theta)$ for a fixed θ with high probability for n large enough.

Example. The assumptions outlined above hold for all generalized linear models under some regularity conditions. We give one concrete examples here (more can be found in Appendix I.4).

Logistic Regression: Let $\mathcal{Z} \subset B_{p,M} \times \{\pm 1\}$ for some $M > 0$. Consider the loss $\ell(z, \theta) = \log\big(1 + \exp(-y\langle\theta, x\rangle)\big)$ and let $\sigma(z) = \frac{1}{1+e^{-z}}$. Assume that $H(\theta_\star) \succ 0$.

(a) Pseudo self-concordance. Note that $\nabla^2_\theta \ell(z, \theta) = \sigma(\theta^\top x)[1 - \sigma(\theta^\top x)]xx^\top$ and $D^3_\theta \ell(z, \theta)[u, u, v] = \sigma(\theta^\top x)[1 - \sigma(\theta^\top x)][1 - 2\sigma(\theta^\top x)](u^\top x)^2(v^\top x)$. It follows that $|D^3_\theta \ell(z, \theta)[u, u, v]| \leq M\|v\|_2\|u\|^2_{\nabla^2 \ell(z,\theta)}$ and thus ℓ is pseudo self-concordant with $R \geq M$.

(b) Sub-Gaussian gradient. Note that $\|\nabla_\theta \ell(Z, \theta_\star)\|_2 = \|[1 - \sigma(Y\theta_\star^\top X)]YX\|_2 \leq M$. Therefore, the normalized gradient $H(\theta_\star)^{-1/2}\nabla \ell(Z, \theta_\star)$ is sub-Gaussian (cf. Lemma 36 from Appendix I).

(c) Bernstein Hessian. Note that $\|\nabla^2_\theta \ell(Z, \theta_\star)\|_2 \leq \|XX^\top\|_2/4 \leq M^2/4$. It follows that the standardized Hessian $H(\theta_\star)^{-1/2}\nabla^2_\theta \ell(Z, \theta_\star)H(\theta_\star)^{-1/2} - I_p$ satisfies the matrix Bernstein condition (cf. Lemma 39 from Appendix I).

Statistical Error Bound. Below and throughout, we omit absolute constants.

Theorem 1. *Suppose the assumptions above hold and*

$$n \geq C_{K_1, K_2, \sigma_H}\left(\frac{R^2 p_\star}{\mu_\star}\log\frac{1}{\delta} + \log\frac{p}{\delta}\right)\,,$$

where $\mu_\star = \lambda_{\min}(H_\star)$ *and* C_{K_1,K_2,σ_H} *is a constant depending on* K_1, K_2, *and* σ_H. *Then, with probability at least* $1-\delta$, *we have* $\frac{1}{4}H_\star \preceq H_n(\theta_n) \preceq 3H_\star$ *and*

$$\|I_n(z) - I(z)\|_{H_\star}^2 \leq C_{K_1,K_2,\sigma_H} \frac{R^2 p_\star^2}{\mu_\star n} \log^3\left(\frac{p}{\delta}\right).$$

Remark. In this result, we view z as a random element following the data distribution P. The quantities $\|\nabla\ell(z,\theta_\star)\|_{H_\star^{-1}}$ and $\|H_\star^{-1/2} H(z,\theta_\star) H_\star^{-1/2}\|_2$ are controlled using the sub-Gaussian gradient and matrix Bernstein assumptions. A similar result holds if we treat z as a fixed datapoint, since these quantities are now fixed as well.

Theorem 1 has several merits. First, it is adapted to the eigenspectrum of $G_\star$ and $H_\star$ via the effective dimension $p_\star$; the bound only has a logarithmic dependence on the ambient dimension p. The effective dimension $p_\star$ is also affine-invariant, similar to the error criterion. The only geometry-dependent (i.e., not affine-invariant) term in Theorem 1 is the minimal eigenvalue $\mu_\star$ of the Hessian $H_\star$. Third, we get a fast $1/n$ rate, faster than the $1/\sqrt{n}$ rate typical of uniform convergence arguments.

We now sketch the key aspects of its proof. The full proof is given in Appendix D.

Proof Sketch of Theorem 1. We use the triangle inequality to bound $\|I_n(z) - I(z)\|_{H_\star}$ by

$$\begin{aligned}
&\left\|\left(H_n(\theta_n)^{-1} - H_\star^{-1}\right)\left(\nabla\ell(z,\theta_n) - \nabla\ell(z,\theta_\star)\right)\right\|_{H_\star} \\
&+ \left\|\left(H_n(\theta_n)^{-1} - H_\star^{-1}\right)\nabla\ell(z,\theta_\star)\right\|_{H_\star} \\
&+ \left\|H_\star^{-1}\left(\nabla\ell(z,\theta_n) - \nabla\ell(z,\theta_\star)\right)\right\|_{H_\star}.
\end{aligned}$$

The proof follows from arguing that $\theta_n \to \theta_\star$, $\nabla\ell(z,\theta_n) \to \nabla\ell(z,\theta_\star)$, and $H_n(\theta_n) \to H_\star$ in the appropriate sense. The first comes from a localization result of Ostrovskii and Bach (2021) that states that θ_n lies in a Dikin ellipsoid of radius $\sqrt{p_\star/n}$ around $\theta_\star$ for n large enough, i.e., $\|\theta_n - \theta_\star\|_{H_\star}^2 \lesssim p_\star/n$. The second comes from arguing using pseudo self-concordance that the gradient $\nabla\ell(z,\cdot)$ is Lipschitz w.r.t. $\|\cdot\|_{H_\star}$ in the Dikin Ellipsoid around $\theta_\star$. For the last one, we argue that $H_n(\theta_n) \approx H_n(\theta_\star)$ from pseudo self-concordance, and formalize $H_n(\theta_\star) \to H_\star$ by matrix concentration. □

In addition to the statistical error bound in Theorem 1, we also provide a bound for the approximation error in (3). Here, we treat z as a fixed data point and make the following boundedness assumptions in addition to the assumptions above.

(d) The normalized gradient is bounded in a neighborhood of $\theta_\star$, i.e., there exist $M_1 \geq 1, \rho \in (0, R^{-1}]$ such that $\|\nabla\ell(z,\theta)\|_{H_\star^{-1}} \leq M_1$ for all $\|\theta - \theta_\star\|_{H_\star} \leq \rho$.

(e) The normalized Hessian is bounded in a neighborhood of $\theta_\star$, i.e., there exist $M_2 \geq 1, \rho \in (0, R^{-1}]$ such that $\|H(z,\theta)\|_{H_\star^{-1}} \leq M_2$ for all $\|\theta - \theta_\star\|_{H_\star} \leq \rho$.

Theorem 2. *Suppose that the assumptions above hold,* $\varepsilon \leq C\min\{\rho/M_1, 1/M_2, \sqrt{\mu_\star}/RM_1\}$, *and*

$$n \geq C_{K_1,K_2,\sigma_H}\left[\left(\frac{R^2}{\mu_\star} + \frac{1}{\rho}\right) p_\star \log\frac{1}{\delta} + \log\frac{p}{\delta}\right].$$

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

$$\left\|\frac{\theta_{n,\varepsilon,z} - \theta_n}{\varepsilon} - I_n(z)\right\|_{H_n(\theta_n)}^2 \leq C_{M_1,M_2} \times \left\{\left[\exp\left(\frac{C_{K_1,M_1} R}{\sqrt{\mu_\star}}\left(\sqrt{\frac{p_\star}{n}\log\frac{1}{\delta}} + \varepsilon\right)\right) - 1\right]^2 + \varepsilon^2\right\}.$$

A full proof can be found in Appendix E.

3.2 Computational and Total Error Bounds

We consider iterative first-order algorithms to compute the influence function $I_n(z) = \arg\min_u g_n(u)$ by minimizing the convex quadratic $g_n(u)$ defined in (9).

We aim to find an ε-approximate minimizer u that satisfies $\mathbb{E}[\|u - I_n(z)\|_{H_n(\theta_n)}^2 | Z_{1:n}] \leq \varepsilon$. This error criterion is not only affine-invariant, but is also equivalent to $\mathbb{E}[g_n(u) - \min g_n | Z_{1:n}] \leq 2\varepsilon$. Throughout this section, we assume for all $z \in \mathcal{Z}$ that $\ell(z,\cdot)$ is L-smooth, i.e., $\|\nabla^2\ell(z,\theta)\|_2 \leq L$ for all θ. The complexity of minimizing g_n with first order algorithms depends on the condition number $\kappa_n := L/\lambda_{\min}\left(H_n(\theta_n)\right)$. The corresponding condition number of the population Hessian $H_\star$ is $\kappa_\star := L/\lambda_{\min}(H_\star) = L/\mu_\star$.

Any ε-approximate minimizer $\hat{I}_n(z)$ of g_n satisfies the following total error bound.

Proposition 3. *Consider the setting of Theorem 1, and let* $\mathcal{G}$ *denote the event under which its conclusions hold. Let* $\hat{I}_n(\theta)$ *be an estimate of* $I_n(\theta)$ *that satisfies* $\mathbb{E}\left[\|\hat{I}_n(z) - I_n(z)\|_{H_n(\theta_n)}^2 | Z_{1:n}\right] \leq \varepsilon$. *Then,*

$$\mathbb{E}\left[\|\hat{I}_n(z) - I(z)\|_{H_\star}^2 \,\middle|\, \mathcal{G}\right] \leq 8\varepsilon + C\frac{R^2 p_\star^2}{\mu_\star n}\log^3\frac{p}{\delta},$$

where $C = C_{K_1,K_2,\sigma_H}$ *is as in Theorem 1.*

This bound is obtained by translating the approximation error in the $H_n(\theta_n)$-norm to the $H_\star$-norm using the spectral Hessian approximation under $\mathcal{G}$ and the triangle inequality.

The conjugate gradient method is known to require $T_n(\varepsilon) := \sqrt{\kappa_n}\log\left(\|I_n(z)\|_{H_n(\theta_n)}^2/\varepsilon\right)$ iterations (ignoring constants) to return an ε-approximate minimizer (e.g. Saad, 2003; Chen, 2005; Bai and Pan, 2021). Since each iteration requires n Hessian-vector products, the total computational complexity to obtain an ε-approximate

Table 1: The number of calls to a Hessian-vector product oracle $u \mapsto \nabla^2 \ell(z,\theta)u$ so that (a) the computational error is at most ε, and (b) the total error is at most ε in the sense of Proposition 3. We show the dependence of the former on the condition number $\kappa_n = L/\lambda_{\min}(H_n(\theta_n))$, the optimal magnitude $\Delta_n = \|I_n(z)\|^2_{H_n(\theta_n)}$, and the SGD noise σ_n^2, defined in Appendix F.3. The total error bound depends on the corresponding population quantities $\kappa_\star = L/\lambda_{\min}(H_\star)$, $\Delta_\star = \|I(z)\|^2_{H_\star}$, and $\sigma_\star^2$, as well the effective dimension $p_\star$. We omit the dependence on problem constants $R, L, K_1, K_2, \sigma_H^2$, as well as logarithmic terms in $p, p_\star, \delta$. For the low-rank approximation, we assume that the total complexity to obtain a rank-k approximation is $O(k)$ full batch Hessian-vector products. We present computational error bounds assuming the eigenvalues $\lambda_i(H_n(\theta_n))$ of $H_n(\theta_n)$ decay polynomially as $i^{-\beta}$ ($\beta > 1$) or exponentially as $e^{-\nu i}$ ($\nu > 0$). The same decay is assumed for $H_\star$ for the total error bound. The full proofs of these bounds are given in Appendix F.

Method	**Computational Error**	**Total Error**	**Reference**
Conjugate Gradient	$n\sqrt{\kappa_n}\log\frac{\Delta_n}{\varepsilon}$	$\frac{\kappa_\star^{3/2}p_\star^2}{\varepsilon}\log\frac{\Delta_\star}{\varepsilon}$	Corollary 16
SGD	$\frac{\sigma_n^2}{\varepsilon} + \kappa_n \log\frac{\kappa_n\Delta_n}{\varepsilon}$	$\frac{\sigma_\star^2}{\varepsilon} + \kappa_\star \log\frac{\kappa_\star\Delta_\star}{\varepsilon}$	Corollary 20
SVRG	$(n+\kappa_n)\log\frac{\kappa_n\Delta_n}{\varepsilon}$	$\kappa_\star\left(1+\frac{p_\star^2}{\varepsilon}\right)\log\frac{\kappa_\star\Delta_\star}{\varepsilon}$	Corollary 23
Accelerated SVRG	$(n+\sqrt{n\kappa_n})\log\frac{\kappa_n\Delta_n}{\varepsilon}$	$\kappa_\star\left(\sqrt{\frac{p_\star^2}{\varepsilon}}+\frac{p_\star^2}{\varepsilon}\right)\log\frac{\kappa_\star\Delta_\star}{\varepsilon}$	Corollary 23
Low-Rank Approx. ($\lambda_i \propto i^{-\beta}$)	$n\left(\frac{\kappa_n\Delta_n}{\varepsilon}\right)^{\frac{1}{\beta-1}}$	$\left(\frac{\kappa_\star}{\varepsilon}\right)^{\frac{\beta}{\beta-1}} p_\star^2 \Delta_\star^{\frac{1}{\beta-1}}$	Corollary 25
Low-Rank Approx. ($\lambda_i \propto e^{-\nu i}$)	$\frac{n}{\nu}\log\frac{\kappa_n\Delta_n}{\varepsilon}$	$\frac{\kappa_\star p_\star^2}{\nu\varepsilon}\log\frac{\kappa_\star\Delta_\star}{\varepsilon}$	Corollary 25

minimizer is $O\big(n\,T_n(\varepsilon)\big)$. To make the statistical error $\|I_n(z) - I(z)\|^2_{H_\star}$ to be smaller than ε, we must choose $n \geq n(\varepsilon) = \tilde{O}\big(R^2 p_\star^2/(\mu_\star\varepsilon)\big)$ (ignoring constants and logarithmic factors). Proposition 3 now says that the overall computational complexity to reduce the total error under $O(\varepsilon)$ is $O\big(n(\varepsilon)T(\varepsilon)\big)$.

Table 1 presents this bound with sample-dependent quantities such as κ_n and $\|I_n(z)\|_{H_n(\theta_n)}$ translated to their population versions. Table 1 also lists the corresponding bounds for the other algorithms we consider. We discuss the implications of the total error bounds. We use $\tilde{O}(\cdot)$ to suppress logarithmic terms in $1/\varepsilon$ below.

Marginal Benefits of Variance Reduction. For a fixed n, the computational error bounds agree with the conventional wisdom that SVRG is significantly faster than SGD, especially for small ε. Indeed, the error $\tilde{O}(n+\kappa_n)$ of SVRG only depends logarithmically on $1/\varepsilon$, while the SGD error $\tilde{O}(\sigma_n^2/\varepsilon + \kappa_n)$ is polynomial. However, the statistical error bounds suggest that the sample size must be $n = \tilde{O}(R^2 p_\star^2/\mu_\star\varepsilon)$, so the total error of SVRG scales as $1/\varepsilon$. This matches SGD up to constants. SVRG has better constants only if the SGD noise $\sigma_\star^2 > p_\star^2/\mu_\star$ is large.

Marginal Benefits of Acceleration. For fixed n, accelerated SVRG's rate of $\tilde{O}(n+\sqrt{n\kappa_n})$ is faster than SVRG for ill-conditioned problems where $\kappa_n > n$, but is no worse for well-conditioned problems where $\kappa_n \leq n$. To have a small total error, we need $n = \tilde{O}(1/\varepsilon)$, while the condition numbers satisfy $\kappa_n \leq 4\kappa_\star$ for κ_n a constant (under Theorem 1). Thus, for ε small, the problem is well-conditioned and acceleration offers marginal benefits.

Stochastic Methods Outperform Full Batch Methods. The total error of the conjugate gradient method is $\tilde{O}(\kappa_\star^{3/2}p_\star^2/\varepsilon)$ while SVRG is $\tilde{O}(\kappa_\star p_\star^2/\varepsilon)$. Thus, SVRG always has better constants than the conjugate gradient method. This is also true of accelerated SVRG.

Low-rank Approximations Work for Faster Eigendecay. For a slow polynomial decay $\lambda_i(H_\star) \propto i^{-\beta}$ of the eigenvalues of $H_\star$ for $\beta > 1$, the total error scales as $\varepsilon^{-\beta/(\beta-1)}$, which is worse than the $1/\varepsilon$ rate for all other methods considered. However, for a faster exponential decay $\lambda_i(H_\star) \propto e^{-\nu i}$ for $\nu > 0$, its $1/\varepsilon$ rate matches SVRG exactly up to a factor of ν, despite being a full batch method.

4 MOST INFLUENTIAL DATA SUBSETS

We now turn to the subset influence defined in Section 2. We start by formalizing the population limit and then establish statistical error bounds. Let $h : \Theta \to \mathbb{R}$ be a continuously differentiable test function and $\alpha \in (0,1)$ be fixed throughout. We only consider n where αn is an integer.

Population Limit. In order to derive the population limit of the subset influence $I_{\alpha,n}(h)$ from (8), we interpret the weights $w \in W_\alpha \subset \Delta^{n-1}$ as a probability distribution over the n datapoints. This gives

$$I_{\alpha,n}(h) = \max_{w \in W_\alpha} \mathbb{E}_{i\sim w}[\phi_n(Z_i)],$$

where $\phi_n(z) = -\langle \nabla h(\theta_n), H_n(\theta_n)^{-1}\nabla\ell(z,\theta_n)\rangle$.

This suggests that the population limit should be $\sup_{Q \in \mathcal{Q}} \mathbb{E}_{Z \sim Q}[\phi(Z)]$ over an appropriate set of distributions $\mathcal{Q}$, where $\phi(z) = -\langle \nabla h(\theta_\star), H_\star^{-1} \nabla \ell(z, \theta_\star) \rangle$.

Since the maximum of a linear program occurs at a corner, we can pass from the max over W_α to its convex hull

$$\operatorname{conv} W_\alpha = \{ w \in \Delta_{n-1} \,:\, w_i (1-\alpha) n \leq 1 \;\; \forall\, i \} \,.$$

Compared to the uniform distribution $\mathbf{1}_n / n$ over $Z_{1:n}$, $w \in \operatorname{conv} W_\alpha$ allows for weights that are a factor of $(1-\alpha)^{-1}$ larger. If P is a continuous distribution with density f_P, then a natural choice for $\mathcal{Q}$ is the set of distributions with density $f_Q(z) \leq f_P(z)/(1-\alpha)$.

We can formalize this discussion through the notion of a tail statistic known as the *superquantile* or the *conditional value at risk* (Rockafellar and Uryasev, 2000). The superquantile of a random variable $Z \sim P$ at level α is defined as

$$S_\alpha(Z) := \sup \left\{ \mathbb{E}_{Z \sim Q}[Z] \,:\, \frac{\mathrm{d}Q}{\mathrm{d}P} \leq \frac{1}{1-\alpha} \right\} ,$$

where $\mathrm{d}Q/\mathrm{d}P$ denotes the Radon-Nikodym derivative of Q w.r.t. P. This constraint subsumes both the density ratio constraint in the continuous case and the weight ratio constraint in the discrete case. The superquantile has a long and storied history in economics and quantitative finance, with recent applications in machine learning; we refer to (Laguel et al., 2021) for a survey. We overload notation to denote the superquantile of the empirical measure over $v_1, \ldots, v_n$ as $S_\alpha(v_1, \ldots, v_n)$.

We formalize the connection between the maximum subset influence $I_{n,\alpha}$ and the superquantile.

Proposition 4. *If αn is an integer, then $I_{n,\alpha}(h) = S_\alpha(v_1, \ldots, v_n)$ where $v_i = -\big\langle \nabla h(\theta_n), H_n(\theta_n)^{-1} \nabla \ell(Z_i, \theta_n) \big\rangle$.*

Proposition 4 motivates us to define the **population subset influence** as

$$I_\alpha(h) = S_\alpha \Big[-\nabla h(\theta_\star)^\top H(\theta_*)^{-1} \nabla \ell(Z, \theta_\star) \Big] \,. \qquad (12)$$

Assumptions. We need to use the strengthen assumptions made in Theorem 2 for technical reasons. We also add the following

(f) The test function h is bounded as $\|\nabla h(\theta)\|_{H_\star^{-1}} \leq M_1'$ and $\|H_\star^{-1/2} \nabla^2 h(\theta) H_\star^{-1/2}\|_2 \leq M_2'$ for all $\|\theta - \theta_\star\|_{H_\star} \leq \rho$.

Assumption (f) asserts the boundedness of the test function h. We make this assumption in a neighborhood around $\theta_\star$.

Statistical Bound. We now state our main bound.

Theorem 5. *Suppose the assumptions above hold and the sample size n satisfies the condition in Theorem 1. Then, with probability at least $1-\delta$, we have*

$$\big(I_{\alpha,n}(h) - I_\alpha(h)\big)^2 \leq \frac{C_{M_1, M_2, M_1', M_2'}}{(1-\alpha)^2} \frac{R^2 p_\star}{\mu_\star n} \log \frac{n \vee p}{\delta} \,.$$

Theorem 5 has the same merits as Theorem 1: it uses the effective dimension $p_\star$ and exhibits only a logarithmic dependence on the ambient dimension p. We square the left side so that it scales for $\alpha \to 0$ as the squared norm $\|(1/n) \sum_{i=1}^n I_n(Z_i) - \mathbb{E}_{Z \sim P}[I(Z)]\|_{H_\star}^2$, comparable to Theorem 1. We get a fast $\log n / n$ rate rather than a slow $1/\sqrt{n}$ rate.

The proof relies on the equivalent expression

$$S_\alpha(Z) = \inf_{\eta \in \mathbb{R}} \left\{ \Phi(Z, \eta) := \eta + \frac{1}{1-\alpha} \mathbb{E}(Z - \eta)_+ \right\}$$

of the superquantile where $(\cdot)_+ = \max\{\cdot, 0\}$. We analyze the convergence of $\Phi\big(\phi_n(Z_{1:n}), \eta\big)$ to $\Phi(\phi(Z), \eta)$ for fixed η using the same techniques as Theorem 1. Then, we construct an ε-net so the bound holds for all η, including the minimizer. The full proof is given in Appendix G.

Related Work. Influence functions or curves have originally been proposed by Hampel (1974), and partly motivated by Jaeckel (1972)'s "infinitesimal jackknife". Cook and Weisberg (1982) showed that the influence function can be computed using inverse Hessian gradient products. Recent works on influence functions include (Cook, 1986; Hadi et al., 1995; Zhu and Zhang, 2004; Ma et al., 2014; Zhao et al., 2019). The theoretical statistical analysis has mostly focused on large-sample asymptotics hence in small dimensions, and we refer to the recent work (Avella-Medina, 2017) for a comprehensive survey.

Efficiently computing influence functions, or related inverse-Hessian-vector products, has received attention recently in the context of the training of deep neural networks using natural gradient or Newton-like algorithms (Henriques et al., 2019). Specifically, on influence functions, stochastic convex optimization algorithms (Agarwal et al., 2017), conjugate gradient methods (Saad, 2003), and low-rank variants (Schioppa et al., 2022) have been applied. The recent discovery of linear convergence for variance-reduced optimization algorithms makes them potentially competitive for the efficient computation of influence functions.

5 EXPERIMENTS

We explore the convergence of the empirical influence function to its population counterpart for classical linear models. We also report the findings from large attention based models, for which little statistical theory is known, yet maximum influential subsets can still be computed as for any black-box model. Appendix H contains the full details of this section. The code as well as the scripts to reproduce the experiments are made publicly available online https://github.com/jfisher52/influence_theory.

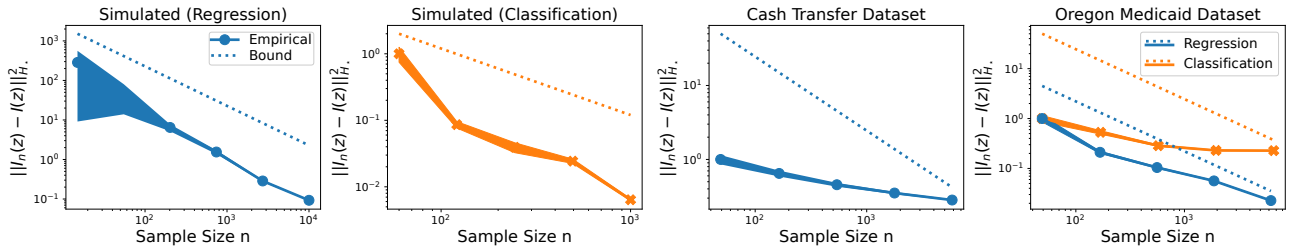

Figure 2: Convergence of the empirical influence function to the population (solid line) compared to the bound of Theorem 1 (dotted line) with linear regression and classification models for simulated (left two) and real data (right two). We plot the mean over 100 repetitions, and the shaded area denotes the 95% standard error.

5.1 Linear Models

We consider synthetic ridge regression and binary logistic regression in $\mathbb{R}^9$. The input $x \sim \mathcal{N}(0, \mathbf{I})$ is normal, and the outputs are generated with a linear or logistic model from i.i.d. noise based on a fixed $\theta_\star$. We also consider two real datasets: (1) Oregon Medicaid (Finkelstein et al., 2012), where the goal is to estimate the overall health (classification) and the number of good health days in the last month (regression) of an individual, and (2) Cash Transfer (Angelucci and De Giorgi, 2009), where the goal is to estimate the total consumption of an individual (regression). Both datasets use some economic and demographic features and treatments as inputs to the model; they contain 20K and 50K points respectively.

We plot the statistical convergence of the exact empirical influence $I_n(z)$ to the population influence $I(z)$ for fixed z using various sample sizes n as well as the bound of Theorem 1. For the real data, we use the full dataset as the population. We measure the influence of points z that are outliers added to the training set for the simulations and a random sample for the real data.

Results: Tightness of Theorem 1. The results are given in Figure 2. We see for the simulated datasets (left two plots) that the empirical observations for a straight line in log-log scale whose slope matches that of the bound. This indicates that the $1/n$ rate of our bounds is also observed empirically.[1] This is also approximately true for the regression line in the Oregon Medicaid dataset. We note that its classification line and the Cash Transfer dataset have slopes that differ from the bound. This phenomenon could be due to the error in the population influence used for the plots: we approximate it from a larger data sample because we do not have access to the population distribution. Note that we do not see such a behavior in the simulated classification task, where we can more accurately approximate the population. In all of these cases, Theorem 1 is still an upper bound on the empirical error.

[1] A log-log plot of $y = cx^a$ is a straight line with slope a.

5.2 Large Transformer Language Models

Setup. We consider (a) a question-answering task where the goal is to respond to a natural language question with a factually correct answer, and (b) a text continuation task where the goal is to generate ten tokens following a given context. We use a BART-base model (Lewis et al., 2020) on the zsRE dataset (Levy et al., 2017) and a DistilGPT-2 model (Sanh et al., 2019) on the WikiText-103 dataset (Merity et al., 2017) respectively. We subsample the training set size for various n and finetune a pretrained model to get θ_n. We take the largest value of n as the population version: this value was 5K and 2K respectively. We estimate the population influence with 100 epochs of SVRG, while we use 50 passes through the data for the approximate methods. We compute the influence $I_n(z)$ for 5 points $z_1, \cdots, z_5$. The quadratic g_n from (9) is nonconvex and unbounded below if the Hessian $H_n(\theta_n)$ is not positive semidefinite; we find this to be the case for our experiments with the deep nets. To overcome this, we consider

$$I_{n,\lambda}(z) = -\big(H_n(\theta_n) + \lambda \mathbf{I}\big)^{-1} \nabla \ell(z, \theta_n)\,.$$

We choose the smallest λ so that the quadratic objective $g_n(u_t)$ from (9) is bounded below for iterates u_t obtained from SGD, ensuring that $H + \lambda \mathbf{I}$ is positive semidefinite.

Error Criterion. The norm $\|\hat{I}_n(z) - I(z)\|$ bound may be vacuous for failing to capture the permutation symmetries of the parameters of a deep network. Instead, we measure the effect of a point z on a test function $h(\theta) = \ell(z_{\text{test}}, \theta)$ as

$$G_n(z) = \langle \nabla h(\theta_n), I_{n,\lambda}(z) \rangle\,, \tag{13}$$

and compare it against its population counterpart $G(z)$. From the chain rule, it follows that $G(z)$ is the linearization $\frac{\mathrm{d}}{\mathrm{d}\varepsilon} h(\theta_{n,\varepsilon,z})|_{\varepsilon=0}$ similar to (3). In our experiments, $h(\theta)$ is the loss on the test set. The results are given in Figure 3.

Results: Total Error Versus n. For the question-answering task, the error reduces by a factor of 10 as n increases from 40 to 300 (slope ≈ -1.5) indicating an empirical $n^{-1.5}$ rate. For the text continuation task, we find that the error in influence estimation does not vary significantly with n and

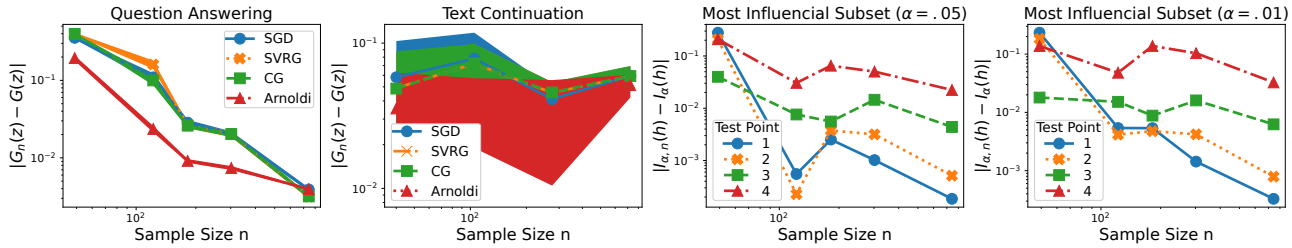

Figure 3: **Left two**: Convergence of the approximate empirical influence to the population for text generation tasks measured in terms of predictions as in (13). The solid line denotes the mean of $|G_n(z_i) - G(z_i)|$ for $i = 1, \cdots, 4$ and the shaded area denotes its standard deviation. **Right two**: Convergence of the influence value $I_{\alpha,n}(h_i)$ found by the most influential subset method to its population version $I_\alpha(h_i)$ on the question-answering task for different test functions $h_i = \ell(z_{\text{test},i}, \theta)$.

has a high variance. Indeed, the open-ended nature of the text continuation task suggests that no one point z should have a large influence on the predictions of a test point z_{test}, leading to noisy influence estimates.

Comparing Computational Approximations. We observe that SGD $\approx$ SVRG in Figure 3. This corroborates the total error bounds of Table 1 which show that variance-reduced SVRG has the same total error as SGD despite being significantly faster in optimization. At a large computation budget, we find that the conjugate gradient method also exhibits an error comparable to SGD and SVRG. The benefits of stochastic algorithms such as SGD become evident for large datasets where SGD gives a reasonable estimate without even making a full pass (its error is independent of n, cf. Table 1). For the question-answering dataset, we find that the low-rank approximation provided by the Arnoldi method (Schioppa et al., 2022) has the smallest error for $n \leq 200$, while it is identical to the others for large n.

Most Influential Subsets. We repeat the question-answering experiment to find the most influential subset of data for different n with test function $h_i(\theta) = \ell(z_{\text{test},i}, \theta)$ for four chosen test points. We use the low-rank (Arnoldi) method to approximate the inverse Hessian-vector product because this method has the best error properties in Figure 3 (left two). For different values of α, we observe that the estimation error tends to decrease with n. We note that a few outliers are to be expected with large-scale deep nets with real data where theoretical assumptions are not precisely met.

The type of influential examples recovered varied from surface-level attributes to deeper features, such as topics, as n increased; see Figure 4 for examples. In some cases, the most influential examples were semantically related questions with different answers. For instance, for the test question "Was Goldmoon male or female" (female), a highly negatively influential questions was "What is the gender of Jacques Rivard?" (male). However, for others the relations seemed more structural. For example, the test question "The nationality of Jean-Louis Laya was what?" (French), we

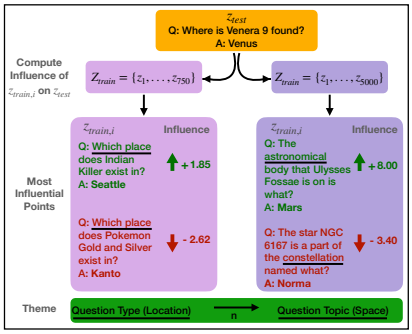

Figure 4: As the sample size n increases, we see a shift in the quality of the most influential questions. Lower n results in surface-level attributes, such as question type, while larger n results in deeper features, such as the topic.

recovered as highly negatively influential, "The nationality of Yitzhak Rabin is?" (Hebrew).

6 CONCLUSION

As statistical learning models and deep nets are being increasingly used, influence diagnostics are precious tools to study the influence of datapoints on predictions, decisions, and outcomes. In this paper, we presented statistical and computational guarantees for influence functions for generalized linear models. We established the statistical consistency of most influential subsets method (Broderick et al., 2020) together with nonasymptotic bounds. We illustrated our results on simulated and real datasets. Extending our results to sparse regularized models as well as deep neural network models are interesting venues for future work.

Acknowledgements This work was supported by NSF DMS-2023166, CCF-2019844, DMS-2052239, DMS-2134012, DMS-2133244, NIH, CIFAR-LMB, and research awards. Part of this work was done while Z. Harchaoui was visiting the Simons Institute for the Theory of Computing, and while K. Pillutla was at the University of Washington.

References

N. Agarwal, B. Bullins, and E. Hazan. Second-Order Stochastic Optimization for Machine Learning in Linear Time. *Journal of Machine Learning Research*, 18: 116:1–116:40, 2017.

Z. Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):8194–8244, 2017.

M. Angelucci and G. De Giorgi. Indirect effects of an aid program: How do cash transfers affect ineligibles' consumption? *American Economic Review*, 99(1):486–508, March 2009.

W. E. Arnoldi. The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quarterly of applied mathematics*, 9(1):17–29, 1951.

M. Avella-Medina. Influence functions for penalized M-estimators. *Bernoulli*, 23(4B):3178 – 3196, 2017. doi: 10.3150/16-BEJ841.

O. Axelsson and I. Kaporin. On the sublinear and superlinear rate of convergence of conjugate gradient methods. *Numerical Algorithms*, 25(1), 2000.

F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4, 2010.

F. Bach. *Learning Theory from First Principles*. Online version, 2021.

Z.-Z. Bai and J.-Y. Pan. *Matrix analysis and computations*. SIAM, 2021.

D. Bertsekas. *Convex optimization algorithms*. Athena Scientific, 2015.

T. Broderick, R. Giordano, and R. Meager. An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference? *arXiv Preprint*, 2020.

K. Chen. *Matrix Preconditioning Techniques and Applications*. Cambridge University Press, 2005.

R. Cook and S. Weisberg. *Residuals and influence in regression*. New York: Chapman and Hall, New York: Chapman Hall, 1982.

R. D. Cook. Assessment of local influence. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48 (2):133–155, 1986.

G. B. Dantzig. Discrete-variable extremum problems. *Operations research*, 5(2):266–288, 1957.

N. De Cao, W. Aziz, and I. Titov. Editing factual knowledge in language models, 2021. arXiv Preprint.

A. Finkelstein, S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. Allen, K. Baicker, and O. H. S. Group. The Oregon Health Insurance Experiment: Evidence from the First Year. *The Quarterly Journal of Economics*, 127(3):1057–1106, 07 2012.

S. Fortunati, F. Gini, and M. S. Greco. The misspecified Cramer-Rao bound and its application to scatter matrix estimation in complex elliptically symmetric distributions. *IEEE Transactions on Signal Processing*, 64:2387–2399, 2016.

R. Giordano, W. Stephenson, R. Liu, M. Jordan, and T. Broderick. A Swiss Army Infinitesimal Jackknife. In *International Conference on Artificial Intelligence and Statistics*, pages 1139–1147. PMLR, 2019.

A. S. Hadi, W. D. Jones, and R. F. Ling. A unifying representation of some case-deletion influence measures in univariate and multivariate linear regression. *Journal of statistical planning and inference*, 46(1):123–135, 1995.

F. R. Hampel. Contributions to the theory of robust estimation. *PhD Dissertation*, 1968.

F. R. Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

J. F. Henriques, S. Ehrhardt, S. Albanie, and A. Vedaldi. Small steps and giant leaps: Minimal newton solvers for deep learning. In *IEEE/CVF International Conference on Computer Vision*, pages 4763–4772, 2019.

T. Hofmann, A. Lucchi, S. Lacoste-Julien, and B. McWilliams. Variance Reduced Stochastic Gradient Descent with Neighbors. *Advances in Neural Information Processing Systems*, 28, 2015.

L. A. Jaeckel. *The infinitesimal jackknife*. Bell Telephone Laboratories, 1972.

P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, V. K. Pillutla, and A. Sidford. A Markov Chain Theory Approach to Characterizing the Minimax Optimality of Stochastic Gradient Descent (for Least Squares). In *Conference on Foundations of Software Technology and Theoretical Computer Science*, pages 2:1–2:10, 2017a.

P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Parallelizing stochastic gradient descent for least squares regression: Mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 18:223:1–223:42, 2017b.

R. Johnson and T. Zhang. Accelerating Stochastic Gradient Descent using Predictive Variance Reduction. *Advances in neural information processing systems*, 26, 2013.

P. W. Koh and P. Liang. Understanding Black-box Predictions via Influence Functions. In *International Conference on Machine Learning*, volume 70, pages 1885–1894, 2017.

Y. Laguel, K. Pillutla, J. Malick, and Z. Harchaoui. Superquantiles at Work: Machine Learning Applications and Efficient Subgradient Computation. *Set-Valued and Variational Analysis*, 29(4):967–996, 2021.

C. Lanczos. An Iteration Method for the Solution of the Eigenvalue Problem of Linear Differential and Integral

Operators. *Journal of Research of the National Bureau of Standards*, 1950.

O. Levy, M. Seo, E. Choi, and L. Zettlemoyer. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics. doi: 10.18653/v1/K17-1034. URL https://aclanthology.org/K17-1034.

M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *ACL*, pages 7871–7880, 2020.

H. Lin, J. Mairal, and Z. Harchaoui. Catalyst Acceleration for First-order Convex Optimization: from Theory to Practice. *Journal of Machine Learning Research*, 18 (212):1–54, 2018.

G. Louvet, J. Raymaekers, G. Van Bever, and I. Wilms. The influence function of graphical lasso estimators. *COMPSTAT*, 2022.

P. Ma, M. Mahoney, and B. Yu. A statistical perspective on algorithmic leveraging. In *International Conference on Machine Learning*, pages 91–99. PMLR, 2014.

Y. Ma. distilgpt2-finetuned-wikitext2. https://huggingface.co/MYX4567/distilgpt2-finetuned-wikitext2, 2021.

L. Mackey, M. I. Jordan, R. Y. Chen, B. Farrell, and J. A. Tropp. Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability*, 42(3): 906–945, 2014.

S. Merity, C. Xiong, J. Bradbury, and R. Socher. Pointer Sentinel Mixture Models. In *ICLR*, 2017.

Y. E. Nesterov and A. Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13. SIAM, 1994.

D. M. Ostrovskii and F. Bach. Finite-sample analysis of M-estimators using self-concordance. *Electronic Journal of Statistics*, 15(1), 2021.

R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

P. J. Rousseeuw, F. R. Hampel, E. M. Ronchetti, and W. A. Stahel. *Robust statistics: the approach based on influence functions*. John Wiley and Sons, 2011.

C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Y. Saad. *Iterative methods for sparse linear systems*. SIAM, 2003.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *NeurIPS EMC^2 Workshop*, 2019.

A. Schioppa, P. Zablotskaia, D. Vilar, and A. Sokolov. Scaling up influence functions. In *AAAI Conference on Artificial Intelligence*, volume 36, pages 8179–8186, 2022.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

J. Zhao, C. Liu, L. Niu, and C. Leng. Multiple influential point detection in high dimensional regression spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(2):385–408, 2019.

H. Zhu and H. Zhang. A diagnostic procedure based on local influence. *Biometrika*, 91(3):579–589, 2004.

Appendix

Table of Contents

A Notation Review

Setup. We review notation from the paper, which will be used throughout the appendix. We define the parameter of interest $\theta_\star \in \Theta = \mathbb{R}^p$ as

$$\theta_\star := \arg\min_{\theta\in\Theta} \Big[F(\theta) := \mathbb{E}_{Z\sim P}\left[\ell(Z,\theta)\right] \Big],$$

where P is an unknown probability distribution over a data space $\mathcal{Z}$ and $\ell : \mathcal{Z} \times \Theta \to \mathbb{R}_+$ is a loss function. We define the estimate of $\theta_\star$ using an i.i.d. sample $Z_{1:n} := (Z_1, \cdots, Z_n) \sim P^n$ as

$$\theta_n := \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \ell(Z_i,\theta).$$

We define the gradient of the loss function as $S(z,\theta) = \nabla_\theta \ell(Z,\theta)$ and the empirical gradient of the loss function as $S_n(\theta) := \frac{1}{n}\sum_{i=1}^n \nabla_\theta \ell(Z_i,\theta)$.

We define the population Hessian $H_\star = \nabla^2_{\theta_\star}\ell(z,\theta_\star)$ of the population objective and the estimate of the Hessian as $H_n(\theta) := \frac{1}{n}\sum_{i=1}^n \nabla^2_\theta \ell(Z_i,\theta)$.

Influence Function. We define $G_\star = \mathrm{Cov}_{Z\sim P}(\nabla_{\theta_\star}\ell(Z,\theta_\star))$ the gradient covariance at $\theta_\star$ and the effective dimension $p_\star = \mathbf{Tr}(H_\star^{-1/2} G_\star H_\star^{-1/2})$. We define the population influence function as $I(z) := H_\star^{-1}\nabla_{\theta_\star}\ell(z,\theta_\star)$. We quantify the influence of a fixed data point z on the estimator θ_n as $I_n(z)$ defined as

$$I_n(z) = -H_n(\theta_n)^{-1}\nabla\ell(z,\theta_n)\,.$$

Most Influential Subset. Let $\alpha \in (0,1)$ and $h : \mathbb{R}^p \to \mathbb{R}$ be a continuously differentiable test function. Then we define the weights w in the probability simplex $\Delta^{n-1}\theta_{n,w} := \arg\min_{\theta\in\Theta}\sum_{i=1}^n w_i\ell(Z_i,\theta)$ and use them to define W_α as

$$W_\alpha := \left\{ w \in \Delta^{n-1} \;:\; \begin{array}{l}\text{at most } \alpha n \text{ elements of } w \text{ are}\\ \text{zero and the rest are equal}\end{array} \right\}.$$

The maximum influence of any subset of data of size at most αn for a test function h is expressed by

$$I_{\alpha,n}(h) = \max_{w\in W_\alpha}\left\{ -\sum_{i=1}^n w_i \big\langle \nabla h(\theta_n), H_n(\theta_n)^{-1}\nabla\ell(Z_i,\theta_n)\big\rangle \right\}.$$

The population subset influence is defined as,

$$I_\alpha(h) = S_\alpha\Big[-\nabla h(\theta_\star)^\top H(\theta_*)^{-1}\nabla\ell(Z,\theta_\star)\Big], \tag{14}$$

where S_α is the superquantile at level α. We refer to Appendix I.6.

Miscellaneous. An unqualified norm $\|\cdot\|$ refers to the Euclidean norm $\|v\|_2$ for a vector v and the spectral norm $\|M\|_2$ for a matrix M. We define a vector norm $\|x\|_A = \langle x, Ax\rangle$ and matrix norm $\|B\|_A = \|A^{1/2}BA^{1/2}\|_2$ for a positive definite A. We define the convex hull as $\mathrm{conv}\, T$ for a set $T \subset \mathbb{R}^n$.

We define $\mathbb{V}(M) = \mathbb{E}[MM^\top] - \mathbb{E}[M]\mathbb{E}[M]^\top$ for a random matrix M. We also denote $\mathrm{d}Q/\mathrm{d}P$ as the Radon-Nikodym derivative of Q w.r.t. P. When P and Q have respective densities p, q, we have $\mathrm{d}Q/\mathrm{d}P(z) = q(z)/p(z)$ as simply the density ratio or likelihood ratio.

Lastly, we define the condition number of a positive definite matrix A with spectral norm $\|A\|_2 \le L$ and minimum eigenvalue $\lambda_{\min}(A)$ as $\kappa = L/\lambda_{\min}(A)$.

B Review of Computational Approaches

We present the pseudocode of the various computational approaches we consider in this work:

- Algorithm 1: Conjugate gradient method,

- Algorithm 2: Stochastic gradient descent,
- Algorithm 3: LiSSA,
- Algorithm 4: Stochastic variance-reduced gradient (SVRG) method,
- Algorithm 5: Low-rank approximation via the Arnoldi/Lanczos iterations.

Algorithm 1 Conjugate Gradient Method to Compute the Influence Function

Input: vector v, batch Hessian vector product oracle $\text{HVP}_n(u) = H_n(\theta_n)u$, number of iterations T

$u_0 = 0$, $r_0 = -v - \text{HVP}_n(u_0)$, $d_0 = r_0$
for $t = 0, ..., T-1$ **do**
  $\alpha_t = \frac{d_t^\top r_t}{d_t^\top \text{HVP}_n(d_t)}$
  $u_{t+1} = u_t + \alpha_t d_t$
  $r_{t+1} = -v - \text{HVP}_n(u_{t+1})$
  $\beta_t = \frac{r_{t+1}^\top r_{t+1}}{r_t^\top r_t}$
  $d_{t+1} = r_{t+1} + \beta_t d_t$
return u_T

Algorithm 2 Stochastic Gradient Descent Method to Compute the Influence Function

Input: vector v, Hessian vector product oracle $\text{HVP}(i, u) = \nabla^2 \ell(z_i, \theta_n)u$, number of iterations T, learning rate γ

$u_0 = 0$
for $t = 0, ..., T-1$ **do**
  Sample $i_t \sim \text{Unif}([n])$
  $u_{t+1} = u_t - \gamma(\text{HVP}(i_t, u_t) + v)$
return u_T

Algorithm 3 The LiSSA Method to Compute the Influence Function (Agarwal et al., 2017)

Input: vector v, Hessian vector product oracle $\text{HVP}(i, u) = \nabla^2 \ell(z_i, \theta_n)u$, number of approximations S, number of iterations T, scaling factor γ

for $s = 1, ..., S$ **do**
  $u_0^{(s)} = -v$
  for $t = 0, ..., T-1$ **do**
    Sample $i_t \sim \text{Unif}([n])$
    $u_{t+1}^{(s)} = -\gamma v + u_t^{(s)} - \gamma\, \text{HVP}(i_t, u_t^{(s)})$
$u_T = \frac{1}{S}\left(\sum_{s=1}^S u_T^{(s)}\right)$
return u_T

Connection between SGD and LiSSA. Observe that the updates of LiSSA for a fixed s are identical to that of SGD:

$$u_{t+1}^{(s)} = -\gamma v + u_t^{(s)} - \gamma \text{HVP}(i_t, u_t^{(s)}) = u_t^{(s)} - \gamma(\text{HVP}(i_t, u_t^{(s)}) + v)\,.$$

Formally, we show that the sequence $u_1, ..., u_t$ produced by stochastic gradient descent with initial guess $u_0 = -v$ (instead of $u_0 = 0$ as required by Algorithm 2) and $u_1', ..., u_t'$ produced by LiSSA with number of repetitions $S = 1$ are identical. Note that $u_0 = u_0' = -v$. We show by induction that the two sequences (u_t) and (u_t') are identical provided the same samples $i_0, \cdots, i_{T-1}$ are drawn. Suppose $u_t = u_t'$ for some $t \geq 0$. We have,

$$u_{t+1}' = -\gamma v + u_t' - \gamma \text{HVP}(i_t, u_t') = u_t' - \gamma(\text{HVP}(i_t, u_t') + v) = u_t - \gamma(\text{HVP}(i_t, u_t) + v) = u_{t+1},$$

showing that the sequences are identical.

Algorithm 4 Stochastic Variance Reduced Gradient Method to Compute the Influence Function

Input: vector v, Hessian vector product oracle $\text{HVP}(i, u) = \nabla^2 \ell(z_i, \theta_n)u$, number of epochs S, number of iterations per epoch T, learning rate γ

$u_T^{(0)} = 0$
for $s = 1, 2, ..., S$ **do**
  $u_0^{(s)} = u_T^{(s-1)}$
  $\tilde{u}_0^{(s)} = \frac{1}{n} \sum_{i=1}^{n} \text{HVP}(u_0^{(s)}) - v$
  for $t = 0, ..., T-1$ **do**
    Sample $i_t \sim \text{Unif}([n])$
    $u_{t+1}^{(s)} = u_t^{(s)} - \gamma(\text{HVP}(i_t, u_t^{(s)}) - \text{HVP}(i_t, u_0^{(s)}) + \tilde{u}_0^{(s)})$
return $u_T^{(S)}$

Algorithm 5 Arnoldi Method to Compute the Influence Function (Schioppa et al., 2022)

Input: vector v, test function h, initial guess u_0, batch Hessian vector product oracle $\text{HVP}_n(u) = H_n(\theta_n)u$, number of top eigenvalues k, number of iterations T

Output: An estimate of $\langle \nabla h(\theta), H_n(\theta_n)^{-1} v \rangle$

Obtain Λ, G = ARNOLDI(u_0, T, k) ▷ Cache the results for future calls
return $\langle G \nabla h(\theta), \Lambda^{-1} G v \rangle$

procedure ARNOLDI(u_0, T, k)
  $w_0 = 1 = u_0 / \|u_0\|_2$
  $A = \mathbf{0}_{T+1 \times T}$
  for $t = 1, ..., T$ **do**
    Set $u_t = \text{HVP}_n(w_t) - \sum_{j=1}^{t} \langle u_t, w_j \rangle w_j$
    Set $A_{j,t} = \langle u_t, w_j \rangle$ for $j = 1, \ldots, t$ and $A_{t+1,t} = \|u_t\|_2$
    Update $w_{t+1} = u_t / \|u_t\|$
  Set $\tilde{A} = A[1:T, :] \in \mathbb{R}^{T \times T}$ (discard the last row)
  Compute an eigenvalue decomposition $\tilde{A} = \sum_{j=1}^{T} \lambda_j e_j e_j^\top$ with λ_j's in descending order
  Define $G : \mathbb{R}^p \to \mathbb{R}^k$ as the operator $Gu = \big(\langle u, W^\top e_1 \rangle, \cdots, \langle u, W^\top e_k \rangle\big)$, where $W = (w_1^\top; \cdots; w_T^\top) \in \mathbb{R}^{T \times p}$
  return diagonal matrix $\Lambda = \text{Diag}(\lambda_1, \cdots, \lambda_k)$ and the operator G

C Effective Dimensions and Eigenspectra of the Hessian and Gradient Covariance

Recall the following definitions, the population Hessian $H_\star = \nabla^2 F(\theta_\star)$ of the population objective and $G_\star = \text{Cov}_{Z \sim P}(\nabla \ell(Z, \theta_\star))$ the gradient covariance at $\theta_\star$. We are interested in how the effective dimension $p_\star = \text{Tr}(H_\star^{-1/2} G_\star H_\star^{-1/2})$ differs from the parameter dimension p due to the eigendecay of $H_\star$. First, we assume that $H_\star$ and $G_\star$ share the same eigenvectors. Then, using the eigenvalue decomposition of a matrix, we can say that for Q containing the eigenvectors as its columns,

$$H_\star = Q \Lambda_H Q^\top,$$
$$G_\star = Q \Lambda_G Q^\top$$

where $\Lambda_A = \text{Diag}\{\lambda_{a,i}\}$ contains the eigenvalues of A in non-increasing order. Therefore, we get

$$H_\star^{-1/2} = Q \Lambda_H^{-1/2} Q^\top .$$

Using these definitions we now show the following,

$$\begin{aligned} H_\star^{-1/2} G_\star H_\star^{-1/2} &= (Q \Lambda_H^{-1/2} Q^\top)(Q \Lambda_G Q^\top)(Q \Lambda_H^{-1/2} Q^\top) \\ &= Q \Lambda_H^{-1/2} \Lambda_G \Lambda_H^{-1/2} Q^\top \\ &= Q \text{Diag}\left\{ \frac{\lambda_{g,1}}{\lambda_{h,1}} ... \frac{\lambda_{g,p}}{\lambda_{h,p}} \right\} Q^\top . \end{aligned}$$

Therefore, due to the cyclic property of traces we define,

$$\mathbf{Tr}(H_\star^{-1/2} G_\star H_\star^{-1/2}) = \sum_{i=1}^{p} \frac{\lambda_{g,i}}{\lambda_{h,i}}.$$

Here we have shown that the dimension dependency of $p_\star$ is dependent on the eigendecay of $G_\star$ and $H_\star$. To illustrate this point, we show four examples of how these calculations continue. All examples are outlined in Table 2.

Polynomial - Polynomial Eigendecay. We assume that both $G_\star$ and $H_\star$ have polynomial eigendecay, that is, $\lambda_{g,i} \lesssim i^{-\alpha}$ and $\lambda_{h,i} \lesssim i^{-\beta}$. Then we can write,

$$p_\star \lesssim \sum_{i=1}^{p} i^{\beta-\alpha} \lesssim \int_1^p x^{\beta-\alpha} \mathrm{d}x \lesssim p^{\beta-\alpha+1}.$$

Polynomial - Exponential Eigendecay. We assume that $G_\star$ has polynomial eigendecay and $H_\star$ have exponential eigendecay, that is $\lambda_{g,i} \lesssim i^{-\alpha}$ and $\lambda_{h,i} \lesssim e^{-\nu i}$. Then we can write,

$$p_\star \lesssim \sum_{i=1}^{p} e^{\nu i} i^{-\alpha} \lesssim p^{1-\alpha} e^{\nu p},$$

where the last inequality holds because $e^{\nu x} x^{-\alpha}$ is increasing when x is large enough.

Exponential - Polynomial Eigendecay. We assume that $G_\star$ has exponential eigendecay and $H_\star$ have polynomial eigendecay, that is $\lambda_{g,i} \lesssim e^{-\mu i}$ and $\lambda_{h,i} \lesssim i^{-\beta}$. Then we can write,

$$p_\star \lesssim \sum_{i=1}^{p} e^{-\mu i} i^{\beta} \lesssim 1,$$

where the last inequality holds because $e^{-\mu x} x^{\beta}$ is decreasing when x is large enough.

Exponential - Exponential Eigendecay. We assume that $G_\star$ has exponential eigendecay and $H_\star$ have exponential eigendecay, that is $\lambda_{g,i} \lesssim e^{-i\mu}$ and $\lambda_{h,i} \lesssim e^{-i\nu}$. Then we can write,

$$p_\star \lesssim \sum_{i=1}^{p} e^{(\nu-\mu)i}.$$

If $\mu > \nu$, then

$$\sum_{i=1}^{p} e^{(\nu-\mu)i} \lesssim 1.$$

If $\mu < \nu$, then

$$\sum_{i=1}^{p} e^{(\nu-\mu)i} \lesssim \int_1^p e^{(\nu-\mu)i} = \frac{1}{\nu-\mu}\left(e^{(\nu-\mu)p} - e^{(\nu-\mu)}\right) \lesssim e^{(\nu-\mu)p}.$$

And if $\mu = \nu$, then

$$\sum_{i=1}^{p} e^{0} = p.$$

D Statistical Error Bounds for Influence Estimation

The main purpose of this section is to prove the statistical error bound Theorem 1. We use C to denote an absolute constant which may change from line to line. We use subscripts to emphasize the dependency on problem-specific constants, e.g., C_{K_1} is a constant that only depends on K_1.

Table 2: Comparison between the effective dimension $p_\star$ and the parameter dimension p in different regimes of eigendecays of $G_\star$ and $H_\star$ assuming they share the same eigenvectors.

	Eigendecay		**Dimension Dependency**		**Ratio**
	$G_\star$	$H_\star$	$p_\star$	p	$p_\star/p$
Poly-Poly	$i^{-\alpha}$	$i^{-\beta}$	$p^{(\beta-\alpha+1)\vee 0}$	p	$p^{(\beta-\alpha)\vee(-1)}$
Poly-Exp	$i^{-\alpha}$	$e^{-\nu i}$	$p^{1-\alpha}e^{\nu p}$	p	$p^{-\alpha}e^{\nu p}$
Exp-Poly	$e^{-\mu i}$	$i^{-\beta}$	1	p	p^{-1}
Exp-Exp	$e^{-\mu i}$	$e^{-\nu i}$	p if $\mu=\nu$ 1 if $\mu>\nu$ $e^{(\nu-\mu)p}$ if $\mu<\nu$	p	1 if $\mu=\nu$ p^{-1} if $\mu>\nu$ $p^{-1}e^{(\nu-\mu)p}$ if $\mu<\nu$

Notation. Let z be a fixed data point not related to the sample $Z_1, \cdots, Z_n \sim P$. Recall that the influence of upweighting an observation z on the model parameter θ is given by

$$I_n(z) = -H_n(\theta_n)^{-1} S(z, \theta_n), \tag{15}$$

where $H_n(\theta) := \frac{1}{n}\sum_{i=1}^n \nabla^2_\theta \ell(Z_i, \theta)$ is the empirical Hessian and $S(z,\theta) := \nabla_\theta \ell(z,\theta)$ is the gradient at z. Let $\theta_\star$ be the minimizer (assumed to exist) of the population risk $\mathbb{E}[\ell(z,\theta)]$ and $H(\theta) := \mathbb{E}[\nabla^2_\theta \ell(z,\theta)]$. We write $H_\star := H(\theta_\star)$ for short. We are interested in bounding the difference

$$\mathcal{E} := \|H_n(\theta_n)^{-1} S(z,\theta_n) - H_\star^{-1} S(z,\theta_\star)\|_{H_\star},$$

where $\|u\|_A := \sqrt{u^\top A u}$ for a vector u and a positive semidefinite matrix A.

D.1 Assumptions

We state the full assumptions under which the statistical bound holds.

Assumption 1. *For any $z \in \mathcal{Z}$, the loss function $\ell(z,\cdot)$ is pseudo self-concordant for some $R \geq 1$:*

$$|D^3_\theta \ell(z,\theta)[u,u,v]| \leq R\|u\|^2_{\nabla^2\ell(z,\theta)} \|v\|_2,$$

where $D^3_x f(x)[u,v,w] := \frac{\mathrm{d}}{\mathrm{d}t}\langle u, \nabla^2 f(x+tw)\, v\rangle|_{t=0}$ for f thrice continuously differentiable.

The most useful consequence of this assumption is a spectral approximation of the Hessian $(1/2)H(\theta') \preceq H(\theta) \preceq 2\,H(\theta')$ for θ and θ' close enough in terms of the L_2 distance.

Assumption 2. *(Sub-Gaussian Gradient). There exists a constant $K_1 \geq 1$ such that the normalized gradient $H(\theta_\star)^{-1/2}\nabla\ell(Z,\theta_\star)$ at $\theta_\star$ is sub-Gaussian with parameter K_1 (see Appendix I.1 for a precise definition).*

Assumption 3. *(Matrix Bernstein of Hessian). The standardized Hessian $H(\theta_\star)^{-1/2}\,\nabla^2\ell(Z,\theta_\star)\,H(\theta_\star)^{-1/2} - \mathbf{I}_p$ at $\theta_\star$ satisfies a Bernstein condition with parameter $K_2 \geq 1$ (see Appendix I.1 for a definition). Moreover,*

$$\sigma_H^2 := \left\| \mathbb{V}\big(H(\theta_\star)^{-1/2}\,\nabla^2\ell(Z,\theta_\star)\,H(\theta_\star)^{-1/2}\big) \right\|_2$$

is finite, where we denote $\mathbb{V}(M) = \mathbb{E}[MM^\top] - \mathbb{E}[M]\,\mathbb{E}[M]^\top$ for a random matrix M.

D.2 Proof of the Statistical Bound of Theorem 1

We now state and prove the full version of Theorem 1. Note that this bound is stated in terms of the $H_\star$ norm, but without the square.

Theorem 1. *Under Assumptions 1,2, and 3, we have, with probability at least $1-\delta$,*

$$\mathcal{E} \leq C_{K_1,K_2,\sigma_H} \log\left(\frac{2p}{\delta}\right)\sqrt{\log\left(\frac{e}{\delta}\right)}\left(1 + R\sqrt{\frac{p_\star}{\mu_\star}}\right)\sqrt{\frac{p_\star}{n}}$$

whenever $n \geq C_{K_1,K_2,\sigma_H}\left(\frac{p_\star}{\mu_\star}R^2\log\left(\frac{e}{\delta}\right) + \log\left(\frac{2p}{\delta}\right)\right)$, where $p_\star := \mathbf{Tr}\{H_\star^{-1/2}G_\star H_\star^{-1/2}\}$ and $\mu_\star = \lambda_{\min}(H_\star)$.

Proof. Define

$$
\begin{aligned}
r_n &:= \sqrt{CK_1^2 \log(2e/\delta)\frac{p_\star}{n}} \\
t_n &:= \frac{2\sigma_H^2}{-K_2 + \sqrt{K_2^2 + 2\sigma_H^2 n/\log(4p/\delta)}}.
\end{aligned} \tag{16}
$$

Note that they both decay as $O(n^{-1/2})$. The proof consists of several key steps.

Step 1. Upper bound $\mathcal{E}$ by basic terms involving the standardized gradient and the standardized Hessian. By the triangle inequality, it holds that

$$
\mathcal{E} \le \|(H_n(\theta_n)^{-1} - H_\star^{-1})S(z,\theta_n)\|_{H_\star} + \|H_\star^{-1}(S(z,\theta_n) - S(z,\theta_\star))\|_{H_\star}. \tag{17}
$$

The first term in (17) can be upper bounded by

$$
\|[H_n(\theta_n)^{-1} - H_\star^{-1}][S(z,\theta_n) - S(z,\theta_\star)]\|_{H_\star} + \|[H_n(\theta_n)^{-1} - H_\star^{-1}]S(z,\theta_\star)\|_{H_\star}. \tag{18}
$$

By the triangle inequality again, it can be shown that, for any $v \in \mathbb{R}^p$,

$$
\begin{aligned}
\|[H_n(\theta_n)^{-1} - H_\star^{-1}]v\|_{H_\star} &= \|[H_\star^{1/2}H_n^{-1}(\theta_n)H_\star^{1/2} - H_\star^{-1/2}H_\star^{1/2}]H_\star^{-1/2}v\|_2 \\
&\le \|H_\star^{1/2}H_n^{-1}(\theta_n)H_\star^{1/2} - \mathbf{I}_p\|_2 \|H_\star^{-1/2}v\|_2.
\end{aligned}
$$

As a result, (18) can be further upper bounded by

$$
\underbrace{\|H_\star^{1/2}H_n(\theta_n)^{-1}H_\star^{1/2} - \mathbf{I}_p\|_2}_{A_3}\left\{\underbrace{\|H_\star^{-1/2}[S(z,\theta_n) - S(z,\theta_\star)]\|_2}_{A_2} + \underbrace{\|H_\star^{-1/2}S(z,\theta_\star)\|_2}_{A_1}\right\}.
$$

Similarly, the second term in (17) can be upper bounded by

$$
\|H_\star^{-1/2}[S(z,\theta_n) - S(z,\theta_\star)]\|_2 = A_2.
$$

Hence, it suffices to bound the three terms A_1, A_2, and A_3. For that purpose, we define the following events

$$
\begin{aligned}
\mathcal{G}_1 &:= \left\{\|H_\star^{-1/2}S(z,\theta_\star)\|_2^2 \le CK_1^2 \log(e/\delta)p_\star\right\} \\
\mathcal{G}_2 &:= \left\{\|\theta_n - \theta_\star\|_{H_\star}^2 \le CK_1^2 \log(e/\delta)\frac{p_\star}{n}\right\} \\
\mathcal{G}_3 &:= \left\{\|H_\star^{-1/2}H(z,\theta_\star)H_\star^{-1/2} - \mathbf{I}_p\|_2 \le t_1\right\} \\
\mathcal{G}_4 &:= \left\{\|H_\star^{1/2}H_n(\theta_n)^{-1}H_\star^{1/2} - \mathbf{I}_p\|_2 \le \frac{Rr_n/\sqrt{\mu_\star} + t_n}{1 - Rr_n/\sqrt{\mu_\star} - t_n}\right\}.
\end{aligned}
$$

Moreover, we assume $n \ge \max\{4(K_2 + 2\sigma_H^2)\log(16p/\delta), CK_1^2\log(e/\delta)p_\star R^2/\mu_\star\}$ throughout the proof. In the following, we bound A_1, A_2, A_3 on the event $\mathcal{G}_1\mathcal{G}_2\mathcal{G}_3\mathcal{G}_4$, and control the probability of this event.

Step 2. Control A_1. On the event $\mathcal{G}_1$, we know

$$
A_1 \le \sqrt{CK_1^2 \log(e/\delta)p_\star}.
$$

Step 3. Control A_2. According to Taylor's theorem, it holds that

$$
S(z,\theta_n) - S(z,\theta_\star) = H(z,\bar{\theta})(\theta_n - \theta_\star),
$$

where $\bar{\theta} \in \mathrm{Conv}\{\theta_n, \theta_\star\}$. Therefore, we can rewrite A_2 as

$$
\begin{aligned}
A_2 &= \|H_\star^{-1/2}H(z,\bar{\theta})(\theta_n - \theta_\star)\|_2 \\
&= \|H_\star^{-1/2}H(z,\bar{\theta})H_\star^{-1/2}H_\star^{1/2}(\theta_n - \theta_\star)\|_2.
\end{aligned}
$$

Consequently,

$$A_2 \leq \|H_\star^{-1/2} H(z, \bar{\theta}) H_\star^{-1/2}\|_2 \|H_\star^{1/2}(\theta_n - \theta_\star)\|_2.$$

According to Proposition 32, we have

$$e^{-R\|\bar{\theta}-\theta_\star\|_2} H(z, \theta_\star) \preceq H(z, \bar{\theta}) \preceq e^{R\|\bar{\theta}-\theta_\star\|_2} H(z, \theta_\star).$$

Note that $R\|\bar{\theta} - \theta_\star\|_2 \leq R\|\theta_n - \theta_\star\|_2 \leq R\mu_\star^{-1/2}\|\theta_n - \theta_\star\|_{H_\star}$. It follows from the event $\mathcal{G}_2$ that

$$\frac{1}{2} H(z, \theta_\star) \preceq H(z, \bar{\theta}) \preceq 2H(z, \theta_\star). \tag{19}$$

As a result, we have

$$\|H_\star^{-1/2} H(z, \bar{\theta}) H_\star^{-1/2}\|_2 \leq 2\|H_\star^{-1/2} H(z, \theta_\star) H_\star^{-1/2}\|_2.$$

On the event $\mathcal{G}_3$, we know

$$\|H_\star^{-1/2} H(z, \theta_\star) H_\star^{-1/2}\|_2 \leq 1 + t_1. \tag{20}$$

Therefore, by the event $\mathcal{G}_2$ and (20), A_2 is upper bounded by

$$A_2 \leq C(1 + t_1) r_n.$$

Step 4. Control A_3. On the event $\mathcal{G}_4$, we have

$$A_3 \leq \frac{Rr_n/\sqrt{\mu_\star} + t_n}{1 - Rr_n/\sqrt{\mu_\star} - t_n}.$$

Step 5. Control the probability of the event $\mathcal{G}_1\mathcal{G}_2\mathcal{G}_3\mathcal{G}_4$.

Event $\mathcal{G}_1$. Since $\theta_\star$ is a minimizer of the population risk, then, by the first order optimality condition, we have $E[S(z, \theta_\star)] = 0$. Moreover, we have

$$\begin{aligned}
\mathrm{Cov}(G_\star^{-1/2} S(z, \theta_\star)) &= E[G_\star^{-1/2} S(z, \theta_\star) S(z, \theta_\star)^\top G_\star^{-1/2}] \\
&= G_\star^{-1/2} E[S(z, \theta_\star) S(z, \theta_\star)^\top] G_\star^{-1/2} \\
&= G_\star^{-1/2} G_\star G_\star^{-1/2} = \mathbf{I}_p.
\end{aligned}$$

It follows that $G_\star^{-1/2} S(z, \theta_\star)$ is an isotropic random vector. Let $J := G_\star^{1/2} H_\star^{-1} G_\star^{1/2}$. It can be checked that

$$\|H_\star^{-1/2} S(z, \theta_\star)\|_2^2 = \|G_\star^{-1/2} S(z, \theta_\star)\|_J^2,$$

where we denote $\|A\|_B = \|B^{1/2} A B^{1/2}\|_2$ for positive semidefinite B. Now it follows from Theorem 38 that, with probability at least $1 - \delta/4$,

$$\|H_\star^{-1/2} S(z, \theta_\star)\|_2^2 \leq C\left[\mathbf{Tr}(J) + K_1^2\left(\|J\|_2 \sqrt{\log(e/\delta)} + \|J\|_\infty \log(1/\delta)\right)\right] \leq C K_1^2 \log(e/\delta) p_\star,$$

since $\|J\|_\infty \leq \|J\|_2 \leq \mathbf{Tr}(J) = p_\star$. Therefore, $\mathbb{P}(\mathcal{G}_1) \geq 1 - \delta/4$.

Event $\mathcal{G}_2$. By Proposition 10, we have $\mathbb{P}(\mathcal{G}_2) \geq 1 - \delta/4$.

Event $\mathcal{G}_3$. By Assumption 3, we know that

$$H_\star^{-1/2} H(z, \theta_\star) H_\star^{-1/2} - \mathbf{I}_p$$

satisfies a Bernstein condition with parameter K_2. It follows from Theorem 40 that $\mathbb{P}(\mathcal{G}_3) \geq 1 - \delta/4$.

Event $\mathcal{G}_4$. It follows directly from Proposition 11 that $\mathbb{P}(\mathcal{G}_4) \geq 1 - \delta/4$.

Now, by a union bound, we obtain $\mathbb{P}(\mathcal{G}_1\mathcal{G}_2\mathcal{G}_3\mathcal{G}_4) \geq 1 - \delta$.

Step 6. Conclusion. Putting all the above results together, we have shown that, with probability at least $1 - \delta$,

$$\mathcal{E} \leq C \frac{Rr_n/\sqrt{\mu_\star} + t_n}{1 - Rr_n/\sqrt{\mu_\star} - t_n} \left[\sqrt{K_1^2 \log(e/\delta) p_\star} + (1 + t_1) r_n\right] + (1 + t_1) r_n.$$

□

D.3 Intermediate Results

The proof of Theorem 1 relies on two key results: 1) the estimator θ_n belongs to a neighborhood of $\theta_\star$ stated in Proposition 10, and 2) the inverse empirical Hessian $H_n(\theta_n)^{-1}$ is close to it population counterpart $H_\star^{-1}$ stated in Proposition 11. Before we prove them, we give several useful lemmas.

Lemma 6. *Under Assumption 1, the empirical risk F_n is pseudo self-concordant with parameter R.*

Proof. By Assumption 1, the loss $\ell(Z_i, \cdot)$ is pseudo self-concordant with parameter R for every $i \in \{1, \dots, n\}$. Since $F_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(Z_i, \theta)$, we have

$$\begin{aligned}|D_\theta^3 F_n(\theta)[u,u,v]| &= \left|\frac{1}{n}\sum_{i=1}^n D_\theta^3 \ell(Z_i,\theta)[u,u,v]\right| \le \frac{1}{n}\sum_{i=1}^n |D_\theta^3 \ell(Z_i,\theta)[u,u,v]| \\ &\le \frac{1}{n}\sum_{i=1}^n R\|v\|_2 u^\top \nabla_\theta^2 \ell(Z_i,\theta) u = R\|v\|_2 u^\top \nabla_\theta^2 F_n(\theta) u.\end{aligned}$$

This completes the proof. □

The next lemma provides a sufficient condition for the estimator θ_n to be close to $\theta_\star$.

Lemma 7. *Under Assumption 1, whenever*

$$\|S_n(\theta_\star)\|_{H_n^{-1}(\theta_\star)} \le \sqrt{\lambda_{\min}(H_n(\theta_\star))}/(2R),$$

the estimator θ_n uniquely exists and satisfies

$$\|\theta_n - \theta_\star\|_{H_n(\theta_\star)} \le 4\|S_n(\theta_\star)\|_{H_n^{-1}(\theta_\star)}.$$

Proof. By Lemma 6, we have F_n is pseudo self-concordant with parameter R. Since θ_n is the empirical risk minimizer, the claim follows from Proposition 34 with $f = F_n$ and $x = \theta_\star$. □

Lemma 8. *Under Assumption 2, it holds that, with probability at least $1-\delta$,*

$$\|S_n(\theta_\star)\|_{H_\star^{-1}}^2 \le \frac{1}{n} C K_1^2 \log(e/\delta) p_\star.$$

Proof. Define $W := \sqrt{n} G_\star^{-1/2} S_n(\theta_\star)$. It can be verified that $\mathbb{E}[W] = \sqrt{n} G_\star^{-1/2} S(\theta_\star) = 0$ and

$$\begin{aligned}\mathbb{E}[WW^\top] &= \frac{1}{n} G_\star^{-1/2} \mathbb{E}\left[\left(\sum_{i=1}^n S(Z_i,\theta_\star)\right)\left(\sum_{i=1}^n S(Z_i,\theta_\star)\right)^\top\right]^2 G_\star^{-1/2} \\ &= G_\star^{-1/2} \frac{1}{n}\sum_{i=1}^n \mathbb{E}[S(Z_i,\theta_\star) S(Z_i,\theta_\star)^\top] G_\star^{-1/2} = \mathbf{I}_p.\end{aligned}$$

Moreover, by Lemma 37 and Assumption 2, we get that W is sub-Gaussian with $\|W\|_{\psi_2} \le C K_1$. Define $J := G_\star^{1/2} H_\star^{-1} G_\star^{1/2}/n$. It is clear that $\|S_n(\theta_\star)\|_{H_\star^{-1}}^2 = \|W\|_J^2$. By Theorem 38, we have, with probability at least $1-\delta$,

$$\|S_n(\theta_\star)\|_{H_\star^{-1}}^2 \le C K_1^2 \log(e/\delta) p_\star.$$

Here we have used $\|J\|_\infty \le \|J\|_2 \le \mathbf{Tr}(J) = p_\star$, $\log(1/\delta) \le \log(e/\delta)$, and $\sqrt{\log(e/\delta)} \le \log(e/\delta)$. □

Lemma 9. *Under Assumption 3, it holds that, with probability at least $1-\delta$,*

$$\frac{1}{2} H_\star \preceq H_n(\theta_\star) \preceq \frac{3}{2} H_\star,$$

whenever $n \ge 4(K_2 + 2\sigma_H^2)\log(2p/\delta)$.

Proof. By Assumption 3 and Theorem 40, it holds that, for any $t > 0$,

$$\mathbb{P}\left(\|H_\star^{-1/2} H_n(\theta_\star) H_\star^{-1/2} - \mathbf{I}_p\|_2 \geq t\right) \leq 2p \exp\left\{-\frac{nt^2}{2(\sigma_H^2 + K_2 t)}\right\}.$$

The claim then follows by setting $t = 1/2$. □

Now we are ready to prove the localization result.

Proposition 10. *Under Assumptions 1,2, and 3, we have, with probability at least $1 - \delta$, the estimator θ_n uniquely exists and satisfies*

$$\|\theta_n - \theta_\star\|_{H_\star}^2 \leq C K_1^2 \frac{p_\star}{n} \log\left(\frac{e}{\delta}\right) \tag{21}$$

whenever $n \geq \max\{4(K_2 + 2\sigma_H^2)\log(4p/\delta), \frac{CK_1^2 p_\star R^2}{\mu_\star} \log(e/\delta)\}$.

Proof. We define two events,

$$\mathcal{G}_1 := \left\{\|S_n(\theta_\star)\|_{H_\star^{-1}}^2 \leq \frac{1}{n} C K_1^2 \log(e/\delta) p_\star\right\}$$
$$\mathcal{G}_2 := \left\{\frac{1}{2} H_\star \preceq H_n(\theta_\star) \preceq \frac{3}{2} H_\star\right\}.$$

It suffices to prove the bound (21) on $\mathcal{G}_1\mathcal{G}_2$ and show $\mathbb{P}(\mathcal{G}_1\mathcal{G}_2) \geq 1 - \delta$.

Step 1. Prove the bound. By the events $\mathcal{G}_2$, we have $\sqrt{\lambda_{\min}(H_n(\theta_\star))}/(2R) \geq \sqrt{\mu_\star}/(2\sqrt{2}R)$. Note that $n \geq CK_1^2 \log(e/\delta) p_\star R^2/\mu_\star$. It follows from the event $\mathcal{G}_1$ that $\|S_n(\theta_\star)\|_{H_\star^{-1}} \leq \sqrt{\lambda_{\min}(H_n(\theta_\star))}/(2\sqrt{2}R)$. By the event $\mathcal{G}_2$, we have

$$\|S_n(\theta_\star)\|_{H_n^{-1}(\theta_\star)} \leq \sqrt{2}\|S_n(\theta_\star)\|_{H_\star^{-1}} \leq \frac{\sqrt{\lambda_{\min}(H_n(\theta_\star))}}{2R}.$$

According to Lemma 7, θ_n uniquely exists and satisfies

$$\|\theta_n - \theta_\star\|_{H_\star}^2 \leq 16\|S_n(\theta_\star)\|_{H_n^{-1}(\theta_\star)}^2.$$

Now the bound (21) follows from the event $\mathcal{G}_1$.

Step 2. Control the probability. According to Lemma 8 and Lemma 9, we know $\mathbb{P}(\mathcal{G}_1) \geq 1 - \delta/2$ and $\mathbb{P}(\mathcal{G}_2) \geq 1 - \delta/2$, respectively. Consequently,

$$\mathbb{P}(\mathcal{G}_1\mathcal{G}_2) = 1 - \mathbb{P}(\mathcal{G}_1^c \mathcal{G}_2^c) \geq 1 - \mathbb{P}(\mathcal{G}_1^c) - \mathbb{P}(\mathcal{G}_2^c) \geq 1 - \delta,$$

which completes the proof. □

We then bound the difference between the inverse empirical Hessian and the inverse population Hessian. Recall that we use the notation $\|A\|_B := \|B^{1/2} A B^{1/2}\|_2$ for B positive semidefinite.

Proposition 11. *Under Assumptions 1, 2, and 3, we have, with probability at least $1 - \delta$,*

$$\|H_n(\theta_n)^{-1} - H_\star^{-1}\|_{H_\star} \leq C_{K_1, K_2, \sigma_H}\left(\sqrt{\log\left(\frac{2p}{\delta}\right)} + R\sqrt{\frac{p_\star}{\mu_\star}\log\left(\frac{e}{\delta}\right)}\right)\frac{1}{\sqrt{n}}$$

whenever $n \geq C_{K_1, K_2, \sigma_H}\left(\log\left(\frac{2p}{\delta}\right) + \frac{p_\star}{\mu_\star} R^2 \log\left(\frac{e}{\delta}\right)\right)$.

Proof. Define

$$r_n := \sqrt{CK_1^2 \log(2e/\delta)\frac{p_\star}{n}}$$
$$t_n := \frac{2\sigma_H^2}{-K_2 + \sqrt{K_2^2 + 2\sigma_H^2 n/\log(4p/\delta)}}.$$

Model	Data	Loss Function	Self-Concordance Parameter R
Linear Regression	$x \in \mathbb{R}^p, y \in \mathbb{R}$	$\ell(\theta, z) := \frac{1}{2}(y - \theta^\top x)^2$	0
Binary Logistic Regression	$x \in \mathbb{R}^p, y \in \{0, 1\}$	$\ell(\theta, z) := -\log(\sigma(y \cdot \theta^\top x))$	$\|x\|_2$
Poisson Regression	$x \in \mathbb{R}^p, y \in \mathbb{N}$	$\ell(\theta, z) := -y(\theta^\top x) + \exp(\theta^\top x) + \log(y!)$	$\|x\|_2$
Multiclass Logistic Regression	$x \in \mathbb{R}^p, y \in \{1, ..., K\}$	$\ell(\theta, z) := \log(1 + \sum_{i=1}^K e^{w_i^T x}) - \sum_{i=2}^K y_i(w_i^\top X)$	$2\|x\|_2$

Table 3: Examples of M-estimation for various generalized linear models and the corresponding values of the pseudo self-concordance parameter R. Each regression estimates a set of parameters θ based on input values x and output values y.

Note that they both decays as $O(n^{-1/2})$. In the following of the proof, we assume that $n \geq \max\{4(K_2 + 3\sigma_H^2)\log(4p/\delta), CK_1^2\log(2e/\delta)p_\star R^2/\mu_\star\}$. According to Lemma 35, it suffices to bound $\|H_n(\theta_n) - H_\star\|_{H_\star^{-1}}$. By the triangle inequality, we have

$$\|H_n(\theta_n) - H_\star\|_{H_\star^{-1}} \leq \underbrace{\|H_n(\theta_n) - H_n(\theta_\star)\|_{H_\star^{-1}}}_{A} + \underbrace{\|H_n(\theta_\star) - H_\star\|_{H_\star^{-1}}}_{B}. \tag{22}$$

We will control these two terms separately. The strategy is similar to the proof of Proposition 10: we prove the bound on some events and control the probability of these events. Define

$$\mathcal{G}_1 := \left\{\|S_n(\theta_\star)\|_{H_\star^{-1}}^2 \leq \frac{1}{n}CK_1^2\log(2e/\delta)p_\star\right\}$$
$$\mathcal{G}_2 := \{(1 - t_n)H_\star \preceq H_n(\theta_\star) \preceq (1 + t_n)H_\star\}.$$

When $n \geq 4(K_2 + 2\sigma_H^2)\log(4p/\delta)$, we have $t_n \leq 1/3$. It then follows from the proof of Proposition 10 that

$$\|\theta_n - \theta_\star\|_{H_\star}^2 \leq \frac{1}{n}CK_1^2\log(2e/\delta)p_\star \tag{23}$$

on the event $\mathcal{G}_1\mathcal{G}_2$ and $\mathbb{P}(\mathcal{G}_1) \geq 1 - \delta/2$.

Step 1. Control A and B. By (23), it holds that $\|\theta_n - \theta_\star\|_{H_\star} \leq r_n$. By Lemma 6 and Lemma 33, we have

$$A = \|H_n(\theta_n) - H_n(\theta_\star)\|_{H_\star^{-1}} \leq Re^{R\|\theta_n - \theta_\star\|_2}\|H_n(\theta_\star)\|_{H_\star^{-1}}\|\theta_n - \theta_\star\|_2.$$

Since $\|\theta_n - \theta_\star\|_2 \leq \mu_\star^{-1/2}r_n$ and $n \geq CK_1^2\log(2e/\delta)p_\star R^2/\mu_\star$, we have $\|\theta_n - \theta_\star\|_2 \leq 1/R$. As a result,

$$A \leq Re\|H_n(\theta_\star)\|_{H_\star^{-1}}r_n/\sqrt{\mu_\star} \leq 3Rer_n/(2\sqrt{\mu_\star}),$$

where the last inequality follows from the event $\mathcal{G}_2$ and $t_n \leq 1/2$. As for B, it follows from the event $\mathcal{G}_2$ that $B \leq t_n$. Therefore, absorbing $3e/2$ into the constant C in r_n, we obtain

$$\|H_n(\theta_n) - H_\star\|_{H_\star^{-1}} \leq Rr_n/\sqrt{\mu_\star} + t_n.$$

And it follows from Lemma 35 that

$$\|H_n(\theta_n)^{-1} - H_\star^{-1}\|_{H_\star} \leq \frac{Rr_n/\sqrt{\mu_\star} + t_n}{1 - Rr_n/\sqrt{\mu_\star} - t_n}.$$

Step 2. Control the probability of $\mathcal{G}_1\mathcal{G}_2$. By the matrix Bernstein inequality Theorem 40, we have $\mathbb{P}(\mathcal{G}_2) \geq 1 - \delta/2$. This implies that $\mathbb{P}(\mathcal{G}_1\mathcal{G}_2) \geq 1 - \delta$ since $\mathbb{P}(\mathcal{G}_1) \geq 1 - \delta/2$. □

E Linearization Error Bound

We control in this section the linearization error in Theorem 2.

E.1 Setup

Recall that

$$\theta_n := \arg\min_{\theta\in\Theta} \left[F_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell(Z_i, \theta) \right]$$

and

$$\theta_{n,\varepsilon,z} := \arg\min_{\theta\in\Theta} \left[(1-\varepsilon)F_n(\theta) + \varepsilon\ell(z,\theta) \right].$$

Since z is a fixed data point, we make the following boundedness assumptions at z in addition to Assumptions 1 to 3.

Assumption 4 (Bounded Gradient at z)**.** *The normalized gradient at z is bounded in a neighborhood of $\theta_\star$, i.e., there exist $M_1 \geq 1, \rho \in (0, R^{-1}]$ such that $\|\nabla\ell(z,\theta)\|_{H_\star^{-1}} \leq M_1$ for all $\|\theta-\theta_\star\|_{H_\star} \leq \rho$.*

Assumption 5 (Bounded Hessian at z)**.** *The normalized Hessian at z is bounded in a neighborhood of $\theta_\star$, i.e., there exist $M_2 \geq 1, \rho \in (0, R^{-1}]$ such that $\|H(z,\theta)\|_{H_\star^{-1}} \leq M_2$ for all $\|\theta-\theta_\star\|_{H_\star} \leq \rho$.*

Remark. When the Hessian $H(z,\theta)$ is well-defined, we know $\nabla\ell(z,\cdot)$ is continuous and thus Assumption 4 is satisfied automatically.

E.2 Proof of the Linearization Error Bound

Theorem 2'. *Under Assumptions 1 to 5, it holds that, with probability at least $1-\delta$,*

$$\left\| \frac{\theta_{n,\varepsilon,z}-\theta_n}{\varepsilon} - I_n(z) \right\|_{H_n(\theta_n)} \leq \frac{\sqrt{2}M_1\left((1-\varepsilon)(e^{R\mathcal{C}_n}-1)+\varepsilon(2M_2+1)\right)}{1-(1-\varepsilon)(e^{R\mathcal{C}_n}-1)-\varepsilon(2M_2+1)},$$

where $\mathcal{C}_n := C\mu_\star^{-1/2}\left[K_1\sqrt{p_\star \log\frac{e}{\delta}/n} + \varepsilon M_1/(1-\varepsilon)\right]$, whenever $\varepsilon \leq \min\{\rho/(CM_1+\rho), C/M_2, \sqrt{\mu_\star}/(\sqrt{\mu_\star}+8RM_1)\}$ and

$$n \geq \max\left\{ 8(K_2+4\sigma_H^2)\log\frac{4p}{\delta}, \frac{CK_1^2 p_\star R^2}{\min\{\mu_\star, \rho^2R^2\}}\log\frac{e}{\delta} \right\}.$$

Proof. The proof is inspired by Giordano et al. (2019). By the optimality of $\theta_{n,\varepsilon,z}$, it holds that

$$(1-\varepsilon)\nabla F_n(\theta_{n,\varepsilon,z}) + \varepsilon\nabla\ell(z,\theta_{n,\varepsilon,z}) = 0.$$

Define $\bar{H}_n(\theta) := \int_0^1 H_n(\theta_n + t(\theta-\theta_n))\mathrm{d}t$ and $\bar{H}(z,\theta) := \int_0^1 H(z,\theta_n+t(\theta-\theta_n))\mathrm{d}t$, where $H(z,\theta) := \nabla^2\ell(z,\theta)$. It follows from the Integral form of the Remainder of Taylor's theorem (defined in Appendix I) that

$$(1-\varepsilon)\bar{H}_n(\theta_{n,\varepsilon,z})(\theta_{n,\varepsilon,z}-\theta_n) + \varepsilon\nabla\ell(z,\theta_n) + \varepsilon\bar{H}(z,\theta_{n,\varepsilon,z})(\theta_{n,\varepsilon,z}-\theta_n) = 0,$$

where we have used $\nabla F_n(\theta_n) = 0$. This implies that

$$\theta_{n,\varepsilon,z} - \theta_n = -\left[(1-\varepsilon)\bar{H}_n(\theta_{n,\varepsilon,z}) + \varepsilon\bar{H}(z,\theta_{n,\varepsilon,z})\right]^{-1}\varepsilon\nabla\ell(z,\theta_n),$$

and thus

$$\begin{aligned}
&\left\| \frac{\theta_{n,\varepsilon,z}-\theta_n}{\varepsilon} - I_n(z) \right\|_{H_n(\theta_n)} \\
&= \left\| \left\{ \left[(1-\varepsilon)\bar{H}_n(\theta_{n,\varepsilon,z}) + \varepsilon\bar{H}(z,\theta_{n,\varepsilon,z})\right]^{-1} - H_n(\theta_n)^{-1} \right\} \nabla\ell(z,\theta_n) \right\|_{H_n(\theta_n)} \\
&= \left\| \left\{ H_n(\theta_n)^{1/2}\left[(1-\varepsilon)\bar{H}_n(\theta_{n,\varepsilon,z}) + \varepsilon\bar{H}(z,\theta_{n,\varepsilon,z})\right]^{-1} H_n(\theta_n)^{1/2} - \mathbf{I}_p \right\} H_n(\theta_n)^{-1/2}\nabla\ell(z,\theta_n) \right\|_2 \\
&\leq \left\| H_n(\theta_n)^{1/2}\left[(1-\varepsilon)\bar{H}_n(\theta_{n,\varepsilon,z}) + \varepsilon\bar{H}(z,\theta_{n,\varepsilon,z})\right]^{-1} H_n(\theta_n)^{1/2} - \mathbf{I}_p \right\|_2 \left\| H_n(\theta_n)^{-1/2}\nabla\ell(z,\theta_n) \right\|_2 \\
&= \underbrace{\left\| \left[(1-\varepsilon)\bar{H}_n(\theta_{n,\varepsilon,z}) + \varepsilon\bar{H}(z,\theta_{n,\varepsilon,z})\right]^{-1} - H_n(\theta_n)^{-1} \right\|_{H_n(\theta_n)}}_{A_1} \underbrace{\left\| H_n(\theta_n)^{-1}\nabla\ell(z,\theta_n) \right\|_{H_n(\theta_n)}}_{A_2}.
\end{aligned}$$

Recall r_n and t_n from (16). To proceed, we define the following events

$$\mathcal{G}_1 := \left\{ \|S_n(\theta_\star)\|^2_{H_\star^{-1}} \le \frac{1}{n} C K_1^2 \log(e/\delta) p_\star \right\}$$
$$\mathcal{G}_2 := \left\{ \frac{1}{2} H_\star \preceq H_n(\theta_\star) \preceq \frac{3}{2} H_\star \right\}$$
$$\mathcal{G}_3 := \left\{ \|H_\star^{1/2} H_n(\theta_n)^{-1} H_\star^{1/2} - \mathbf{I}_p\|_2 \le \frac{R r_n/\sqrt{\mu_\star} + t_n}{1 - R r_n/\sqrt{\mu_\star} - t_n} \right\}.$$

Moreover, we assume $\varepsilon \le \min\{\rho/(CM_1+\rho), C/M_2, \sqrt{\mu_\star}/(\sqrt{\mu_\star}+8RM_1)\}$ and

$$n \ge \max\left\{ 8(K_2 + 4\sigma_H^2) \log\frac{4p}{\delta}, \frac{C K_1^2 p_\star R^2}{\min\{\mu_\star, \rho^2 R^2\}} \log\frac{e}{\delta} \right\}.$$

throughout the proof. Note that $R r_n/\sqrt{\mu_\star} + t_n \le 1/2$ under this requirement of n. Recall from the proof of Proposition 10, Proposition 11, and Proposition 12 that $\mathbb{P}(\mathcal{G}_1\mathcal{G}_2\mathcal{G}_3) \ge 1-\delta$ and

$$\begin{aligned} \|\theta_n - \theta_\star\|^2_{H_\star} &\le C K_1^2 \frac{p_\star}{n} \log\frac{e}{\delta} \\ \|\theta_{n,\varepsilon,z} - \theta_\star\|^2_{H_\star} &\le C K_1^2 \frac{p_\star}{n} \log\frac{e}{\delta} + \frac{128\varepsilon^2}{(1-\varepsilon)^2} M_1^2. \end{aligned} \tag{24}$$

Therefore, it suffices to bound A_1 and A_2 on the event $\mathcal{G}_1\mathcal{G}_2\mathcal{G}_3$.

Step 1. Bound A_1. We will use Lemma 35 to bound A_1. We define

$$\begin{aligned} B &:= \left\| (1-\varepsilon)\bar{H}_n(\theta_{n,\varepsilon,z}) + \varepsilon \bar{H}(z, \theta_{n,\varepsilon,z}) - H_n(\theta_n) \right\|_{H_n(\theta_n)^{-1}} \\ &\le (1-\varepsilon) \underbrace{\left\| \bar{H}_n(\theta_{n,\varepsilon,z}) - H_n(\theta_n) \right\|_{H_n(\theta_n)^{-1}}}_{B_1} + \varepsilon \underbrace{\left\| \bar{H}(z, \theta_{n,\varepsilon,z}) - H_n(\theta_n) \right\|_{H_n(\theta_n)^{-1}}}_{B_2}. \end{aligned}$$

We first bound B_1. By Jensen's inequality, we get

$$\begin{aligned} B_1 &\le \int_0^1 \|H_n(\theta_n + t(\theta_{n,\varepsilon,z} - \theta_n)) - H_n(\theta_n)\|_{H_n(\theta_n)^{-1}} \, \mathrm{d}t \\ &= \int_0^1 \|H_n(\theta_n + t(\theta_{n,\varepsilon,z} - \theta_n))\|_{H_n(\theta_n)^{-1}} \, \mathrm{d}t + 1. \end{aligned}$$

By Lemma 6 and Proposition 32, it holds that

$$e^{-Rt\|\theta_{n,\varepsilon,z} - \theta_n\|_2} H_n(\theta_n) \preceq H_n(\theta_n + t(\theta_{n,\varepsilon,z} - \theta_n)) \preceq e^{Rt\|\theta_{n,\varepsilon,z} - \theta_n\|_2} H_n(\theta_n).$$

It then follows from Proposition 12 and $t \in [0,1]$ that

$$e^{-R\mathcal{C}_n} H_n(\theta_n) \preceq H_n(\theta_n + t(\theta_{n,\varepsilon,z} - \theta_n)) \preceq e^{R\mathcal{C}_n} H_n(\theta_n),$$

where $\mathcal{C}_n := C\mu_\star^{-1/2}\left[K_1\sqrt{p_\star \log\frac{e}{\delta}/n} + \varepsilon M_1/(1-\varepsilon)\right]$. Since $1 - e^{-x} \le e^x - 1$ for all $x \ge 0$, we get

$$B_1 \le e^{R\mathcal{C}_n} - 1.$$

We then bound B_2. We start the same as before using Jensen's inequality, we get

$$B_2 \le \int_0^1 \|H(z, \theta_n + t(\theta_{n,\varepsilon,z} - \theta_n)) - H_n(\theta_n)\|_{H_n(\theta_n)^{-1}} \, \mathrm{d}t.$$

Using the triangle inequality we can write

$$\begin{aligned} B_2 &\le \int_0^1 \left[\|H(z, \theta_n + t(\theta_{n,\varepsilon,z} - \theta_n))\|_{H_n(\theta_n)^{-1}} + \|H_n(\theta_n)\|_{H_n(\theta_n)^{-1}} \right] \mathrm{d}t \\ &= \int_0^1 \|H(z, \theta_n + t(\theta_{n,\varepsilon,z} - \theta_n))\|_{H_n(\theta_n)^{-1}} \, \mathrm{d}t + 1. \end{aligned}$$

Then it follows from the event $\mathcal{G}_3$ and the requirement of n that

$$B_2 \leq \frac{1}{1 - Rr_n/\sqrt{\mu_\star} - t_n} \int_0^1 \|H(z, \theta_n + t(\theta_{n,\varepsilon,z} - \theta_n))\|_{H_\star^{-1}} \, \mathrm{d}t + 1$$
$$\leq 2 \int_0^1 \|H(z, \theta_n + t(\theta_{n,\varepsilon,z} - \theta_n))\|_{H_\star^{-1}} \, \mathrm{d}t + 1$$

Since $\|\theta_n + t(\theta_{n,\varepsilon,z} - \theta_n) - \theta_\star\|_{H_\star} \leq \max\{\|\theta_n - \theta_\star\|_{H_\star}, \|\theta_{n,\varepsilon,z} - \theta_\star\|_{H_\star}\}$ for $t \in [0,1]$, it follows from Proposition 12 that

$$\|\theta_n + t(\theta_{n,\varepsilon,z} - \theta_n) - \theta_\star\|_{H_\star} \leq C \left[K_1 \sqrt{\frac{p_\star}{n} \log \frac{e}{\delta}} + \frac{\varepsilon}{1-\varepsilon} M_1 \right] < \rho$$

by the requirement of n and ε. As a result, we have

$$\|H(z, \theta_n + t(\theta_{n,\varepsilon,z} - \theta_n))\|_{H_\star^{-1}} \leq M_2$$

by Assumption 5. Combining the above results we obtain

$$B_2 \leq 2M_2 + 1,$$

which implies

$$B \leq (1-\varepsilon)(e^{R\mathcal{C}_n} - 1) + \varepsilon(2M_2 + 1) \leq \lambda_{\min}(\mathbf{I}_p) = 1,$$

where the last inequality holds by the requirements of n and ε.

Hence, applying Lemma 35 to $H_n(\theta_n)^{-1/2}[(1-\varepsilon)\bar{H}_n(\theta_{n,\varepsilon,z}) + \varepsilon\bar{H}(z, \theta_{n,\varepsilon,z})]H_n(\theta_n)^{-1/2}$ and $\mathbf{I}_p$ yields

$$A_1 \leq \frac{(1-\varepsilon)(e^{R\mathcal{C}_n} - 1) + \varepsilon(2M_2 + 1)}{1 - (1-\varepsilon)(e^{R\mathcal{C}_n} - 1) - \varepsilon(2M_2 + 1)}.$$

Step 2. Bound A_2. By the event $\mathcal{G}_3$ and the requirement of n, we have (similar to the bound of B_2)

$$A_2 = \|\nabla \ell(z, \theta_n)\|_{H_n(\theta_n)^{-1}} \leq \sqrt{2} \|\nabla \ell(z, \theta_n)\|_{H_\star^{-1}}.$$

By (24) and the requirement of n, it holds that $\|\theta_n - \theta_\star\|_{H_\star} < \rho$ and thus, by Assumption 4,

$$A_2 \leq \sqrt{2} M_1.$$

Step 3. Combine the bounds of A_1 and A_2. Combining the bounds for A_1 and A_2 we arrive at the final result,

$$\left\| \frac{\theta_{n,\varepsilon,z} - \theta_n}{\varepsilon} - I_n(z) \right\|_{H_n(\theta_n)} \leq \frac{\sqrt{2} M_1 \left((1-\varepsilon)(e^{R\mathcal{C}_n} - 1) + \varepsilon(2M_2 + 1) \right)}{1 - (1-\varepsilon)(e^{R\mathcal{C}_n} - 1) - \varepsilon(2M_2 + 1)}.$$

□

E.3 Intermediate Results

The proof of Theorem 2 relies on a key result: the perturbed estimator $\theta_{n,\varepsilon,z}$ is close to θ_n stated in Proposition 12.

Proposition 12. *Under Assumptions 1 to 5, it holds that*

$$\|\theta_{n,\varepsilon,z} - \theta_n\|_{H_\star}^2 \leq C K_1^2 \frac{p_\star}{n} \log \frac{e}{\delta} + \frac{128\varepsilon^2}{(1-\varepsilon)^2} M_1^2,$$

whenever $\varepsilon \leq \sqrt{\mu_\star}/(\sqrt{\mu_\star} + 8RM_1)$ *and*

$$n \geq \max \left\{ 4(K_2 + 2\sigma_H^2) \log \frac{4p}{\delta}, \frac{C K_1^2 p_\star R^2}{\mu_\star} \log \frac{e}{\delta} \right\}.$$

Proof. By the triangle inequality, we have

$$\|\theta_{n,\varepsilon,z} - \theta_n\|_{H_\star} \le \|\theta_{n,\varepsilon,z} - \theta_\star\|_{H_\star} + \|\theta_n - \theta_\star\|_{H_\star}.$$

It remains to control $\|\theta_{n,\varepsilon,z} - \theta_\star\|_{H_\star}$ and $\|\theta_n - \theta_\star\|_{H_\star}$. The second term is controlled by Proposition 10. We will control the first term with a similar argument.

We define two events

$$\mathcal{G}_1 := \left\{ \|S_n(\theta_\star)\|^2_{H_\star^{-1}} \le \frac{1}{n} C K_1^2 \log(e/\delta) p_\star \right\}$$
$$\mathcal{G}_2 := \left\{ \frac{1}{2} H_\star \preceq H_n(\theta_\star) \preceq \frac{3}{2} H_\star \right\},$$

and assume that $\varepsilon \le \sqrt{\mu_\star}/(\sqrt{\mu_\star} + 8RM_1)$ and

$$n \ge \max\left\{ 4(K_2 + 2\sigma_H^2) \log\frac{4p}{\delta}, \frac{CK_1^2 p_\star R^2}{\mu_\star} \log\frac{e}{\delta} \right\}.$$

It follows from Proposition 10 that $\mathbb{P}(\mathcal{G}_1\mathcal{G}_2) \ge 1 - \delta$ and

$$\|\theta_n - \theta_\star\|^2_{H_\star} \le CK_1^2 \frac{p_\star}{n} \log\frac{e}{\delta}.$$

We then control $\|\theta_{n,\varepsilon,z} - \theta_\star\|_{H_\star}$ on the event $\mathcal{G}_1\mathcal{G}_2$. Following the proof of Lemma 6, we know that $(1-\varepsilon)F_n(\cdot) + \varepsilon\ell(z,\cdot)$ is pseudo self-concordant with parameter R. Let

$$S_{n,\varepsilon,z}(\theta) := (1-\varepsilon)S_n(\theta) + \varepsilon S(z,\theta) \quad \text{and} \quad H_{n,\varepsilon,z}(\theta) := (1-\varepsilon)H_n(\theta) + \varepsilon H(z,\theta).$$

Since we assume $\ell(z,\theta)$ is convex then $H(z,\theta) \succeq 0$. Then, by the event $\mathcal{G}_2$, we have

$$H_{n,\varepsilon,z}(\theta_\star) \succeq \left(\frac{1-\varepsilon}{2}\right) H_\star.$$

As a result, it holds that

$$\begin{aligned}
\|S_{n,\varepsilon,z}(\theta_\star)\|_{H_{n,\varepsilon,z}(\theta_\star)^{-1}} &\le \sqrt{\frac{2}{1-\varepsilon}} \|S_{n,\varepsilon,z}(\theta_\star)\|_{H_\star^{-1}} \\
&\le \sqrt{\frac{2}{1-\varepsilon}} \left[(1-\varepsilon)\|S_n(\theta_\star)\|_{H_\star^{-1}} + \varepsilon \|S(z,\theta_\star)\|_{H_\star^{-1}} \right].
\end{aligned}$$

By Assumption 4, we obtain

$$\|S_{n,\varepsilon,z}(\theta_\star)\|_{H_{n,\varepsilon,z}(\theta_\star)^{-1}} \le \sqrt{\frac{2}{1-\varepsilon}} \left[(1-\varepsilon)\|S_n(\theta_\star)\|_{H_\star^{-1}} + \varepsilon M_1 \right]$$

Since $\sqrt{\lambda_{\min}(H_{n,\varepsilon,z}(\theta_\star))} \ge \sqrt{(1-\varepsilon)\mu_\star/2}$, it follows from the event $\mathcal{G}_1$ and the requirement of n that

$$\|S_{n,\varepsilon,z}(\theta_\star)\|_{H_{n,\varepsilon,z}(\theta_\star)^{-1}} \le \frac{\sqrt{\lambda_{\min}(H_{n,\varepsilon,z}(\theta_\star))}}{2R}.$$

According to Proposition 34, $\theta_{n,\varepsilon,z}$ uniquely exists and satisfies

$$\|\theta_{n,\varepsilon,z} - \theta_\star\|^2_{H_{n,\varepsilon,z}(\theta_\star)} \le 16 \|S_{n,\varepsilon,z}(\theta_\star)\|^2_{H_{n,\varepsilon,z}(\theta_\star)^{-1}} \le \frac{64}{1-\varepsilon} \left[(1-\varepsilon)^2 CK_1^2 \frac{p_\star}{n} \log\frac{e}{\delta} + \varepsilon^2 M_1^2 \right],$$

which implies

$$\|\theta_{n,\varepsilon,z} - \theta_\star\|^2_{H_\star} \le CK_1^2 \frac{p_\star}{n} \log\frac{e}{\delta} + \frac{128\varepsilon^2}{(1-\varepsilon)^2} M_1^2. \tag{25}$$

□

F Computational Error Bounds

We analyze the computation error of the algorithms discussed in Section 2 used to compute the empirical influence function. Throughout, we assume that the target precision satisfies $\varepsilon \leq \|I(z)\|_{H_\star}^2$. If not, taking $\hat{I}_n(z) = 0$ satisfies the desired precision and there is nothing to do.

Condition Numbers. Throughout, we assume that the loss function $\ell(\cdot, z)$ is L-smooth for each Z and that $H_n(\theta_n)$ is invertible. Let $\mu_n = \lambda_{\min}(H_n(\theta_n))$ denote the minimal eigenvalue. The computational bounds depend on the condition number

$$\kappa_n := \frac{L}{\mu_n}.$$

The corresponding population condition number is

$$\kappa_\star = \frac{L}{\mu_\star},$$

where $\mu_\star = \lambda_{\min}(H_\star)$. They are related as follows.

K-Condition Numbers. Another useful notion to obtain the convergence rate of the conjugate gradient method is the K-condition number defined as

$$K_n := \frac{[\mathbf{Tr}\, H_n(\theta_n)/p]^p}{\det H_n(\theta_n)}.$$

Its population counterpart is defined as

$$K_\star := \frac{[\mathbf{Tr}\, H_\star/p]^p}{\det H_\star}.$$

Proposition 13. *Consider the setting of Theorem 1, and let $\mathcal{G}$ denote the event under which its conclusions hold. Under this event $\mathcal{G}$, we have,*

(a) $\kappa_n \leq 4\kappa_\star$, *and*

(b) if $\|I_n(z) - I(z)\|_{H_\star}^2 = \varepsilon$, *then* $\|I_n(z)\|_{H_n(\theta_n)}^2 \leq 6\|I(z)\|_{H_\star}^2 + 6\varepsilon$.

Proof. We have under $\mathcal{G}$ that $(1/4)H_\star \preceq H_n(\theta_n) \preceq 3H_\star$. This implies that $\mu_n \geq \mu_\star/4$, $\mathbf{Tr}\, H_n(\theta_n) \leq 3\,\mathbf{Tr}\, H_\star$, and $\det H_n(\theta_n) \geq \det H_\star/4^p$. For the second part, we get from the triangle inequality,

$$\|I_n(z)\|_{H_n(\theta_n)}^2 \leq 3\|I_n(z)\|_{H_\star}^2 \leq 6\|I(z)\|_{H_\star}^2 + 6\|I_n(z) - I(z)\|_{H_\star}^2.$$

□

F.1 Total Error

We combine the computational error with the statistical error to get the total error bound. This is a restatement of Proposition 3 of the main paper.

Proposition 14. *Consider the setting of Theorem 1, and let $\mathcal{G}$ denote the event under which its conclusions hold. Let $\hat{I}_n(\theta)$ be an estimate of $I_n(\theta)$ that satisfies $\mathbb{E}\left[\|\hat{I}_n(z) - I_n(z)\|_{H_n(\theta_n)}^2 \middle| Z_{1:n}\right] \leq \varepsilon$. Then, we have,*

$$\mathbb{E}\left[\|\hat{I}_n(z) - I(z)\|_{H_\star}^2 \,\middle|\, \mathcal{G}\right] \leq 8\varepsilon + C_{K_1,K_2,\sigma_H} \frac{R^2 p_\star^2}{\mu_\star n} \operatorname{poly}\log \frac{p}{\delta},$$

whenever $n \geq C_{K_1,K_2,\sigma_H}\left(\frac{p_\star}{\mu_\star} R^2 \log\left(\frac{e}{\delta}\right) + \log\left(\frac{2p}{\delta}\right)\right)$.

Proof. Following the proof of Theorem 1, we have under $\mathcal{G}$ that

$$\frac{1}{4} H_\star \preceq H_n(\theta_n) \preceq 3H_\star.$$

Therefore, $\|u\|_{H_\star}^2 \leq 4\|u\|_{H_n(\theta_n)}^2$. Combining this with the triangle inequality completes the proof. □

F.2 The Conjugate Gradient Method

We start by recalling the convergence analysis of the conjugate gradient method, providing a full proof for completeness.

Proposition 15. *Consider the sequence (u_t) produced by the conjugate gradient method for solving $u_\star = H_n(\theta_n)^{-1}S(z,\theta_n)$. It holds that*

$$\|u_t - u_\star\|^2_{H_n(\theta_n)} \leq 4\left(\frac{\sqrt{\kappa_n}-1}{\sqrt{\kappa_n}+1}\right)^{2t}\|u_0 - u_\star\|^2_{H_n(\theta_n)}.$$

In other words, we get $\|u_t - u_\star\|^2_{H_n(\theta_n)} \leq \varepsilon$ after t_{cg} iterations, where

$$t_{\mathrm{cg}} \leq \frac{\sqrt{\kappa_n}}{2}\log\left(\frac{4\|u_0 - u_\star\|^2_{H_n(\theta_n)}}{\varepsilon}\right).$$

Proof. We follow the proof template of Chen (2005, Chapter 3.4). Throughout, we use the shorthand $A = H_n(\theta_n)$. By construction, we have $u_k \in \mathrm{Span}\{p_0,\dots,p_{k-1}\}$. It then follows from $p_k = r_k + \beta_{k-1}p_{k-1}$ that $\mathrm{Span}\{p_0,\dots,p_{k-1}\} = \mathrm{Span}\{r_0,\dots,r_{k-1}\}$. Moreover, since $r_k = b - Au_k = r_{k-1} - \alpha_{k-1}Ap_{k-1}$, we get

$$\mathrm{Span}\{r_0,\dots,r_{k-1}\} = \mathrm{Span}\{r_0, Ar_0,\dots,A^{k-1}r_0\} =: \mathcal{K}_k(A,r_0),$$

where $\mathcal{K}_k(A,r_0)$ is known as the Krylov subspace of order k for the matrix A and the generating vector r_0. Since $u_0 = 0$, it holds that $r_0 = b = Au_\star$ and thus

$$\mathcal{K}_k(A,r_0) = \mathrm{Span}\{b, Ab,\dots,A^{k-1}b\}.$$

We will write $\mathcal{K}_k$ for short.

For an arbitrary $x \in \mathcal{K}_k$, there exists $\{\alpha_i\}_{i=0}^{k-1}$ such that $x = \sum_{i=0}^{k-1}\alpha_i A^i b$. Let $f(t) := \sum_{i=0}^{k-1}\alpha_i t^i$. It follows that

$$\|u - u_\star\|^2_A = (f(A)Au_\star - u_\star)^\top A(f(A)Au_\star - u_\star) = u_\star^\top g(A)Ag(A)u_\star,$$

where $g(t) := 1 - f(t)t$ and $A = A^\top$ has been used. Since A is positive semi-definite, it admits an eigenvalue decomposition $A = Q\Lambda Q^\top$. It then follows from $A^k = Q\Lambda^k Q$ that

$$u_\star^\top g(A)Ag(A)u_\star = u_\star^\top Qg(\Lambda)\Lambda g(\Lambda)Q^\top u_\star.$$

Denote $y := Q^\top u_\star$ and $\Lambda = \mathrm{Diag}\{\lambda_j\}$. Then we get

$$u_\star^\top Qg(\Lambda)\Lambda g(\Lambda)Q^\top u_\star = \sum_{j=1}^p \lambda_j g(\lambda_j)^2 y_j^2.$$

Note that

$$\|u - u_\star\|^2_A = u^\top Au - 2u^\top Au_\star + u_\star^\top Au_\star = u^\top Au - 2u^\top b + u_\star^\top Au_\star$$

According to Chen (2005, Equation 3.31),

$$\|u_k - u_\star\|^2_A = \min_{x\in\mathrm{Span}\{p_0,\dots,p_{k-1}\}}\|x - u_\star\|^2_A = \min_{g\in\mathcal{G}_k}\sum_{j=1}^p \lambda_j g(\lambda_j)^2 y_j^2,$$

where $\mathcal{G}_k$ is the collection of polynomials of degree k that take value 1 at 0. Define

$$C(\Lambda) := \min_{g\in\mathcal{G}_k}\max_{j\in[p]}|g(\lambda_j)|.$$

Using properties of Chebyshev polynomials, we obtain (e.g., Chen, 2005, Equation 3.46)

$$C(\Lambda) \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k,$$

where $\kappa := \lambda_{\max}(A)/\lambda_{\min}(A)$. As a result,

$$\|u_k - u_\star\|_A^2 \leq \min_{g \in \mathcal{G}_k} \sum_{j=1}^p \lambda_j \max_{j' \in [p]} g(\lambda_{j'})^2 y_j^2 = C(\Lambda)^2 \sum_{j=1}^p \lambda_j y_j^2 = C(\Lambda)^2 y^\top \Lambda y = C(\Lambda)^2 u_\star^\top A u_\star$$
$$\leq 4 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^{2k} \|u_0 - u_\star\|_A^2.$$

We use the bound $\kappa \leq \kappa_n$ to complete the proof. □

Corollary 16 (Total Computational Cost; Conjugate Gradient Method)**.** *Fix $\varepsilon > 0$. Consider the setting of Theorem 1, and let $\mathcal{G}$ denote the high probability event under which its conclusions hold. Choose a sample size n such that*

$$n = C_{K_1, K_2, \sigma_H} \frac{R^2 p_\star^2}{\mu_\star \varepsilon} \operatorname{poly} \log \frac{p}{\delta}.$$

Then, under $\mathcal{G}$, the number N_{cg} of gradient and Hessian-vector oracle calls required to obtain a point $\hat{I}_n(z)$ using the conjugate gradient method initialized at $u_0 = 0$ such that $\|\hat{I}_n(z) - I(z)\|_{H_\star}^2 \leq \varepsilon$ is bounded by

$$N_{\mathrm{cg}} \leq C_{K_1, K_2, \sigma_H} \frac{R^2 p_\star^2 \kappa_\star^{3/2}}{L\varepsilon} \log \left(\frac{\|I(z)\|_{H_\star}^2}{\varepsilon} + 1 \right) \operatorname{poly} \log \frac{p}{\delta}.$$

Proof. We combine the total error bound of Proposition 14 with the computational bound of Proposition 15. Under $\mathcal{G}$, note that the choice of the sample size n implies that the statistical error is bounded from Theorem 1 by

$$\|I_n(z) - I(z)\|_{H_\star}^2 \leq \frac{\varepsilon}{2}.$$

Let t_{cg} be the number of conjugate gradient iterations t such that the $\|\hat{I}_n(z) - I_n(z)\|_{H_n(\theta_n)}^2 \leq \varepsilon/16$ as given in Proposition 15. By Proposition 14, the total error is then ε and the total number of gradient and Hessian-vector product oracle calls in $N = t_{\mathrm{cg}} n$, since each iteration requires a full pass over the data. To complete the proof, we invoke Proposition 13 to bound the initial gap $\|u_0 - u_\star\|_{H_n(\theta_n)}^2 = \|I_n(z)\|_{H_n(\theta_n)}$ and the condition number κ_n in terms of their respective population quantities. □

Remark 17. *When the spectrum of $H_\star$ decays as $O(i^{-\beta})$ for $\beta \in [0, 1)$, we can obtain a more refined analysis using the K-condition number. In the following, we assume that $p > 1$ and*

$$n \geq C_{K_1, K_2, \sigma_H} (p^2 + \varepsilon^{-1}) R^2 \frac{p_\star}{\mu_\star} \operatorname{poly} \log \frac{p}{\delta}.$$

Following the proof of Proposition 15, it holds that

$$\|u_t - u_\star\|_A^2 \leq C^2(\Lambda) \|u_0 - u_\star\|_A^2.$$

According to Axelsson and Kaporin (2000, Theorem 4.3), we have

$$C(\Lambda) \leq \left(\frac{3 \log K_n}{t} \right)^{t/2}.$$

Using the event $\mathcal{G}_4$ from the proof of Theorem 1, we know that $(1 - p^{-1}) H_\star \preceq H_n(\theta_n) \preceq (1 + p^{-1}) H_\star$. As a result, we have $K_n \leq (1 + p^{-1})^p (1 - p^{-1})^{-p} K_\star \leq C K_\star$. Moreover, it follows from Theorem 1 that the statistical error is controlled by $\varepsilon/2$.

We then control the computational error. Since $\lambda_i \sim i^{-\beta}$, we have $\mathbf{Tr}\, H_\star \sim p^{1-\beta}/(1-\beta)$ and $\det H_\star \sim (p!)^{-\beta}$. Consequently, it follows from Stirling's approximation that $K_\star \sim (2\pi p)^{\beta/2} e^{-\beta p} (1-\beta)^{-p}$. If $t > 6 \log (C K_\star) > 6 \log K_n$, then we only need $t > C \log \left(1 + \frac{\|I(z)\|_{H_\star}^2}{\varepsilon} \right)$ to achieve $\varepsilon/2$ computation error. Therefore, we have

$$t_{cg} \gtrsim 6 \log \left[C (2\pi p)^{\beta/2} e^{-\beta p} (1-\beta)^{-p} \right] + C \log \left(1 + \frac{\|I(z)\|_{H_\star}^2}{\varepsilon} \right),$$

and thus

$$N_{cg} \sim C_{K_1,K_2,\sigma_H}(p^2+\varepsilon^{-1})R^2\frac{p_\star}{\mu_\star}\left\{6\log\left[C(2\pi p)^{\beta/2}e^{-\beta p}(1-\beta)^{-p}\right]+C\log\left(1+\frac{\|I(z)\|^2_{H_\star}}{\varepsilon}\right)\right\}\operatorname{poly}\log\frac{p}{\delta}.$$

F.3 Stochastic Gradient Descent

We consider using SGD to solve the linear system $H_n(\theta_n)u+\nabla\ell(z,\theta_n)=0$. We do so by minimizing the quadratic g_n from (9):

$$g_n(u)=\frac{1}{2}\langle u,H_n(\theta_n)u\rangle+\langle\nabla\ell(z,\theta_n),u\rangle\,.$$

We run SGD by sampling an index i_t uniformly at random to update

$$u_{t+1}=u_t-\gamma\big(H(Z_{i_t},\theta_n)u_t+\ell(z,\theta_n)\big)\,.$$

The bounds depend on the following quantities:

(a) Let $\mu_n=\lambda_{\min}(H_n(\theta_n))$ be the minimal eigenvalue of $H_n(\theta_n)$.
(b) Define the matrix $W_n=\big(H_n(\theta_n)^{-1/2}H(Z_i,\theta_n)H_n(\theta_n)^{-1/2}-\mathbf{I}_p\big)$ and

$$\Sigma_n=\frac{1}{n}\sum_{i=1}^n W_nH_n(\theta_n)^{1/2}I_n(z)I_n(z)^\top H_n(\theta_n)^{1/2}W_n\,.$$

(c) Define the noise term

$$\sigma_n^2:=\mathbf{Tr}\,\Sigma_n+p\|\Sigma_n\|_2\,.$$

We have the following convergence bound for SGD (Jain et al., 2017b,a); cf. Appendix I.5 for details.

Lemma 18. *The sequence $(\bar u_t)$ produced by tail-averaged SGD on the function $g_n(u)$ from (9) with a learning rate of $\gamma=(2L)^{-1}$ satisfies*

$$\mathbb{E}\|\bar u_t-u_\star\|^2_{H_n(\theta_n)}\le C\left(\kappa_n\|u_0-u_\star\|^2_{H_n(\theta_n)}\exp\left(-\frac{t}{4\kappa_n}\right)+\frac{\sigma_n^2}{t}\right)\,.$$

Therefore, it returns a point $\bar u_t$ satisfying $\mathbb{E}\|\bar u_t-u_\star\|^2_{H_n(\theta_n))}\le\varepsilon$ after $t\ge t_{\mathrm{sgd}}$ steps where

$$t_{\mathrm{sgd}}\le C\left(\frac{\sigma_n^2}{\varepsilon}+\kappa_n\log\left(\frac{\kappa_n\|u_0-u_\star\|^2_{H_n(\theta_n)}}{\varepsilon}\right)\right)\,,$$

where $\kappa_n=L/\mu_n$ is the condition number.

Total Error Bound. We give a total error bound under a stronger assumption on the normalized Hessian. We strengthen the matrix Bernstein condition on the normalized Hessian into a spectral norm bound in a neighborhood around $\theta_\star$ as formalized below.

Assumption 3' (Bounded Hessian)**.** *The normalized Hessian is bounded in a neighborhood of $\theta_\star$, i.e., there exist $M_2>1$ and $\rho>0$ such that $\|H(z,\theta)\|_{H_\star^{-1}}\le M_2$ for all $z\in\mathcal{Z}$ and $\|\theta-\theta_\star\|_{H_\star}\le\rho$.*

This gives the following total error bound.

Proposition 19 (Total Error bound for SGD)**.** *Fix $\varepsilon>0$. Consider the setting of Theorem 1 and let $\mathcal{G}$ denote the event under which its conclusions hold. Suppose also that Assumption 3' is true. With probability at least $1-\delta$, the total error of $\hat I_n(z)$ obtained from t iterations of tail-averaged SGD is bounded as*

$$\mathbb{E}\left[\|\hat I_n(z)-I(z)\|^2_{H_\star}\,\middle|\,\mathcal{G}\right]\le C_{K_1,M_2,\sigma_H}\;(\mathcal{A}_1+\mathcal{A}_2+\mathcal{A}_3)\operatorname{poly}\log\frac{p}{\delta}\,,$$

where

$$\mathcal{A}_1 = \frac{R^2 p_\star^2}{n\mu_\star}\left(1 + \kappa_\star \exp\left(-\frac{t}{16\kappa_\star}\right)\right)$$
$$\mathcal{A}_2 = \kappa_\star \|I(z)\|_{H_\star}^2 \exp\left(-\frac{t}{16\kappa_\star}\right)$$
$$\mathcal{A}_3 = \frac{p_\star p^2}{nt} + \frac{R^2 p_\star p^2}{\mu_\star nt} + \frac{p_\star}{t}\|I(z)\|_{H_\star}^2$$

whenever

$$n \geq C_{K_1,M_2,\sigma_H}\, p_\star \left(\frac{R^2}{\mu_\star} + \frac{1}{\rho}\right)\log\frac{p}{\delta}\,.$$

Before proving Proposition 19, we state the final total error bound in terms of the number of calls to a Hessian-vector product oracle. To this end, define the coefficient $\sigma_\star^2$ as

$$\sigma_\star^2 := p_\star^2\left(\frac{R^2}{\mu_\star} + 1\right) + p^2\|I(z)\|_{H_\star}^2\,. \tag{26}$$

Corollary 20 (Total Oracle Complexity for SGD)**.** *Consider the setting of Proposition 19. If we choose*

$$n \geq \max\left\{1, \frac{R^2}{\mu_\star}\right\}\frac{p_\star^2}{\varepsilon}\operatorname{poly}\log\frac{p}{\delta} \quad \textit{and} \quad t \geq \left(\frac{p^2\|I(z)\|_{H_\star}^2}{\varepsilon} + \kappa_\star\log\left(\frac{\kappa_\star\|I(z)\|_{H_\star}^2}{\varepsilon}\right)\right)\operatorname{poly}\log\frac{p}{\delta}\,,$$

we have $\mathbb{E}\left[\|\hat{I}_n(z) - I(z)\|_{H_\star}^2 \,\middle|\, \mathcal{G}\right] \leq \varepsilon$*. Then, the minimal total number of calls to a Hessian-vector product oracle is*

$$N_{sgd} \leq \left(\frac{\sigma_\star^2}{\varepsilon} + \kappa_\star\log\left(\frac{\kappa_\star\|I(z)\|_{H_\star}^2}{\varepsilon}\right)\right)\operatorname{poly}\log\frac{p}{\delta}\,.$$

Proof. We use the shorthand $\Delta_\star := \|I(z)\|_{H_\star}^2$. We have that the total error is bounded as $\mathbb{E}\left[\|\hat{I}_n(z) - I(z)\|_{H_\star}^2 \,\middle|\, \mathcal{G}\right] \leq 6\varepsilon$ if each of the terms of Proposition 19 is bounded by ε. These conditions are (ignoring constants and the $\operatorname{poly}\log(p/\delta)$ term):

(a) $R^2 p_\star^2/(n\mu_\star) \leq \varepsilon$ holds, or the stronger condition $n \geq \max\{1, R^2/\mu_\star\}p_\star^2/\varepsilon$ holds.

(b) $R^2 p_\star^2\kappa_\star/(n\mu)\exp(-t/(16\kappa_\star)) \leq \varepsilon$ holds.

(c) $\Delta_\star\kappa_\star\exp(-t/(16\kappa_\star)) \leq \varepsilon$ or $t \geq 16\kappa_\star\log(\Delta_\star\kappa_\star/\varepsilon)$ holds.

(d) $p^2 p_\star/(nt) \leq \varepsilon$ or that $nt \geq p^2 p_\star/\varepsilon$.

(e) $R^2 p_\star p^2/(\mu_\star nt) \leq \varepsilon$ or that $nt \geq \frac{R^2 p_\star p^2}{\mu_\star\varepsilon}$.

(f) $p^2\Delta_\star/t \leq \varepsilon$ or that $t \geq p^2\Delta_\star/\varepsilon$.

Under the assumption that $\varepsilon < \Delta_\star$ (or else there is nothing to estimate), the conditions (a) and (f) together imply that the conditions (d) and (e) hold. Similarly, the conditions (a) and (c) together imply that condition (b) holds. Therefore, it suffices to have conditions (a), (c), and (f), which is the first claim. For the second one, note that the total number of Hessian-vector product calls is $\max\{n, t\} \leq n + t$. □

We now prove Proposition 19.

Proof of Proposition 19. We denote $\Delta_\star := \|I(z)\|_{H_\star}^2$ and $\Delta_n := \|I_n(z)\|_{H_n(\theta_n)}^2$ in this proof. Under the event $\mathcal{G}$, we have

$$\|I_n(z) - I(z)\|_{H_\star}^2 \leq \frac{R^2 p_\star^2}{n\mu_\star}\operatorname{poly}\log\frac{p}{\delta} =: E_n\,. \tag{27}$$

The computational bound Lemma 18 implies that

$$\mathbb{E}\left[\|\hat{I}_n(z) - I_n(z)\|^2_{H_n(\theta_n)} \middle| Z_{1:n}\right] \leq \kappa_n \Delta_n \exp\left(-\frac{t}{4\kappa_n}\right) + \frac{\sigma_n^2}{t}\,.$$

Invoking Proposition 13 and Lemma 21 (which requires n large enough as assumed), we can write

$$\mathbb{E}\left[\|\hat{I}_n(z) - I_n(z)\|^2_{H_\star} \middle| \mathcal{G}\right] \leq C\kappa_\star \Delta_\star \exp\left(-\frac{t}{16\kappa_\star}\right) + C_{K_1,M_2}\frac{p^2}{t}\left(\frac{p_\star}{n} + \frac{\Delta_\star R^2 p_\star}{\mu_\star n} + \Delta_\star\right)\log\frac{p}{\delta}\,. \tag{28}$$

We invoke the triangle inequality to complete the proof. □

The total error bounds rely on the following upper bound of the noise term σ_n^2 in terms of the population quantities. Recall that, for $A, J \in \mathbb{R}^{p\times p}$ with J being p.s.d., the weighted spectral norm $\|A\|_J := \left\|J^{1/2}AJ^{1/2}\right\|_2$.

Lemma 21. *Under Assumptions 1, 2, 3', we have, with probability at least* $1-\delta$*,*

$$\sigma_n^2 \leq C_{K_1,M_2}\cdot p^2\left[\frac{p_\star}{n}\log\frac{e}{\delta} + \frac{\|I(z)\|^2_{H_\star}}{n}\left[\frac{R^2 p_\star}{\mu_\star}\log\frac{e}{\delta} + \log\frac{2p}{\delta}\right] + \|I(z)\|^2_{H_\star}\right]$$

whenever $n \geq C_{K_1,M_2}\big(p_\star(R^2/\mu_\star + 1/\rho)\log(e/\delta) + \log(2p/\delta)\big)$.

Proof. Let $\mathcal{H}_n(Z) := H_n(\theta_n)^{-1/2}H(Z,\theta_n)H_n(\theta_n)^{-1/2}$. Then

$$\begin{aligned}
\mathbf{Tr}(\Sigma_n) &= \mathbf{Tr}\left\{\frac{1}{n}\sum_{i=1}^n [\mathcal{H}_n(Z_i) - \mathbf{I}_p]H_n(\theta_n)^{1/2}I_n(z)I_n(z)^\top H_n(\theta_n)^{1/2}[\mathcal{H}_n(Z_i) - \mathbf{I}_p]\right\} \\
&= \mathbf{Tr}\left\{\frac{1}{n}\sum_{i=1}^n [\mathcal{H}_n(Z_i) - \mathbf{I}_p]^2 H_n(\theta_n)^{1/2}I_n(z)I_n(z)^\top H_n(\theta_n)^{1/2}\right\} \\
&= I_n(z)^\top H_n(\theta_n)^{1/2}\left\{\frac{1}{n}\sum_{i=1}^n [\mathcal{H}_n(Z_i) - \mathbf{I}_p]^2\right\}H_n(\theta_n)^{1/2}I_n(z).
\end{aligned}$$

Note that $n^{-1}\sum_{i=1}^n \mathcal{H}_n(Z_i) = \mathbf{I}_p$. It follows that

$$\begin{aligned}
\mathbf{Tr}(\Sigma_n) &= I_n(z)^\top H_n(\theta_n)^{1/2}\left[\frac{1}{n}\sum_{i=1}^n \mathcal{H}_n(Z_i)^2\right]H_n(\theta_n)^{1/2}I_n(z) - \|I_n(z)\|^2_{H_n(\theta_n)} \\
&= I_n(z)^\top H_n(\theta_n)^{1/2}\left[\frac{1}{n}\sum_{i=1}^n H(Z_i,\theta_n)H_n(\theta_n)^{-1}H(Z_i,\theta_n)\right]H_n(\theta_n)^{1/2}I_n(z) - \|I_n(z)\|^2_{H_n(\theta_n)} \\
&= I_n(z)^\top H_n(\theta_n)^{1/2}H_n(\theta_n)^{-1/2}H_\star^{1/2}\mathcal{A}_n H_\star^{1/2}H_n(\theta_n)^{-1/2}H_n(\theta_n)^{1/2}I_n(z) - \|I_n(z)\|^2_{H_n(\theta_n)} \\
&\leq \left[\|\mathcal{A}_n\|_2\left\|H_n(\theta_n)^{-1/2}H_\star H_n(\theta_n)^{-1/2}\right\|_2 - 1\right]\|I_n(z)\|^2_{H_n(\theta_n)}\,,
\end{aligned} \tag{29}$$

where

$$\mathcal{A}_n := \frac{1}{n}\sum_{i=1}^n H_\star^{-1/2}H(Z_i,\theta_n)H_\star^{-1/2}H_\star^{1/2}H_n(\theta_n)^{-1}H_\star^{1/2}H_\star^{-1/2}H(Z_i,\theta_n)H_\star^{-1/2}.$$

The term $\left\|H_n(\theta_n)^{-1/2}H_\star H_n(\theta_n)^{-1/2}\right\|_2$ has been controlled in Proposition 11. Since

$$\|I_n(z)\|^2_{H_n(\theta_n)} \leq 2\|I_n(z) - I(z)\|^2_{H_n(\theta_n)} + 2\|I(z)\|^2_{H_n(\theta_n)}$$

it can be controlled using Theorem 1. It remains to control $\|\mathcal{A}_n\|_2$. Note that

$$\begin{aligned}\|\mathcal{A}_n\|_2 \leq \mathbf{Tr}(\mathcal{A}_n) &= \mathbf{Tr}\left\{\left[\frac{1}{n}\sum_{i=1}^n \left(H_\star^{-1/2}H(Z_i,\theta_n)H_\star^{-1/2}\right)^2\right] H_\star^{1/2}H_n(\theta_n)^{-1}H_\star^{1/2}\right\} \\ &\leq p\left\|\left[\frac{1}{n}\sum_{i=1}^n \left(H_\star^{-1/2}H(Z_i,\theta_n)H_\star^{-1/2}\right)^2\right] H_\star^{1/2}H_n(\theta_n)^{-1}H_\star^{1/2}\right\|_2 \\ &\leq p\left\|\frac{1}{n}\sum_{i=1}^n \left(H_\star^{-1/2}H(Z_i,\theta_n)H_\star^{-1/2}\right)^2\right\|_2 \left\|H_\star^{1/2}H_n(\theta_n)^{-1}H_\star^{1/2}\right\|_2. \end{aligned} \tag{30}$$

Again, the term $\left\|H_\star^{1/2}H_n(\theta_n)^{-1}H_\star^{1/2}\right\|_2$ can be controlled via Proposition 11. As for the term

$$\left\|\frac{1}{n}\sum_{i=1}^n \left(H_\star^{-1/2}H(Z_i,\theta_n)H_\star^{-1/2}\right)^2\right\|_2, \tag{31}$$

it can be bounded by 1) using the Lipschitzness of the Hessian to replace θ_n by $\theta_\star$, and 2) using the Matrix Bernstein inequality.

Let us prove the result rigorously. Define

$$r_n := \sqrt{CK_1^2\log(8e/\delta)\frac{p_\star}{n}} \quad \text{and} \quad t_n := \frac{CM_2}{-1+\sqrt{1+Cn/\log(16p/\delta)}}.$$

Define the following events

$$\begin{aligned}\mathcal{G}_1 &:= \left\{\|\theta_n-\theta_\star\|_{H_\star}^2 \leq r_n^2\right\} \\ \mathcal{G}_2 &:= \left\{\|H_\star^{1/2}H_n(\theta_n)^{-1}H_\star^{1/2}-\mathbf{I}_p\|_2 \leq \frac{Rr_n/\sqrt{\mu_\star}+t_n}{1-Rr_n/\sqrt{\mu_\star}-t_n}\right\} \\ \mathcal{G}_3 &:= \left\{\|I_n(z)-I(z)\|_{H_\star}^2 \leq \left[M_2r_n+(\|S(z,\theta_\star)\|_{H_\star^{-1}}+M_2r_n)\frac{Rr_n/\sqrt{\mu_\star}+t_n}{1-Rr_n/\sqrt{\mu_\star}-t_n}\right]^2\right\} \\ \mathcal{G}_4 &:= \left\{\left\|\frac{1}{n}\sum_{i=1}^n[H_\star^{-1/2}H(Z_i,\theta_\star)H_\star^{-1/2}]^2-\mathbb{E}\left\{[H_\star^{-1/2}H(Z,\theta_\star)H_\star^{-1/2}]^2\right\}\right\|_2 \leq \frac{1}{2}\right\}.\end{aligned}$$

Let $Q := [H_\star^{-1/2}H(z,\theta_\star)H_\star^{-1/2}]^2 - \mathbb{E}\left\{[H_\star^{-1/2}H(Z,\theta_\star)H_\star^{-1/2}]^2\right\}$. Under Assumption 3', it holds that

$$\left\|[H_\star^{-1/2}H(Z,\theta_\star)H_\star^{-1/2}]^2\right\|_2 \leq \left\|H_\star^{-1/2}H(Z,\theta_\star)H_\star^{-1/2}\right\|_2^2 \leq M_2^2.$$

As a result, it holds that $\|Q\|_2 \leq 2M_2^2$. Moreover, we have

$$\left\|\mathbb{E}[QQ^\top]\right\|_2 \leq \mathbb{E}\left\|QQ^\top\right\|_2 \leq \mathbb{E}\|Q\|_2^2 \leq 4M_2^4$$

and, similarly, $\left\|\mathbb{E}[Q]\mathbb{E}[Q^\top]\right\|_2 \leq 4M_2^4$. Consequently, $\|\mathbb{V}(Q)\|_2 \leq 8M_2^4$. This, together with Lemma 39 implies that Q satisfies a matrix Bernstein condition with $K_2 = 2M_2^2$ and $\sigma_H^2 = 8M_2^4$. Analogously, Assumption 3 holds true with $K_2 = 2M_2$ and $\sigma_H^2 = 4M_2^2$. In the following of the proof, we assume $n \geq C\max\{M_2^4\log(2p/\delta), K_1^2\log(e/\delta)p_\star(R^2/\mu_\star+1/\rho)\}$. This implies that $\|\theta_n-\theta_\star\|_{H_\star} < \rho$ on the event $\mathcal{G}_1$. Furthermore, we have $Rr_n/\sqrt{\mu_\star} \leq 1/6$ and $t_n \leq 1/6$, and thus

$$\frac{Rr_n/\sqrt{\mu_\star}+t_n}{1-Rr_n/\sqrt{\mu_\star}-t_n} \leq 1/2. \tag{32}$$

Step 1. Prove the bound on the event $\mathcal{G}_1\mathcal{G}_2\mathcal{G}_3\mathcal{G}_4$. By the event $\mathcal{G}_2$ and (32), it holds that

$$\|H_\star^{1/2}H_n(\theta_n)^{-1}H_\star^{1/2}\|_2, \|H_n(\theta_n)^{-1/2}H_\star H_n(\theta_n)^{-1/2}\|_2 \leq \frac{3}{2}, \tag{33}$$

and $H_n(\theta_n) \preceq 2H_\star$. It follows that

$$\|I_n(z) - I(z)\|^2_{H_n(\theta_n)} \le 2\|I_n(z) - I(z)\|^2_{H_\star} \quad \text{and} \quad \|I(z)\|^2_{H_n(\theta_n)} \le 2\|I(z)\|^2_{H_\star}.$$

As a result,

$$\|I_n(z)\|^2_{H_n(\theta_n)} \le 2\,\|I_n(z) - I(z)\|^2_{H_n(\theta_n)} + 2\,\|I(z)\|^2_{H_n(\theta_n)} \le 4\|I_n(z) - I(z)\|^2_{H_\star} + 4\|I(z)\|^2_{H_\star}. \tag{34}$$

By the event $\mathcal{G}_3$ and (32), it holds that

$$\|I_n(z) - I(z)\|^2_{H_\star} \le \frac{9}{2}M_2^2 r_n^2 + 2\|S(z,\theta_\star)\|^2_{H_\star^{-1}} \left(\frac{Rr_n/\sqrt{\mu_\star} + t_n}{1 - Rr_n/\sqrt{\mu_\star} - t_n}\right)^2. \tag{35}$$

On the event $\mathcal{G}_4$, we get

$$\left\|\frac{1}{n}\sum_{i=1}^n \left(H_\star^{-1/2} H(Z_i,\theta_\star) H_\star^{-1/2}\right)^2\right\|_2 \le \frac{1}{2} + \left\|\mathbb{E}\left\{[H_\star^{-1/2} H(Z,\theta_\star) H_\star^{-1/2}]^2\right\}\right\|_2 \le \frac{1}{2} + M_2^2.$$

Furthermore, by Lemma 33, it holds that

$$\|H(Z_i,\theta_n) - H(Z_i,\theta_\star)\|_{H_\star^{-1}} \le Re^{R\|\theta_n - \theta_\star\|_2}\|H(Z_i,\theta_\star)\|_{H_\star^{-1}}\|\theta_n - \theta_\star\|_2.$$

Note that $\|H(z,\theta_\star)\|_{H_\star^{-1}} \le M_2$ and $R\|\theta_n - \theta_\star\|_2 \le R\|\theta_n - \theta_\star\|_{H_\star}/\sqrt{\mu_\star} \le 1/2$ by the event $\mathcal{G}_1$. It follows that

$$\left\|H_\star^{-1/2} H(Z_i,\theta_n) H_\star^{-1/2} - H_\star^{-1/2} H(Z_i,\theta_\star) H_\star^{-1/2}\right\|_2 = \|H(Z_i,\theta_n) - H(Z_i,\theta_\star)\|_{H_\star^{-1}} \le M_2.$$

Since $\|A^2 - B^2\|_2 \le \|A(A-B)\|_2 - \|(A-B)B\|_2 \le (\|A\|_2 + \|B\|_2)\|A - B\|_2$, we get

$$\left\|\left(H_\star^{-1/2} H(Z_i,\theta_n) H_\star^{-1/2}\right)^2 - \left(H_\star^{-1/2} H(Z_i,\theta_\star) H_\star^{-1/2}\right)^2\right\|_2 \le 2M_2^2,$$

and thus

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^n \left(H_\star^{-1/2} H(Z_i,\theta_n) H_\star^{-1/2}\right)^2\right\|_2 \\
\le &\left\|\frac{1}{n}\sum_{i=1}^n \left(H_\star^{-1/2} H(Z_i,\theta_\star) H_\star^{-1/2}\right)^2\right\|_2 + \\
&\left\|\frac{1}{n}\sum_{i=1}^n \left(H_\star^{-1/2} H(Z_i,\theta_n) H_\star^{-1/2}\right)^2 - \frac{1}{n}\sum_{i=1}^n \left(H_\star^{-1/2} H(Z_i,\theta_\star) H_\star^{-1/2}\right)^2\right\|_2 \le 4M_2^2.
\end{aligned} \tag{36}$$

Putting (29), (30), (33), (34), (35), and (36) together, we obtain

$$\mathbf{Tr}(\Sigma_n) \le (CpM_2^2 - 1)\left[18M_2^2 r_n^2 + 8\|S(z,\theta_\star)\|^2_{H_\star^{-1}}\left(\frac{Rr_n/\sqrt{\mu_\star} + t_n}{1 - Rr_n/\sqrt{\mu_\star} - t_n}\right)^2 + 4\|I(z)\|^2_{H_\star}\right].$$

Now the claim follows from $\|\Sigma_n\|_2 \le \mathbf{Tr}(\Sigma_n)$ and $I(z) = H_\star^{-1} S(z,\theta_\star)$.

Step 2. Control the probability of $\mathcal{G}_1\mathcal{G}_2\mathcal{G}_3\mathcal{G}_4$. According to Propositions 10 and 11, we have $\mathbb{P}(\mathcal{G}_1) \ge 1 - \delta/4$ and $\mathbb{P}(\mathcal{G}_2) \ge 1 - \delta/4$. Following a similar proof as Theorem 1 and noticing that $\|H(z,\theta)\|_{H_\star^{-1}} \le M_2$ for all $\|\theta - \theta_\star\|_{H_\star} \le \rho$, we obtain $\mathbb{P}(\mathcal{G}_3) \ge 1 - \delta/4$. Finally, invoking the matrix Bernstein inequality yields $\mathbb{P}(\mathcal{G}_4) \ge 1 - \delta/4$. Hence, we have $\mathbb{P}(\mathcal{G}_1\mathcal{G}_2\mathcal{G}_3\mathcal{G}_4) \ge 1 - \delta$. □

F.4 Variance Reduction: SVRG and Accelerated SVRG

We minimize the quadratic g_n from (9) with SVRG (Johnson and Zhang, 2013) or its accelerated variant (Lin et al., 2018; Allen-Zhu, 2017). Let $u_\star = \arg\min_u f(u)$ denote the minimizer of $f_n(u)$. A Taylor expansion gives us the expression

$$f(u) - f(u_\star) = \frac{1}{2}\|u - u_\star\|^2_{H_n(\theta_n)}.$$

Combining this fact with standard convergence bounds of SVRG and accelerated SVRG (cf. Appendix I.5 for a review) give us the following computational bound.

Theorem 22. *Suppose that the loss function ℓ is convex and L-smooth, i.e., $0 \preceq \nabla^2 \ell(\cdot, z) \preceq L\mathbf{I}_d$ for all $z \in \mathcal{Z}$. Further, assume that f_n is μ_n strongly convex, i.e., $H_n(\theta_n) \succeq \mu_n \mathbf{I}_d$. Then, SVRG starting at $u_0 \in \mathbb{R}^d$ returns an iterate u_t satisfying $\mathbb{E}\left[\|u_t - u_\star\|^2_{H_n(\theta_n)} \middle| Z_{1:n}\right] \leq \varepsilon$ after t_{svrg} steps where*

$$t_{\mathrm{svrg}} \leq C(n + \kappa_n) \log\left(\frac{\kappa_n \|u_0 - u_\star\|^2_{H_n(\theta_n)}}{\varepsilon}\right),$$

where $\kappa_n = L/\mu_n$ and C is an absolute constant. Accelerated SVRG satisfies the same condition after t_{asvrg} steps where steps where

$$t_{\mathrm{asvrg}} \leq C\left(n + \sqrt{n\kappa_n}\right) \log\left(\frac{\kappa_n \|u_0 - u_\star\|^2_{H_n(\theta_n)}}{\varepsilon}\right).$$

This gives us the following full error bound.

Corollary 23 (Total Computational Cost; Variance Reduction)**.** *Fix $\varepsilon > 0$. Consider the setting of Theorem 1, and let $\mathcal{G}$ denote the high probability event under which its conclusions hold. Choose a sample size n such that*

$$n = C_{K_1, K_2, \sigma_H} \frac{R^2 p_\star^2}{\mu_\star \varepsilon} \operatorname{poly}\log \frac{p}{\delta}.$$

Then, the number N_{svrg} of gradient and Hessian-vector oracle calls required to obtain a point $\hat{I}_n(z)$ using SVRG initialized at $u_0 = 0$ such that $\mathbb{E}\left[\|\hat{I}_n(z) - I(z)\|^2_{H_\star} | \mathcal{G}\right] \leq \varepsilon$ is bounded by

$$N_{\mathrm{svrg}} \leq C_{K_1, K_2, \sigma_H}\ \kappa_\star \left(1 + \frac{R^2 p_\star^2}{L\varepsilon}\right)\ \log\left(\frac{\kappa_\star \|I(z)\|^2_{H_\star}}{\varepsilon} + \kappa_\star\right) \operatorname{poly}\log \frac{p}{\delta}.$$

The corresponding number N_{asvrg} for accelerated SVRG is

$$N_{\mathrm{asvrg}} \leq C_{K_1, K_2, \sigma_H}\ \kappa_\star \left(\sqrt{\frac{R^2 p_\star^2}{L\varepsilon}} + \frac{R^2 p_\star^2}{L\varepsilon}\right)\ \log\left(\frac{\kappa_\star \|I(z)\|^2_{H_\star}}{\varepsilon} + \kappa_\star\right) \operatorname{poly}\log \frac{p}{\delta}.$$

Proof. The proof is identical to that of Corollary 16 with Theorem 22 invoked instead of Proposition 15. □

F.5 Low Rank Approximation

Consider the eigenvalue decomposition $H_n(\theta_n) = Q\Lambda Q^\top$, where $\Lambda = (\lambda_1, \cdots, \lambda_p)$ contains the eigenvalues of $H_n(\theta_n)$ in non-increasing order. Recall that this method relies on approximating $H_n(\theta_n)$ with its low-rank approximation $Q\Lambda_k Q^\top$ where $\Lambda_k = \operatorname{Diag}(\lambda_1, \cdots, \lambda_k, 0, \cdots, 0)$ to approximate the product with a vector v as $H_n(\theta_n)^{-1} v = Q\Lambda^{-1} Q^\top v \approx Q\Lambda_k^+ Q^\top v$, where $\Lambda_k^+ = \operatorname{Diag}(\lambda_1^{-1}, \cdots, \lambda_k^{-1}, 0, \cdots, 0)$ is the pseudoinverse of Λ. The rank-k approximation of $v = H_n(\theta_n)^{-1} u$ is given by $v_k = Q \operatorname{Diag}(\lambda_1^{-1}, \cdots, \lambda_k^{-1}, 0 \cdots, 0) Q^\top u$.

Consequently, this section gives bounds for the method of Schioppa et al. (2022), who compute the low-rank approximation of the Hessian using the Lanczos/Arnoldi iterations (Lanczos, 1950; Arnoldi, 1951).

The computational bound we obtain depends on the low rank k.

Proposition 24. *Let $\lambda_1 \geq \cdots \geq \lambda_d$ denote the eigenvalues of $H_n(\theta_n)$. Then, the low-rank estimate $\hat{I}_{n,k}(z)$ of $I_n(z)$ satisfies*

$$\left\|\hat{I}_{n,k}(z) - I_n(z)\right\|^2_{H_n(\theta_n)} \leq \|I_n(z)\|_2^2 \sum_{i=k+1}^{p} \lambda_i.$$

We have the following two regimes depending on the decay of eigenvalues $\lambda_i(H_n(\theta_n))$:

- *If $\lambda_i(H_n(\theta_n)) \leq L\, i^{-\beta}$ for some $\beta > 1$, we have*

$$\left\|\hat{I}_{n,k}(z) - I_n(z)\right\|^2_{H_n(\theta_n)} \leq C_\beta \frac{\kappa_n \|I_n(z)\|^2_{H_n(\theta_n)}}{k^{\beta - 1}}.$$

- *If* $\lambda_i(H_n(\theta_n)) \le L\exp(-\nu(k-1))$ *for some* $\nu > 0$*, we have*

$$\left\|\hat{I}_{n,k}(z) - I_n(z)\right\|^2_{H_n(\theta_n)} \le C_\nu \kappa_n \exp(-\nu k)\|I_n(z)\|^2_{H_n(\theta_n)}\,.$$

Proof. Denote $v = \nabla\ell(\theta_n, z)$ and $u_\star = -H_n(\theta_n)^{-1}v$. Let $q_1, \cdots, q_p$ denote the columns of Q. Using $Q^\top Q = \mathbf{I}_p$, we get

$$\begin{aligned}\left\|\hat{I}_{n,k}(z) - I_n(z)\right\|^2_{H_n(\theta_n)} &= v^\top Q(\Lambda^{-1} - \Lambda_k^+)\Lambda(\Lambda^{-1} - \Lambda_k^+)Q^\top v \\ &= u_\star^\top Q\Lambda(\Lambda^{-1} - \Lambda_k^+)\Lambda(\Lambda^{-1} - \Lambda_k^+)Qu_\star \\ &= \sum_{i=k+1}^p \lambda_i\langle q_i, u\rangle_2^2 \le \sum_{i=k+1}^p \lambda_i\|u_\star\|_2^2\,,\end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality and $\|q_i\|_2 = 1$. For the second part of the proof, we use the bound $\|u\|_2^2 \le \|u\|_A^2/\lambda_{\min}(A)$ together with

$$\sum_{i=k+1}^p i^{-\beta} \le \int_k^\infty x^{-\beta}\mathrm{d}x = \frac{k^{-(\beta-1)}}{\beta-1}\,, \quad \text{and} \quad \sum_{i=k+1}^p \exp(-\nu(i-1)) \le \frac{\exp(-\nu k)}{1-\exp(-\nu)}\,.$$

□

Corollary 25 (Total Computational Cost; Low-Rank Approximation)**.** *Fix* $\varepsilon > 0$*. Consider the setting of Theorem 1, and let* $\mathcal{G}$ *denote the high probability event under its conclusions hold. Choose a sample size*

$$n \ge C_{K_1,K_2,\sigma_H,R}\frac{p_\star^2}{\mu_\star\varepsilon}\,\mathrm{poly}\log\frac{p}{\delta}\,.$$

Then, under $\mathcal{G}$*, the rank-*k *approximation* $\hat{I}_{n,k}(z)$ *satisfies* $\|\hat{I}_{n,k}(z) - I(z)\|^2_{H_\star} \le \varepsilon$ *for all* k *no smaller than*

$$k_\star = \min\left\{k \,:\, \sum_{i=k+1}^p \lambda_i(H_\star)\,\|I_n(z)\|_2^2 \le \varepsilon/32\right\}\,.$$

We have the following two regimes depending on the decay of eigenvalues $\lambda_i(H_\star)$*:*

- *If* $\lambda_i(H_\star) \le L\,i^{-\beta}$ *for some* $\beta > 1$*, we have*

$$k_\star \le C_\beta\left(\frac{\kappa_\star\|I(z)\|^2_{H_\star}}{\varepsilon} + \kappa_\star\right)^{\frac{1}{\beta-1}}\,.$$

- *If* $\lambda_i(H_\star) \le L\exp(-\nu(k-1))$ *for some* $\nu > 0$*, we have*

$$k_\star \le \frac{1}{\nu}\log\left(\frac{\kappa_\star\|I(z)\|^2_{H_\star}}{\varepsilon} + \kappa_\star\right)\,.$$

Proof. The proof follows from combining Proposition 24 with Proposition 14. □

G Most Influential Subset: Statistical Error Bound

Our goal in this section is to prove Theorem 5.

G.1 Setup

Throughout, we assume that the Hessian $\nabla^2_\theta F(\theta)$ of the population is invertible for all $\theta \in \Theta$. For a continuously differentiable test function h such as the loss of a test example $h(\theta) = \ell(z_{\text{test}}, \theta)$, recall that we define the population influence as

$$I_\alpha(h) = \sup_{Q \ll P} \left\{ -\nabla_\theta h(\theta_\star)^\top \nabla^2_\theta H_\star^{-1} \, \mathbb{E}_{Z\sim Q}[\nabla_\theta \ell(Z, \theta_\star)] \, : \, \frac{\mathrm{d}Q}{\mathrm{d}P} \le \frac{1}{1-\alpha} \right\}. \tag{37}$$

We characterize the convergence of $I_{n,\alpha}(h)$ towards $I_\alpha(h)$ via finite sample bounds. Recall that, for $A, J \in \mathbb{R}^{p\times p}$ with J being p.s.d., the weighted spectral norm $\|A\|_J := \left\|J^{1/2} A J^{1/2}\right\|_2$.

We retain Assumption 1 but strengthen the other assumptions.

Assumption 2' (Bounded Gradient)**.** *The normalized gradient is bounded in a neighborhood of $\theta_\star$, i.e., there exist $M_1 \ge 1, \rho \in (0,1]$ such that $\|\nabla \ell(z,\theta)\|_{H_\star^{-1}} \le M_1$ for all $z \in \mathcal{Z}$ and $\|\theta - \theta_\star\|_{H_\star} \le \rho$.*

If the normalized gradient $H_\star^{-1/2}\nabla\ell(z,\theta_\star)$ is bounded, then it is also sub-Gaussian, as required by Assumption 2. In addition, we make this assumption in a neighborhood of $\theta_\star$. For the next assumption, we strengthen the Bernstein condition on the normalized Hessian into a spectral norm bound in a neighborhood around $\theta_\star$.

Assumption 3' (Bounded Hessian)**.** *The normalized Hessian is bounded in a neighborhood of $\theta_\star$, i.e., there exist $M_2 \ge 1, \rho \in (0,1]$ such that $\|H(z,\theta)\|_{H_\star^{-1}} \le M_2$ for all $z \in \mathcal{Z}$ and $\|\theta - \theta_\star\|_{H_\star} \le \rho$.*

Finally, we also require that the gradient and Hessian of the test function h are bounded.

Assumption 4 (Bounded Test Function)**.** *There exist $M_1', M_2', \rho > 0$ such that $\|\nabla h(\theta)\|_{H_\star^{-1}} \le M_1'$ and $\left\|\nabla^2 h(\theta)\right\|_{H_\star^{-1}} \le M_2'$ for all $\|\theta - \theta_\star\|_{H_\star} \le \rho$.*

G.2 Proof of the Statistical Bound of Theorem 5

Recall that the maximum subset influence is defined as

$$I_{\alpha,n}(h) = \max_{w \in W_\alpha} \sum_{i=1}^n w_i v_i, \quad \text{where } v_i = -\left\langle \nabla h(\theta_n), H_n(\theta_n)^{-1} \nabla \ell(Z_i, \theta_n) \right\rangle.$$

Here $H_n(\theta_n)^{-1}\nabla\ell(Z_i,\theta_n) = -I_n(Z_i)$. Hence, the maximum subset influence can be equivalently defined as

$$I_{\alpha,n}(h) = \max_{w \in W_\alpha} \sum_{i=1}^n w_i \langle \nabla h(\theta_n), I_n(Z_i)\rangle.$$

We state and prove the precise version of Theorem 5 below. Note that we give a bound in terms of $|I_{\alpha,n}(h) - I_\alpha(h)|$ while the main paper gave a bound in terms of the square.

Theorem 5. *Under Assumptions 1, 2', 3', and 4, it holds that, with probability at least $1-\delta$,*

$$|I_{\alpha,n}(h) - I_\alpha(h)| \le \frac{C_{M_1,M_2,M_1',M_2'}}{(1-\alpha)\sqrt{n}} \left(R\sqrt{\frac{p_\star}{\mu_\star}\log\left(\frac{e}{\delta}\right)} + \sqrt{\log\left(\frac{2p}{\delta}\right)} + \sqrt{\log\left(\frac{n}{\delta}\right)} \right).$$

whenever $n \ge C_{M_1,M_2}\left(\left(\frac{R^2}{\mu_\star} + \frac{1}{\rho}\right) p_\star \log\left(\frac{e}{\delta}\right) + \log\left(\frac{2p}{\delta}\right)\right)$.

The proof centrally relies on the following duality property of the superquantile.

Lemma 26 (Rockafellar and Uryasev (2000))**.** *For any integrable random variable $Z \sim P$ and any $\alpha \in (0,1)$, the superquantile satisfies the equivalent expressions*

$$S_\alpha(Z) = \inf_{\eta \in \mathbb{R}} \left\{ \eta + \frac{1}{1-\alpha}\mathbb{E}(Z-\eta)_+ \right\} = \sup_{Q \ll P} \left\{ \mathbb{E}_{Z\sim Q}[Z] \, : \, \frac{\mathrm{d}Q}{\mathrm{d}P} \le \frac{1}{1-\alpha} \right\}.$$

We now prove Theorem 5.

Proof of Theorem 5. Define the shorthand for the per-point influence as

$$\psi_n(z,\theta) := \nabla h(\theta)^\top H_n(\theta)^{-1} \nabla \ell(z,\theta) \quad \text{and} \quad \psi(z,\theta) := \nabla h(\theta)^\top H(\theta)^{-1} \nabla \ell(z,\theta).$$

Motivated by the alternate expression for the superquantile in Lemma 26, we will define

$$\varphi_{n,n}(\theta,\eta) := \eta + \frac{1}{(1-\alpha)n} \sum_{i=1}^n \left(-\psi_n(Z_i,\theta) - \eta\right)_+ ,$$
$$\varphi_n(\theta,\eta) := \eta + \frac{1}{(1-\alpha)n} \sum_{i=1}^n \left(-\psi(Z_i,\theta) - \eta\right)_+ ,$$
$$\varphi(\theta,\eta) := \eta + \frac{1}{1-\alpha} \mathbb{E}_{Z\sim P} \left(-\psi(Z,\theta) - \eta\right)_+ .$$

According to Lemma 26, it holds that

$$|I_{\alpha,n}(h) - I_\alpha(h)| = \left| \inf_{\eta\in\mathbb{R}} \varphi_{n,n}(\theta_n,\eta) - \inf_{\eta\in\mathbb{R}} \varphi(\theta_\star,\eta) \right|,$$

By the triangle inequality,

$$\left| \inf_{\eta\in\mathbb{R}} \varphi_{n,n}(\theta_n,\eta) - \inf_{\eta\in\mathbb{R}} \varphi(\theta_\star,\eta) \right| \le \underbrace{\left| \inf_{\eta\in\mathbb{R}} \varphi_{n,n}(\theta_n,\eta) - \inf_{\eta\in\mathbb{R}} \varphi_n(\theta_n,\eta) \right|}_{\mathcal{A}} + \underbrace{\left| \inf_{\eta\in\mathbb{R}} \varphi_n(\theta_n,\eta) - \inf_{\eta\in\mathbb{R}} \varphi(\theta_\star,\eta) \right|}_{\mathcal{B}}. \tag{38}$$

As before, we prove the bound on some events and control the probability of these events. Before we start, we make two observations. First, according to Lemma 36 and Assumption 2', the sub-Gaussian gradient assumption, Assumption 2, holds true with $K_1 = CM_1$. Second, let $Q := H_\star^{-1/2} H(Z,\theta_\star) H_\star^{-1/2} - I_p$. Under Assumption 3', it holds that $\|Q\|_2 = \|H(Z,\theta_\star) - H_\star\|_{H_\star^{-1}} \le 1 + M_2 \le CM_2$. Moreover, we have

$$\left\|\mathbb{E}[QQ^\top]\right\|_2 \le \mathbb{E}\left\|QQ^\top\right\|_2 \le \mathbb{E}\|Q\|_2^2 \le C^2 M_2^2$$

and, similarly, $\left\|\mathbb{E}[Q]\mathbb{E}[Q^\top]\right\|_2 \le C^2 M_2^2$. Consequently, $\|\mathbb{V}(Q)\|_2 \le 2C^2 M_2^2$. This, together with Lemma 39, implies that Assumption 3 holds true with $K_2 = M_2$ and $\sigma_H^2 = 2C^2 M_2^2$.

Fix $\varepsilon > 0$ and denote $M := eM_1 M_1'$. Let $\mathcal{R}_\varepsilon$ be an ε-net of $[-M, M]$. It is clear that $|\mathcal{R}_\varepsilon| \le \frac{M}{\varepsilon} + 1$. Denote

$$r_n := \sqrt{CM_1^2 \frac{p_\star}{n} \log(2e/\delta)} \quad \text{and} \quad t_n := \frac{CM_2}{-1 + \sqrt{1 + Cn/\log(4p/\delta)}}.$$

Define the following events

$$\mathcal{G}_1 := \left\{ \|\nabla \ell_n(\theta_\star)\|_{H_\star^{-1}}^2 \le \frac{1}{n} CM_1^2 p_\star \log(3e/\delta) \right\}$$
$$\mathcal{G}_2 := \{(1-t_n)H_\star \preceq H_n(\theta_\star) \preceq (1+t_n)H_\star\}$$
$$\mathcal{G}_3 := \left\{ |\varphi_n(\theta_\star,\eta) - \varphi(\theta_\star,\eta)| \le \frac{M}{1-\alpha} \sqrt{\frac{2\log(6\,|\mathcal{R}_\varepsilon|/\delta)}{n}} \text{ for all } \eta \in \mathcal{R}_\varepsilon \right\}.$$

In what follows, we assume that

$$n \ge \max\left\{ CM_2^2 \log(6p/\delta), CM_1^2 p_\star \left(\frac{R^2}{\mu_\star} + \frac{1}{\rho}\right) \log(3e/\delta) \right\}. \tag{39}$$

From the proof of Proposition 11, we know that $t_n \le 1/3$,

$$\|\theta_n - \theta_\star\|_{H_\star}^2 \le r_n^2 = \frac{1}{n} CM_1^2 p_\star \log(2e/\delta) \quad \text{on the event } \mathcal{G}_1\mathcal{G}_2, \tag{40}$$

and $\mathbb{P}(\mathcal{G}_k) \geq 1 - \delta/3$ for $k \in \{1, 2\}$.

Step 1. Control $\mathcal{A}$. Since $(\cdot)_+$ is 1-Lipschitz, we get

$$\begin{aligned}|\varphi_{n,n}(\theta_n, \eta) - \varphi_n(\theta_n, \eta)| &\leq \frac{1}{(1-\alpha)n} \sum_{i=1}^n |\psi_n(Z_i, \theta_n) - \psi(Z_i, \theta_n)| \\ &\leq \frac{1}{(1-\alpha)n} \sum_{i=1}^n \|\nabla h(\theta_n)\|_{H_\star^{-1}} \left\|H_n(\theta_n)^{-1} - H(\theta_n)^{-1}\right\|_{H_\star} \|\nabla \ell(Z_i, \theta_n)\|_{H_\star^{-1}}, \end{aligned} \tag{41}$$

where the last inequality follows from the definition of matrix spectral norm. By (39) and (40), we have the $\|\theta_n - \theta_\star\|_{H_\star} \leq 1$. It then follows from Assumptions 2' and 4 that $\|\nabla \ell(Z_i, \theta_n)\|_{H_\star^{-1}} \leq M_1$ and $\|\nabla h(\theta_n)\|_{H_\star^{-1}} \leq M_1'$. It remains to control $\left\|H_n(\theta_n)^{-1} - H(\theta_n)^{-1}\right\|_{H_\star}$. By the triangle inequality, we have

$$\left\|H_n(\theta_n)^{-1} - H(\theta_n)^{-1}\right\|_{H_\star} \leq \left\|H_n(\theta_n)^{-1} - H_\star^{-1}\right\|_{H_\star} + \left\|H(\theta_n)^{-1} - H_\star^{-1}\right\|_{H_\star}.$$

The first term above has been taken care of in Proposition 11:

$$\left\|H_n(\theta_n)^{-1} - H_\star^{-1}\right\|_{H_\star} \leq \frac{Rr_n/\sqrt{\mu_\star} + t_n}{1 - Rr_n/\sqrt{\mu_\star} - t_n}.$$

The second term can be controlled similarly:

$$\left\|H(\theta_n)^{-1} - H_\star^{-1}\right\|_{H_\star} \leq \frac{Rr_n/\sqrt{\mu_\star}}{1 - Rr_n/\sqrt{\mu_\star}}.$$

Putting all together, we obtain

$$\mathcal{A} \leq \sup_{\eta \in \mathbb{R}} |\varphi_{n,n}(\theta_n, \eta) - \varphi_n(\theta_n, \eta)| \leq \frac{M_1 M_1'}{(1-\alpha)} \left(\frac{Rr_n/\sqrt{\mu_\star} + t_n}{1 - Rr_n/\sqrt{\mu_\star} - t_n} + \frac{Rr_n/\sqrt{\mu_\star}}{1 - Rr_n/\sqrt{\mu_\star}} \right). \tag{42}$$

Step 2. Control $\mathcal{B}$. On a high level, we first apply a covering number argument to restrict η to a finite number of values. We then control the absolute difference $|\varphi_n(\theta_n, \eta) - \varphi(\theta_\star, \eta)|$ on this finite subset.

Step 2.1. Restrict η to a compact subset. According to Assumptions 2' and 4, it holds that, for any $\|\theta - \theta_\star\|_{H_\star} \leq 1$,

$$|\psi(z, \theta)| \leq M_1 M_1' \left\|H(\theta)^{-1}\right\|_{H_\star} \leq M_1 M_1' e^{R\|\theta - \theta_\star\|_2},$$

where the last inequality follows from Proposition 32. Recall that we have shown $\|\theta_n - \theta_\star\|_{H_\star} \leq 1$ and $\|\theta_n - \theta_\star\|_2 \leq 1/R$. It then follows that $|\psi(z, \theta)| \leq e M_1 M_1' = M$. Consequently, we have

$$\varphi_n(\theta_n, \eta) = \begin{cases} \eta \geq \varphi_n(\theta_n, M) & \text{if } \eta \geq M \\ \eta + \frac{1}{(1-\alpha)n} \sum_{i=1}^n [\psi(Z_i, \theta) - \eta] \geq \varphi_n(\theta_n, -M) & \text{if } \eta \leq -M. \end{cases}$$

Therefore, it holds that $\inf_{\eta \in \mathbb{R}} \varphi_n(\theta_n, \eta) = \inf_{|\eta| \leq M} \varphi_n(\theta_n, \eta)$. Similarly, it can be shown that $\inf_{\eta \in \mathbb{R}} \varphi(\theta_\star, \eta) = \inf_{|\eta| \leq M} \varphi(\theta_\star, \eta)$.

Step 2.2. Restrict η to a finite subset. By the triangle inequality, we have

$$\begin{aligned}|\varphi_n(\theta_n, \eta) - \varphi_n(\theta_n, \eta')| &\leq \frac{1}{(1-\alpha)n} \sum_{i=1}^n |(-\psi(Z_i, \theta_n) - \eta)_+ - (-\psi(Z_i, \theta_n) - \eta')_+| + |\eta - \eta'| \\ &\leq \frac{1}{1-\alpha} |\eta - \eta'| + |\eta - \eta'|, \quad (\cdot)_+ \text{ is 1-Lipschitz} \\ &= \frac{2-\alpha}{1-\alpha} |\eta - \eta'|. \end{aligned}$$

For any $\eta \in [-M, M]$, we define $\pi(\eta)$ to be the projection of η onto $\mathcal{R}_\varepsilon$, i.e., $|\eta - \pi(\eta)| \leq \varepsilon$. As a result,

$$\varphi_n(\theta_n, \pi(\eta)) \leq \varphi_n(\theta_n, \eta) + \frac{2-\alpha}{1-\alpha} \varepsilon,$$

which implies

$$\inf_{\eta\in[-M,M]}\varphi_n(\theta_n,\eta)\le\inf_{\eta\in\mathcal{R}_\varepsilon}\varphi_n(\theta_n,\eta)\le\inf_{\eta\in[-M,M]}\varphi_n(\theta_n,\eta)+\frac{2-\alpha}{1-\alpha}\varepsilon.$$

Similarly,

$$\inf_{\eta\in[-M,M]}\varphi(\theta_\star,\eta)\le\inf_{\eta\in\mathcal{R}_\varepsilon}\varphi(\theta_\star,\eta)\le\inf_{\eta\in[-M,M]}\varphi(\theta_\star,\eta)+\frac{2-\alpha}{1-\alpha}\varepsilon.$$

From these results we can further conclude that

$$\begin{aligned}\left|\inf_{\eta\in[-M,M]}\varphi_n(\theta_n,\eta)-\inf_{\eta\in[-M,M]}\varphi(\theta_\star,\eta)\right|&\le\left|\inf_{\eta\in\mathcal{R}_\varepsilon}\varphi_n(\theta_n,\eta)-\inf_{\eta\in\mathcal{R}_\varepsilon}\varphi(\theta_\star,\eta)\right|+\frac{2-\alpha}{1-\alpha}\varepsilon\\&\le\sup_{\eta\in\mathcal{R}_\varepsilon}|\varphi_n(\theta_n,\eta)-\varphi(\theta_\star,\eta)|+\frac{2-\alpha}{1-\alpha}\varepsilon.\end{aligned}$$

Therefore, using the results from Step 2.1, we obtain

$$\begin{aligned}\mathcal{B}&=\left|\inf_{\eta\in[-M,M]}\varphi_n(\theta_n,\eta)-\inf_{\eta\in[-M,M]}\varphi(\theta_\star,\eta)\right|\\&\le\underbrace{\sup_{\eta\in\mathcal{R}_\varepsilon}|\varphi_n(\theta_n,\eta)-\varphi_n(\theta_\star,\eta)|}_{\mathcal{B}_1}+\underbrace{\sup_{\eta\in\mathcal{R}_\varepsilon}|\varphi_n(\theta_\star,\eta)-\varphi(\theta_\star,\eta)|}_{\mathcal{B}_2}+\frac{2-\alpha}{1-\alpha}\varepsilon.\end{aligned}\tag{43}$$

Step 2.3. Control $\mathcal{B}_1$. By the 1-Lipschitzness of $(\cdot)_+$, we have

$$|\varphi_n(\theta_n,\eta)-\varphi_n(\theta_\star,\eta)|\le\frac{1}{(1-\alpha)n}\sum_{i=1}^n|\psi(Z_i,\theta_n)-\psi(Z_i,\theta_\star)|.$$

It follows from the triangle inequality that

$$|\psi(Z_i,\theta_n)-\psi(Z_i,\theta_\star)|\le D_1+D_2+D_3,$$

where

$$\begin{aligned}D_1&:=\left|\nabla h(\theta_n)^\top[H(\theta_n)^{-1}-H_\star^{-1}]\nabla\ell(Z_i,\theta_n)\right|\\D_2&:=\left|\nabla h(\theta_n)^\top H_\star^{-1}[\nabla\ell(Z_i,\theta_n)-\nabla\ell(Z_i,\theta_\star)]\right|\\D_3&:=\left|[\nabla h(\theta_n)-\nabla h(\theta_\star)]^\top H_\star^{-1}\nabla\ell(Z_i,\theta_\star)\right|.\end{aligned}$$

Following the derivation of Step 1, it holds that

$$D_1\le M_1M_1'\frac{Rr_n/\sqrt{\mu_\star}}{1-Rr_n/\sqrt{\mu_\star}}.$$

To control D_2, we use the mean value theorem to write $\nabla\ell(Z_i,\theta_n)-\nabla\ell(Z_i,\theta_\star)=\nabla^2\ell(Z_i,\bar\theta)(\theta_n-\theta_\star)$ for some $\bar\theta\in\mathrm{conv}\{\theta_n,\theta_\star\}$. As a result,

$$D_2\le\|\nabla h(\theta_n)\|_{H_\star^{-1}}\left\|\nabla^2\ell(Z_i,\bar\theta)\right\|_{H_\star^{-1}}\|\theta_n-\theta_\star\|_{H_\star}\le M_2M_1'r_n,$$

where the last inequality follows from (40) and Assumptions 2' and 4. Similarly, we can show that $D_3\le M_1M_2'r_n$. Therefore,

$$\mathcal{B}_1\le\frac{1}{1-\alpha}\left[M_1M_1'\frac{Rr_n/\sqrt{\mu_\star}}{1-Rr_n/\sqrt{\mu_\star}}+M_1M_2'r_n+M_2M_1'r_n\right].\tag{44}$$

Step 2.4. Control $\mathcal{B}_2$. By the event $\mathcal{G}_3$, it holds that

$$\mathcal{B}_2\le\frac{M}{1-\alpha}\sqrt{\frac{2\log\left(6\left|\mathcal{R}_\varepsilon\right|/\delta\right)}{n}}\le\frac{M}{1-\alpha}\sqrt{\frac{2\log\left(12M/(\delta\varepsilon)\right)}{n}}\tag{45}$$

since $|\mathcal{R}_\varepsilon| \leq M/\varepsilon + 1 \leq 2M/\varepsilon$. Setting $\varepsilon = 1/\sqrt{n}$ and combining (38), (42), (43), (44), and (45) lead to, after simplification,

$$\left| \inf_{\eta \in \mathbb{R}} \varphi_{n,n}(\theta_n, \eta) - \inf_{\eta \in \mathbb{R}} \varphi(\theta_\star, \eta) \right| \leq \frac{C_{M_1, M_2, M_1', M_2'}}{(1-\alpha)\sqrt{n}} \left(R \sqrt{\frac{p_\star}{\mu_\star} \log\left(\frac{e}{\delta}\right)} + \sqrt{\log\left(\frac{2p}{\delta}\right)} + \sqrt{\log\left(\frac{n}{\delta}\right)} \right)$$

Step 2.5. Control $\mathbb{P}(\mathcal{G}_1 \mathcal{G}_2 \mathcal{G}_3)$**.** Recall from Step 2.1 that $|\psi(z, \theta_\star)| \leq M$ for all $z \in \mathcal{Z}$. This yields, for all $\eta \in \mathcal{R}_\varepsilon$,

$$0 \leq (-\psi(z, \theta_\star) - \eta)_+ \leq M - \eta \leq 2M.$$

Consequently, it follows from Hoeffding's inequality that $\mathbb{P}(\mathcal{G}_3) \geq 1 - \delta/3$. Since $\mathbb{P}(\mathcal{G}_k) \geq 1 - \delta/3$ for $k \in \{1, 2\}$ (Proposition 11), we obtain $\mathbb{P}(\mathcal{G}_1 \mathcal{G}_2 \mathcal{G}_3) \geq 1 - \delta$, which completes the proof. □

H Experimental Details

We conduct our experimentation on six datasets (two simulated, two small datasets from economics, and two natural language datasets). Here, we provide full details of the experimentation used in this paper. We start with the dataset and model details in Appendix H.1, hyperparameter choices in Appendix H.2, and evaluation methodology in Appendix H.3.

H.1 Data and Models

H.1.1 Linear Regression Simulation

We simulate a linear model with orthogonal design, which we solve using penalized ridge regression to illustrate the theoretical influence function bound results in Theorem 1. Following (Avella-Medina, 2017), we simulate a model $y_i = x_i^T \theta + \mu_i$ for varying sample sizes $n \in [15, 10000]$. Each x_i is i.i.d. standard normal variables and $\theta \in \mathbb{R}^9$ is fixed ahead of time. We introduce contamination into the dataset with $\mu_i = (1 - b_i)\mathcal{N}(0, 1) + b_i \mathcal{N}(0, 10)$ where $b_i \sim \text{Bernoulli}(.1)$. All experimental results are the average of 100 simulations.

H.1.2 Logistic Regression Simulation

We simulate a simple logistic regression model to illustrate the theoretical influence function bound results in Theorem 1. We simulate a model $y_i \sim \text{Binomial}(p_i)$, where $p_i = \left(1 + \exp(-(x_i^\top \theta + \mu_i))\right)^{-1}$ for varying sample sizes $n \in [15, 1000]$. Each x_i is i.i.d. standard normal variables and $\theta \in \mathbb{R}^9$ is fixed ahead of time. Similar to the linear regression case, we introduce contamination into the dataset with $\mu_i = (1 - b_i)\mathcal{N}(0, 1) + b_i \mathcal{N}(0, 10)$ where $b_i \sim \text{Bernoulli}(.1)$. All experimental results are the average of 100 simulations.

H.1.3 Oregon Medicaid Dataset

The dataset's covariates contains economic and demographic factors, as well as whether treatment was given. The goal is to predict various attributes of the health of a person.

Data. This dataset comes from the Oregon Medicaid study (Finkelstein et al., 2012). In 2008, Oregon instituted a lottery system for choosing low-income adult resident to enroll in the Medicaid program. Due to the nature of the lottery, it simulates a randomized controlled design study. A year later, a comprehensive survey was conducted on both the treatment group (those who had won the lottery) and the control group (those who did not win the lottery). We analyzed the effects of the treatment (L) on two different health outcomes: overall health indicated by a binary self-reported measure of positive (not fair, good, very good, or excellent) or negative (poor), and the number of days with good physical or mental health in the past 30 days. After removing all datapoints without entries for each response variable, we used $n = 22517$ for the overall health indicator model and $n = 20902$ for the number of days of good health model.

Models. We use ordinary least squares to solve a linear system where outcomes per individual i in a household h is denoted by y_{ih}. Since all individuals in a household chosen by the lottery can apply for Medicaid, the variable L_h is equal to one if the household h won the Medicaid lottery and zero otherwise. Lastly, we use a set of demographic and economic covariates x_i (shown in the Table 4). Using these, we estimate the following model for each response variable y_{ih} using the model:

$$y_{ih} = \theta_0 + \theta_1 L_h + \theta_2 x_i + \varepsilon_{ih} \,.$$

Variable Name	Description
hhsize	Household size including adults and children
wave_survey	Weights used for each draw of the survey (out of 8 draws)
employ_hrs	Average hours worked per week
edu	Highest level of education completed
dia_dx*	Diagnosed by a health professional with diabetes/sugar diabetes
ast_dx*	Diagnosed by a health professional with asthma
hbp_dx*	Diagnosed by a health professional with high blood pressure
emp_dx*	Diagnosed by a health professional with COPD
dep_dx*	Diagnosed by a health professional with depression or anxiety
ins_any	Currently have any type of insurance
ins_ohp*	Currently have OHP insurance
ins_private*	Currently have private insurance
ins_other*	Currently have other insurance
ins_months	Number of months (in last 6 months) have had insurance

Table 4: **Explanatory variables used in the Oregon Medicaid experimentation.** The "Variable Name" corresponds to the name used in the original analysis (Finkelstein et al., 2012), and then a brief description is given. Variables with a (*) are binary.

Therefore, the covariates for each person are $x_{ih} = (1, x_i, L_h)$, where ε_{ih} is assumed to be zero mean Gaussian noise.

We ran each model with increasing sample size; for the overall health indicator model (binary classification task) we used $n = 49, 169, 575, 1954, 6634$, and for the number of days of good health model (regression) we used $n = 49, 167, 559, 1869, 6251$. The model that ran using all the training data for each model was considered the population results. All experimental results are the average of 5 repetitions.

H.1.4 Cash Transfer

Data. The cash transfer dataset comes from a study of the impact of Progresa, a social program in Mexico that gives cash gifts to low income households (Angelucci and De Giorgi, 2009). Although, the effects on the population receiving the cash transfers is important, Angelucci and De Giorgi (2009) argue that we must also analyze the impact on the remaining members of the village that are not eligible in order to understand the full impact of the program. However, due to concerns that the non-poor households might have a large influence, the authors decided to limit the range of consumption outcomes for these households (less than 10,000). This results in robustness in the analysis for the poor household, but sensitive results for the non-poor households. For our analysis, we will only use data from time period 8. After removing all entries with no response variable (household consumption), we used the remaining $n = 19180$ datapoints.

Model. Following the analysis in Table 1 from (Angelucci and De Giorgi, 2009), we use total household consumption C_i for an individual i as the response variable, and a set of demographic and variables X_i as covariates (shown in Table 5). Lastly, we use Poor_i and Nonpoor_i, which are interaction terms between the treatment (getting cash transfer) and being a poor (non-poor) household, as our dependent variables of interest. The model is as below,

$$C_i = \theta_0 + \theta_1 \text{Poor}_i + \theta_2 \text{Nonpoor}_i + \theta_3 X_i \tag{46}$$

The model was run with increasing sample size $n = 49, 164, 540, 1775, 5835$. The model ran using all the training data for each model was considered the population results. All experimental results are the average of 5 repetition.

H.1.5 Question-Answering with zsRE

Data. This is a question-answering task, in which the inputs x_i are factual questions and the targets y_i are the answers. We used the Zero-Shot Relation Extraction (zsRE) dataset (Levy et al., 2017), with custom test/train split provided by (De Cao et al., 2021). An example of this data can be found in Table 6. We use a subsample of size 4499 for our experiments. We

Variable Name	Description
hhhsex*	Sex of head of household
hectareas	Land size (hecta-acres)
vhhnum	Number of household in the village
hhhage_cl	Age of head of household
hhhspouse_cl*	Head of household is married

Table 5: **Explanatory variables used in the Cash Transfer experimentation.** The "Variable Name" corresponds to the name used in the original analysis (Angelucci and De Giorgi, 2009), and then a brief description is given. Variables with a (*) are binary.

Task	Input (x_i)	Output (y_i)
zsRE	What country did The Laughing Cow originate?	France
WikiText	The interchange is considered by Popular Mechanics to be one of "The World's 18 Strangest Roadways" because of its height (as high as a 12-story building), its 43 permanent bridges and other unusual...	design and construction features. In 2006, the American Public Works Association named the High Five Interchange

Table 6: **Examples of the zsRE and WikiTextdataset**. The zsRE data consists of an input question x_i, and target answer y_i. The WikiText data has a paragraph as the input x_i and the next 10 token continuation as the output y_i.

take the full dataset of $n = 4499$ as the population and experiment with subsamples of size $49, 122, 182, 302$, and 743. The test dataset has size $n_{\text{test}} = 200$. All experimental results are the average of 5 repetitions.

Model. For these experiments, we use a BART-base model, which was fine-tuned on the zsRE dataset by De Cao et al. (2021). BART-base models have 12-layers, 768-hidden units, 16 heads, and 139M parameters (Lewis et al., 2020). Each model was fine-tuned on a subset of the full data of size $n \in \{49, 122, 182, 302, 743, 4499\}$. Fine-tuning was done using stochastic gradient descent using the Adam optimizer with a learning rate of $\gamma = 10^{-6}$ for 20 iterations.

H.1.6 Wikitext

Data. The next task is an open-ended text continuation task. The prompt x_i is a natural language text sequence, while the generation y_i is a 10 token continuation of the prompt. The dataset consists of random passages from WikiText-103. We use a subsample of size 1903 for our experiments. We take the full dataset of $n = 1903$ as the population and experiment with subsamples of size $40, 105, 275, 724$, and 1903. The test dataset has size $n_{\text{test}} = 200$. All experimental results are the average of 5 repetitions. An example of this data can be found in Table 6.

Model. We use a DistilGPT-2 model for this experiment, which was finetuned on the WikiText-103 dataset (Merity et al., 2017). DistilGPT2 models have 6-layers, 768-hidden units, 12 heads, and 82M parameters (Ma, 2021). Each model was fine-tuned on a subset of the full data of size $n \in \{40, 105, 275, 724\}$. Fine-tuning was done using stochastic gradient descent using Adam optimizer with a learning rate of $\gamma = 10^{-6}$ for 20 iterations.

H.2 Hyperparameters

The hyperparameters for each experimentation are detailed below.

Linear Regression Simulation. The linear simulation was run with a penalization hyperparameter for the Ridge regression, $\alpha = 10^{-3}$.

Oregon Medicaid Dataet. This was run with a regularization parameter of 0.01.

Cash Transfer Dataset. This was run with a regularization parameter of 0.01.

zsRE. Each of the methods requires a different set of hyperparameters, we list these in Table 7. We note that we use the same regularization parameter for each method $\lambda_1 = 100$. We used twice as many SGD epochs as SVRG epochs, because one iteration in SVRG takes twice as many Hessian-vector product class as SGD. We ran the Arnoldi method for 30 iteration,

Approx. Method	Hyperparameter	zsRE	WikiText
Conjugate Gradient	Max. Iterations	100	100
	Early stopping	0.01	0.01
SGD	Number of epochs	50	50
	Learning rate	5×10^{-4}	1×10^{-2}
SVRG	Number of epochs	25	25
	Learning rate	5×10^{-4}	1×10^{-3}
Arnoldi	Number of iterations	30	30
	Top_k eigen.	10	10
	Number of iterations	30	50

Table 7: **Hyperparameters for the language model experiements; zsRE and WikiTExt**.

which is less than SGD, this was due to lack of memory to run the Arnoldi method for more iterations (discussed in our limitations for this method).

WikiText. Similar to zsRE, each method requires a different set of hyperparameters, refer to Table 7. We note that we use the same regularization parameter for each method $\lambda_1 = 1$.

H.3 Evaluation Methodology and Other Details

Here, we specify the quantities that appear on the x and y axes of the plots in this paper. We also give some extra details of the experimentation.

x **Axis.** We are interested in how the empirical influence function differs from the population influences functions as sample size increases. Therefore, on the x axis we place the size of the subset (sample size) of the original population that was used to calculate the empirical influence.

y **Axis.** In each of our experimentation's we demonstrate how certain quantities change as the sample size increases. For both of the simulations and the small economic datasets, we calculate the normalized Hessian difference between the empirical influence and population's influence, $||I_n(z) - I(z)||^2_{H_\star}$.Lastly, for the y axis for both of the language model experiments (zsRE and WikiText), we compute the difference in the influence on the test set between the empirical and population influence, $G_n(z) - G(z)$.

Software. We used Python 3.7.11, Pytorch 1.10.2 and HuggingFace Transformers 4.16.2.

Hardware. All experiments were run on 4 NIVIDIA Titan V GPU with 12GB memory.

I Technical Definitions, Tools, and Results

I.1 Definitions

Theorem 27 (Integral (Cauchy) form of remainder)**.** *Let $f(x)$ be a differentiable function on interval I around a real number a and $T_{n,a}(b)$ be the nth Taylor polynomial of a real number b around a. For $n \geq 0$ and $b \neq a$ in the interval I*

$$f(b) = T_{n,a}(b) + \int_a^b \frac{f^{(n+1)}(t)}{n!}(b-t)^n dt.$$

Moreover, if $n = 0$ then

$$f(b) = f(a) + \int_a^b f'(t)dt.$$

Definition 28 (Sub-Gaussian variable)**.** *Let $S \in \mathbb{R}$ be a mean-zero random variable. We say S is sub-Gaussian with*

variance parameter σ^2, if for any $\lambda \in \mathbb{R}$

$$\mathbb{E}[\exp(\lambda S)] \leq \exp\left(\frac{\sigma^2\lambda^2}{2}\right).$$

Moreover, we define the sub-Gaussian norm of S as

$$\|S\|_{\psi_2} := \inf\left\{t > 0 : \mathbb{E}\left[\exp\left(\frac{S^2}{t^2}\right)\right] \leq 2\right\}.$$

Definition 29 (Sub-Gaussian vector)**.** *Let $S \in \mathbb{R}^p$ be a mean-zero random vector. We say S is sub-Gaussian if $\langle S, s\rangle$ is sub-Gaussian for every $s \in \mathbb{R}^p$. Moreover, we define the sub-Gaussian norm of S as*

$$\|S\|_{\psi_2} := \sup_{\|s\|_2=1} \|\langle S, s\rangle\|_{\psi_2}.$$

Note that $\|.\|_{\psi_2}$ is a norm and satisfies, e.g., the triangle inequality.

Definition 30 (Matrix Bernstein condition)**.** *Let $H \in \mathbb{R}^{p\times p}$ be a zero-mean symmetric random matrix. We say H satisfies a Bernstein condition with parameter $b > 0$ if, for all $j \geq 3$,*

$$\mathbb{E}[H^j] \preceq \frac{1}{2}j!b^{j-2}\mathbb{V}(H).$$

Definition 31 (Pseudo self-concordance)**.** *Let $\mathcal{X} \subset \mathbb{R}^p$ be open and $f : \mathcal{X} \to \mathbb{R}$ be a closed convex function. For a constant $R > 0$, we say f is pseudo self-concordant on $\mathcal{X}$ if*

$$|D^3_x f(x)[u, u, v]| \leq R\|u\|^2_{\nabla^2 f(x)}\|v\|_2$$

I.2 Implications of Pseudo Self-Concordance

We give in this section useful properties of pseudo self-concordant functions. We denote by $f : \mathbb{R}^p \to \mathbb{R}$ a pseudo self-concordant function with parameter R throughout this section.

The next result shows that the Hessian of a pseudo self-concordant function cannot vary too fast.

Proposition 32 (Bach (2010), Prop. 1)**.** *For any $x, y \in \mathbb{R}^p$, we have*

$$e^{-R\|y-x\|_2}\nabla^2 f(x) \preceq \nabla^2 f(y) \preceq e^{R\|y-x\|_2}\nabla^2 f(x).$$

We prove below a Lipschitz-type property for the normalized Hessian of a pseudo self-concordant function. Let $A, J \in \mathbb{R}^{p\times p}$ where J is p.s.d. We denote $\|A\|_J := \|J^{1/2}AJ^{1/2}\|$.

Lemma 33. *Let $J \in \mathbb{R}^{p\times p}$ be p.s.d. For any $x_1, x_2, x_\star \in \mathbb{R}^p$, we have*

$$\|\nabla^2 f(x_2) - \nabla^2 f(x_1)\|_J \leq Re^{R\|x_1-x_\star\|_2\vee\|x_2-x_\star\|_2}\|\nabla^2 f(x_\star)\|_J\|x_2 - x_1\|_2.$$

Proof. Take an arbitrary $v \in \mathbb{R}^p$ with $\|v\|_2 = 1$, and denote $\bar{v} := J^{1/2}v$. It holds that

$$|\bar{v}^\top\nabla^2 f(x_2)\bar{v} - \bar{v}^\top\nabla^2 f(x_1)\bar{v}| = |D^2 f(x_2)[\bar{v}, \bar{v}] - D^2 f(x_1)[\bar{v}, \bar{v}]| = |D^3 f(\bar{x})[\bar{v}, \bar{v}, x_2 - x_1]|$$

for some $\bar{x} \in \mathrm{Conv}\{x_1, x_2\}$ by the mean value theorem. By the pseudo self-concordance of f, we obtain

$$|D^3 f(\bar{x})[\bar{v}, \bar{v}, x_2 - x_1]| \leq R\|\bar{v}\|^2_{\nabla^2 f(\bar{x})}\|x_2 - x_1\|_2.$$

According to Proposition 32, we know $\nabla^2 f(\bar{x}) \preceq e^{R\|\bar{x}-x_\star\|_2}\nabla^2 f(x_\star)$. As a result,

$$R\|\bar{v}\|^2_{\nabla^2 f(\bar{x})}\|x_2 - x_1\|_2 \leq Re^{R\|x_1-x_\star\|_2\vee\|x_2-x_\star\|_2}\bar{v}^\top\nabla^2 f(x_\star)\bar{v}\|x_2 - x_1\|_2.$$

Therefore,

$$\begin{aligned}
\|\nabla^2 f(x_2) - \nabla^2 f(x_1)\|_J &= \sup_{\|v\|=1} |\bar{v}^\top\nabla^2 f(x_2)\bar{v} - \bar{v}^\top\nabla^2 f(x_1)\bar{v}| \\
&\leq \sup_{\|v\|=1} Re^{R\|x_1-x_\star\|_2\vee\|x_2-x_\star\|_2}\bar{v}^\top\nabla^2 f(x_\star)\bar{v}\|x_2 - x_1\|_2 \\
&\leq Re^{R\|x_1-x_\star\|_2\vee\|x_2-x_\star\|_2}\|\nabla^2 f(x_\star)\|_J\|x_2 - x_1\|_2.
\end{aligned}$$

□

The next result shows that the local distance between the minimizer of f and an arbitrary point x only depends on the local information at x. Its original version was given by Bach (2010, Proposition 2) and we state here a variant of it.

Proposition 34. *Let $x \in \mathbb{R}^p$ be such that $\nabla^2 f(x) \succ 0$. Whenever $\|\nabla f(x)\|_{\nabla^2 f(x)^{-1}} \leq \sqrt{\lambda_{\min}(\nabla^2 f(x))}/(2R)$, the function f has a unique minimizer $x_\star$ and*

$$\|x_\star - x\|_{\nabla^2 f(x)} \leq 4\|\nabla f(x)\|_{\nabla^2 f(x)^{-1}}.$$

The lemma below is an inequality for the spectral norm used in the proof of Proposition 11. Even though we prove it for general matrices A and B, we will only use it for $B = I_d$.

Lemma 35. *Let A and B be two p.d. matrices of size $p \times p$. Assume that $\|A - B\| \leq s < \lambda_{\min}(B)$. Then we have*

$$\|A^{-1} - B^{-1}\| \leq \frac{s}{\lambda_{\min}(B)\big(\lambda_{\min}(B) - s\big)}.$$

In particular, if $B = I_p$ and $\|I - A\| \leq 1$, we have

$$\|A^{-1} - I\| \leq \frac{\|I - A\|}{1 - \|I - A\|}.$$

Proof. Since $\|A - B\| \leq s$, it holds that

$$B - sI_p \preceq A \preceq B + sI_p.$$

It then follows from $\lambda_{\min}(B)I_p \preceq B$ that

$$[1 - s/\lambda_{\min}(B)]B \preceq A \preceq [1 + s/\lambda_{\min}(B)]B.$$

As a result, we obtain

$$\frac{1}{1 + s/\lambda_{\min}(B)}B^{-1} \preceq A^{-1} \preceq \frac{1}{1 - s/\lambda_{\min}(B)}B^{-1}.$$

Hence,

$$\|A^{-1} - B^{-1}\| \leq \frac{s/\lambda_{\min}(B)}{1 - s/\lambda_{\min}(B)}\|B^{-1}\| \leq \frac{s}{\lambda_{\min}(B)[\lambda_{\min}(B) - s]}.$$

□

I.3 Concentration of Random Vectors and Matrices

It follows from Vershynin (2018, Eq. (2.17)) that a bounded random vector is sub-Gaussian.

Lemma 36. *Let S be a random vector such that $\|S\|_2 \overset{a.s.}{\leq} M$ for some constant $M > 0$. Then S is sub-Gaussian with $\|S\|_{\psi_2} \leq M/\sqrt{\log 2}$.*

As a direct consequence of Vershynin (2018, Prop. 2.6.1), the sum of i.i.d. sub-Gaussian random vectors is also sub-Gaussian.

Lemma 37. *Let $S_1, \ldots, S_n$ be i.i.d. sub-Gaussian random vectors, then we have $\|\sum_{i=1}^n S_i\|_{\psi_2}^2 \leq C\sum_{i=1}^n \|S_i\|_{\psi_2}^2$.*

We call a random vector $S \in \mathbb{R}^d$ isotropic if $\mathbb{E}[S] = 0$ and $\mathbb{E}[SS^\top] = \mathbf{I}_d$. The following theorem is a tail bound for quadratic forms of isotropic sub-Gaussian random vectors.

Theorem 38 (Ostrovskii and Bach (2021), Theorem A.1)**.** *Let $S \in \mathbb{R}^d$ be an isotropic random vector with $\|S\|_{\psi_2} \leq K$, and let $J \in \mathbb{R}^{d\times d}$ be positive semi-definite. Then,*

$$\mathbb{P}(\|S\|_J^2 - \mathbf{Tr}(J) \geq t) \leq \exp\left(-c\min\left\{\frac{t^2}{K^2\|J\|_2^2}, \frac{t}{K\|J\|_\infty}\right\}\right).$$

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

$$\|S\|_J^2 - \mathbf{Tr}(J) \leq CK^2\left(\|J\|_2\sqrt{\log(e/\delta)} + \|J\|_\infty \log(1/\delta)\right), \tag{47}$$

where C is an absolute constant.

The next lemma, which follows from Wainwright (2019, Eq. (6.30)), shows that a matrix with bounded spectral norm satisfies the matrix Bernstein condition.

Lemma 39. *Let H be a zero-mean random matrix such that $\|H\|_2 \overset{a.s.}{\leq} M$ for some constant $M > 0$. Then H satisfies the matrix Bernstein condition with $b = M$ and $\sigma_H^2 = \|\mathbb{V}(H)\|_2$. Moreover, $\sigma_H^2 \leq 2M^2$.*

The next theorem is the Bernstein bound for random matrices.

Theorem 40 (Wainwright (2019), Theorem 6.17)**.** *Let $\{H_i\}_{i=1}^n$ be a sequence of zero-mean independent symmetric random matrices that satisfies the Bernstein condition with parameter $b > 0$. Then, for all $t > 0$, it holds that*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n H_i\right\| \geq t\right) \leq 2\,\mathbf{Rank}\left(\sum_{i=1}^n \mathbb{V}(H_i)\right)\exp\left\{-\frac{nt^2}{2(\sigma^2 + bt)}\right\}, \tag{48}$$

where $\sigma^2 := \frac{1}{n}\|\sum_{i=1}^n \mathbb{V}(H_i)\|_2$.

I.4 Generalized Linear Models Satisfy Theorem 1 Assumptions

The assumptions used to prove Theorem 1 hold for generalized linear models under some regularity conditions. We give two concrete examples here.

1. Least Squares: Let $\mathcal{Z} \subset B_{p,M} \times B_{1,M}$, where $B_{p,M} := \{x \in \mathbb{R}^p : \|x\|_2 \leq M\}$ for some $M > 0$. Consider the loss $\ell(z,\theta) := \frac{1}{2}(y - \theta^\top x)^2$ where $z = (x, y)$ denotes an input-output pair. Assume that $H(\theta_\star) = \mathbb{E}[XX^\top] \succ 0$.

(a) Pseudo self-concordance. Since $\nabla_\theta^2 \ell(z,\theta) = xx^\top \succeq 0$ and $\nabla_\theta^3 \ell(z,\theta) = 0$, the loss ℓ is pseudo self-concordant for all $R \geq 0$.
(b) Sub-Gaussian gradient. Note that $\|\nabla_\theta \ell(Z,\theta_\star)\|_2 = \|XX^\top\theta_\star - XY\|_2 \leq M^2(\|\theta_\star\|_2 + 1)$ and $H(\theta_\star) = \mathbb{E}[XX^\top] \succ 0$. This is sufficient to guarantee that the normalized gradient $H(\theta_\star)^{-1/2}\nabla\ell(Z,\theta_\star)$ is sub-Gaussian (cf. Lemma 36).
(c) Bernstein Hessian. Note that $\|\nabla_\theta^2 \ell(Z,\theta_\star)\|_2 = \|XX^\top\|_2 \leq M^2$, the standardized Hessian $H(\theta_\star)^{-1/2}\nabla_\theta^2\ell(Z,\theta_\star)H(\theta_\star)^{-1/2} - I_p$ satisfies the matrix Bernstein condition (cf. Lemma 39).

2. Logistic Regression: Let $\mathcal{Z} \subset B_{p,M} \times \{\pm 1\}$ for some $M > 0$. Consider the loss $\ell(z,\theta) = \log\left(1 + \exp(-y\langle\theta, x\rangle)\right)$ and let $\sigma(z) = \frac{1}{1+e^{-z}}$. Assume that $H(\theta_\star) \succ 0$.

(a) Pseudo self-concordance. Note that $\nabla_\theta^2\ell(z,\theta) = \sigma(\theta^\top x)[1 - \sigma(\theta^\top x)]xx^\top$ and $D_\theta^3\ell(z,\theta)[u,u,v] = \sigma(\theta^\top x)[1 - \sigma(\theta^\top x)][1 - 2\sigma(\theta^\top x)](u^\top x)^2(v^\top x)$. It follows that $|D_\theta^3\ell(z,\theta)[u,u,v]| \leq M\|v\|_2\|u\|^2_{\nabla^2\ell(z,\theta)}$ and thus ℓ is pseudo self-concordant with $R \geq M$.
(b) Sub-Gaussian gradient. Note that $\|\nabla_\theta\ell(Z,\theta_\star)\|_2 = \|[1 - \sigma(Y\theta_\star^\top X)]YX\|_2 \leq M$. Therefore, the normalized gradient $H(\theta_\star)^{-1/2}\nabla\ell(Z,\theta_\star)$ is sub-Gaussian (cf. Lemma 36).
(c) Bernstein Hessian. Note that $\|\nabla_\theta^2\ell(Z,\theta_\star)\|_2 \leq \|XX^\top\|_2/4 \leq M^2/4$. It follows that the standardized Hessian $H(\theta_\star)^{-1/2}\nabla_\theta^2\ell(Z,\theta_\star)H(\theta_\star)^{-1/2} - I_p$ satisfies the matrix Bernstein condition (cf. Lemma 39).

I.5 Convergence Bounds of Optimization Algorithms

We recall here the convergence bounds of various linear system solvers.

Stochastic Gradient Descent. We give here the convergence bounds of tail-averaged stochastic gradient descent (SGD) for general strongly convex quadratics from (Jain et al., 2017b,a).

Suppose we wish to minimize the function

$$f(u) = \frac{1}{2}\langle u, Au\rangle + \langle b, u\rangle\,, \tag{49}$$

where $A \in \mathbb{R}^{d\times d}$ is strictly positive definite and $b \in \mathbb{R}^d$ is given. Denote $u_\star = \arg\min_u f(u) = -A^{-1}b$.

Starting from some $u_0 \in \mathbb{R}^d$, consider the SGD iterations

$$u_{t+1} = u_t - \gamma(\hat{A}_t u_t + b)\,, \tag{50}$$

where $\hat{A}_t$ is a stochastic estimator of the Hessian A. We make the following assumptions:

(a) The Hessian estimator $\hat{A}$ of A is unbiased, i.e., $\mathbb{E}[\hat{A}] = A$. Further, we have the second moment bound $\mathbb{E}[\hat{A}^2] \preceq B^2 A$ for some $B^2 > 0$. If $\hat{A} \preceq L\mathbf{I}$ almost surely, then $B^2 \leq L$ is always true.
(b) The minimal eigenvalue of the Hessian A is bounded $\lambda_{\min}(A) \geq \mu$ for some $\mu > 0$.

The bounds depend on the covariance matrix of the stochastic gradients at $u = u_\star$:

$$\Sigma := \mathbb{E}\left[(\hat{A}u_\star + b)(\hat{A}u_\star + b)^\top\right] = \mathbb{E}\left[\hat{A}A^{-1}bb^\top A^{-1}\hat{A}\right] - bb^\top .$$

The noise contribution is characterized by the trace of the sandwich matrix

$$\sigma^2 := \mathbf{Tr}(A^{-1/2}\Sigma A^{-1/2}) = \mathbb{E}\left[u_\star^\top A^{1/2}(A^{-1/2}\hat{A}A^{-1/2} - I)^2 A^{1/2}u_\star\right] .$$

The degree of misspecification is captured by the scalar

$$\rho = \frac{d\,\|A^{-1/2}\Sigma A^{-1/2}\|_2}{\mathbf{Tr}(A^{-1/2}\Sigma A^{-1/2})} .$$

Theorem 41 ((Jain et al., 2017b,a)). *Consider the sequence $(u_t)_{t=0}^\infty$ produced by stochastic gradient descent (50) on function (49) with a step size $\gamma = 1/(2B^2)$. The tail-averaged iterate $\bar{u}_t = (2/t) = \sum_{\tau=t/2}^t u_\tau$ satisfies*

$$\mathbb{E}\|\bar{u}_\tau - u_\star\|_A^2 \leq 2\kappa \exp\left(-\frac{t}{4\kappa}\right)\|u_0 - u_\star\|_A^2 + 8(1+\rho)\frac{\sigma^2}{t} ,$$

where $\kappa = B^2/\mu$ is a condition number.

Stochastic Variance Reduced Gradient (SVRG) and its Acceleration.

Consider the optimization problem

$$\min_{u\in\mathbb{R}^d}\left[f(u) = \frac{1}{n}\sum_{i=1}^n f_i(u)\right] ,$$

where each f_i is L-smooth and convex, and f is μ-strongly convex. If each f_i is the quadratic

$$f_i(u) = \frac{1}{2}\langle u, A_i u\rangle + b ,$$

then the smoothness is equivalent to $0 \preceq A_i \preceq L\mathbf{I}_d$ for each i and the strong convexity to $A := (1/n)\sum_{i=1}^n A_i \succeq \mu\mathbf{I}_d$. Let $u_\star = \arg\min f(u)$. For the quadratic example above, we have $u_\star = A^{-1}b$

The following is the convergence bound for SVRG (Johnson and Zhang, 2013).

Theorem 42 ((Hofmann et al., 2015)). *The sequence (u_t) produced by SVRG satisfies*

$$\mathbb{E}[f(u_t) - f(u_\star)] \leq C_1\kappa \exp\left(-\frac{t}{C_2(n+\kappa)}\right)(f(u_0) - f(u_\star)) ,$$

for $\kappa = L/\mu$ and some absolute constants C_1 and C_2.

Accelerated SVRG (Lin et al., 2018; Allen-Zhu, 2017) satisfies the following bound.

Theorem 43. *The sequence (u_t) produced by accelerated SVRG satisfies*

$$\mathbb{E}[f(u_t) - f(u_\star)] \leq C_1\kappa \exp\left(-\frac{t}{C_2(n+\sqrt{n\kappa})}\right)(f(u_0) - f(u_\star)) ,$$

where $\kappa = L/\mu$ is the condition number and C_1 and C_2 are absolute constants.

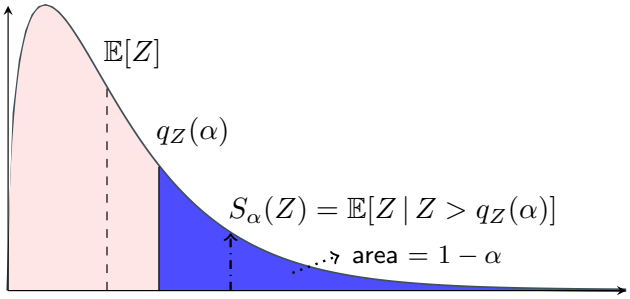

Figure 5: Expectation, quantile, and superquantile of a continuous random variable Z at level $\alpha \in (0, 1)$.

I.6 Superquantile Review

We review the various equivalent expressions of the superquantile. Consider a real-valued random variable Z with distribution P, cumulative distribution function F_Z and quantile function $q_Z(\alpha) = F_Z^{-1}(\alpha)$.

The following are equivalent expressions for the superquantile:

$$\begin{aligned} S_\alpha(Z) &= \sup \left\{ \mathbb{E}_Q[Z] \,:\, \frac{\mathrm{d}Q}{\mathrm{d}P} \leq \frac{1}{1-\alpha} \right\} \\ &= \inf_{\eta \in \mathbb{R}} \left\{ \eta + \frac{1}{1-\alpha} \mathbb{E}_P \left(Z - \eta\right)_+ \right\} \\ &= \frac{1}{1-\alpha} \int_\alpha^1 q_Z(\beta) \, \mathrm{d}\beta \,. \end{aligned} \tag{51}$$

When Z is a continuous random variable, the third expression is equivalent to (see Figure 5)

$$S_\alpha(Z) = \mathbb{E}[Z \mid Z > q_Z(\alpha)] \,.$$

When Z is discrete and takes equiprobable values $z_1, \ldots, z_n$, the three expressions above reduce to the following

$$\begin{aligned} S_\alpha(Z) &= \max \left\{ \sum_{i=1}^n w_i z_i \,:\, 0 \leq w_i \leq \frac{1}{(1-\alpha)n} \text{ for all } i \in [n] \,, \sum_{i=1}^n w_i = 1 \right\} \\ &= \min_{\eta \in \mathbb{R}} \left\{ \eta + \frac{1}{(1-\alpha)n} \sum_{i=1}^n (z_i - \eta)_+ \right\} \\ &= \frac{1}{(1-\alpha)n} \sum_{i \in I} z_i + \frac{\delta_\alpha}{1-\alpha} q_Z(\alpha) \,, \end{aligned} \tag{52}$$

where $I = \{i \,:\, z_i > q_Z(\alpha)\}$ and $\delta_\alpha = F_Z(q_Z(\alpha)) - \alpha$. Note that $\delta_\alpha = 0$ when αn is an integer.