Decomposability and restricted strong convexity

Tianhao Wang, Liangchen He

Department of Statistics and Finance, USTC

November 7, 2022

Problem

Our starting point is an indexed family of probability distributions $\{\mathbb{P}_\theta, \theta \in \Omega\}$, where θ represents some type of "parameter" to be estimated. Suppose that we observe a collection of n samples $Z_1^n = (Z_1, \ldots, Z_n)$, where each sample Z_i takes values in some space $\mathcal{Z}$, and is drawn independently according to some distribution $\mathbb{P}$. In the simplest setting, known as the well-specified case, the distribution $\mathbb{P}$ is a member of our parameterized family say $\mathbb{P} = \mathbb{P}_{\theta^*}$ and our goal is to estimate the unknown parameter θ^*.

Problem

The first ingredient of a general *M*-estimator is a cost function $\mathcal{L}_n : \Omega \times \mathcal{Z}^n \to \mathbb{R}$, where e value $\mathcal{L}_n\left(\theta; Z_1^n\right)$ provides a measure of the fit of parameter θ to the data Z_1^n. Its expectation fines the population cost function-namely the quantity

$$\overline{\mathcal{L}}(\theta) := \mathbb{E}\left[\mathcal{L}_n\left(\theta; Z_1^n\right)\right].$$

- $\mathcal{L}_n\left(\theta; Z_1^n\right) = \frac{1}{n}\sum_{i=1}^n \mathcal{L}\left(\theta; Z_i\right)$, where $\mathcal{L} : \Omega \times \mathcal{Z} \to \mathbb{R}$.

Estimator

- We define the target parameter as the minimum of the population cost function

$$\theta^* = \arg\min_{\theta\in\Omega} \overline{\mathcal{L}}(\theta).$$

- With this set-up, our goal is to estimate θ^* on the basis of the observed samples $Z_1^n = \{Z_1, \ldots, Z_n\}$. In order to do so, we combine the empirical cost function with a regularizer or penalty function $\Phi : \Omega \to \mathbb{R}$.

$$\widehat{\theta} \in \arg\min_{\theta\in\Omega} \left\{\mathcal{L}_n\left(\theta; Z_1^n\right) + \lambda_n \Phi(\theta)\right\},$$

- We will adopt $\mathcal{L}_n(\theta)$ as a shorthand for $\mathcal{L}_n\left(\theta; Z_1^n\right)$
- This chapter turn to the development of techniques for bounding the estimation error $\widehat{\Delta} = \widehat{\theta} - \theta^*$.

Example 9.1 (Linear regression and Lasso)

In this case, each sample takes the form $Z_i = (x_i, y_i)$The data are generated exactly from a linear model, so that $y_i = \langle x_i, \theta^* \rangle + w_i$, where w_i is some type of stochastic noise variable, assumed to be independent of x_i. The least-squares estimator is based on the quadratic cost function

$$\mathcal{L}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \left(y_i - \langle x_i, \theta \rangle\right)^2 = \frac{1}{2n} \|y - \mathbf{X}\theta\|_2^2,$$

Then the population cost function takes the form,

$$\mathbb{E}_{x,y}\left[\frac{1}{2}(y - \langle x, \theta \rangle)^2\right] = \frac{1}{2}\left(\theta - \theta^*\right)^{\mathrm{T}} \mathbf{\Sigma} \left(\theta - \theta^*\right) + \frac{1}{2}\sigma^2$$

Where $\Sigma := \operatorname{cov}(x_1)$ and $\sigma^2 := \operatorname{var}(w_1)$.

Definition

We now turn to the development of techniques for bounding the estimation error $\widehat{\theta} - \theta^*$. The first ingredient in our analysis is a property of the regularizer known as decomposability.

- Given a vector $\theta \in \Omega$ and a subspace S of Ω, we use $\theta_{\mathbb{S}}$ to denote the projection of θ onto S.

$$\theta_S := \arg\min_{\tilde{\theta} \in S} \|\tilde{\theta} - \theta\|^2.$$

- The notion of a decomposable regularizer is defined in terms of a pair of subspaces $M \subseteq \bar{M}$ of $\mathbb{R}^d$. The orthogonal complement of the space $\bar{M}$, namely the set

$$\bar{M}^{\perp} := \left\{ v \in \mathbb{R}^d \mid \langle u, v \rangle = 0 \quad \text{for all } u \in \bar{M} \right\},$$

(Definition 9.9) Given a pair of subspaces $M \subseteq \bar{M}$, a norm-based regularizer Φ is decomposable with respect to $\left(M, \bar{M}^{\perp}\right)$ if

$$\Phi(\alpha + \beta) = \Phi(\alpha) + \Phi(\beta) \quad \text{for all } \alpha \in M \text{ and } \beta \in \bar{M}^{\perp}.$$

- By the triangle inequality for a norm, we always have

$$\Phi(\alpha + \beta) \leq \Phi(\alpha) + \Phi(\beta)$$

- We have

$$\Phi\left(\alpha_M + \beta_{\bar{M}^{\perp}}\right) = \Phi\left(\alpha_M\right) + \Phi\left(\beta_{\bar{M}^{\perp}}\right)$$

Example 9.10 (Decomposability and sparse vectors) We begin with the ℓ_1-norm, which is the canonical example of a decomposable regularizer. Let S be a given subset of the index set $\{1, \ldots, d\}$ and S^c be its complement. We then define the model subspace

$$M \equiv M(S) := \left\{\theta \in \mathbb{R}^d \mid \theta_j = 0 \quad \text{for all } j \in S^c\right\},$$

Observe that

$$M^{\perp}(S) = \left\{\theta \in \mathbb{R}^d \mid \theta_j = 0 \quad \text{for all } j \in S\right\}.$$

With these definitions, it is then easily seen that for any pair of vectors $\alpha \in M(S)$ and $\beta \in M^{\perp}(S)$, we have

$$\|\alpha + \beta\|_1 = \|\alpha\|_1 + \|\beta\|_1,$$

showing that the ℓ_1-norm is decomposable with respect to the pair $(M(S), M^{\perp}(S))$.

A key consequence of decomposability

Given any norm $\Phi : \mathbb{R}^d \to \mathbb{R}$, its dual norm is defined in a variational manner as

$$\Phi^*(v) := \sup_{\Phi(u)\leq 1} \langle u, v\rangle.$$

Under mild regularity conditions, we have $\mathbb{E}\left[\nabla \mathcal{L}_n(\theta^*)\right] = \nabla \overline{\mathcal{L}}(\theta^*)$. Consequently, when the target parameter θ^* lies in the interior of the parameter space Ω, by the optimality conditions, the random vector $\nabla \mathcal{L}_n(\theta^*)$ has zero mean. Under ideal circumstances, we expect that the score function will not be too large.

$$\mathbb{G}(\lambda_n) := \left\{\Phi^*\left(\nabla \mathcal{L}_n(\theta^*)\right) \leq \frac{\lambda_n}{2}\right\}$$

- $\left|\langle \nabla \mathcal{L}_n(\theta^*), \Delta\rangle\right| \leq \Phi^*\left(\nabla \mathcal{L}_n(\theta^*)\right)\Phi(\Delta)$

(Proposition 9.13) Let $\mathcal{L}_n : \Omega \to \mathbb{R}$ be a convex function, let the regularizer $\Phi : \Omega \to [0, \infty)$ be a norm, and consider a subspace pair $(M, \bar{M}^{\perp})$ over which Φ is decomposable. Then conditioned on the event $\mathbb{G}(\lambda_n)$, the error $\widehat{\Delta} = \widehat{\theta} - \theta^*$ belongs to the set

$$\mathbb{C}_{\theta^*}\left(M, \bar{M}^{\perp}\right) := \left\{\Delta \in \Omega \mid \Phi\left(\Delta_{\bar{M}^{\perp}}\right) \leq 3\Phi\left(\Delta_{\bar{M}}\right) + 4\Phi\left(\theta^*_{M^{\perp}}\right)\right\}$$

When $\theta^* \in M$, $\Phi\left(\widehat{\Delta}_{\bar{M}^{\perp}}\right) \leq 3\Phi\left(\widehat{\Delta}_{\bar{M}}\right)$, and hence that

$$\Phi(\widehat{\Delta}) = \Phi\left(\widehat{\Delta}_{\bar{M}} + \widehat{\Delta}_{\bar{M}^{\perp}}\right) \leq \Phi\left(\widehat{\Delta}_{\bar{M}}\right) + \Phi\left(\widehat{\Delta}_{\bar{M}^{\perp}}\right) \leq 4\Phi\left(\widehat{\Delta}_{\bar{M}}\right)$$

Proof

Our argument is based on the function $\mathcal{F} : \Omega \to \mathbb{R}$ given by

$$\mathcal{F}(\Delta) := \mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*) + \lambda_n \{\Phi(\theta^* + \Delta) - \Phi(\theta^*)\}.$$

- By construction, we have $\mathcal{F}(0) = 0$, and so the optimality of $\widehat{\theta}$ implies that the error vector $\widehat{\Delta} = \widehat{\theta} - \theta^*$ must satisfy the condition $\mathcal{F}(\widehat{\Delta}) \leq 0$

Lemma 9.14 (Deviation inequalities) For any decomposable regularizer and parameters θ^* and Δ, we have

$$\Phi\left(\theta^* + \Delta\right) - \Phi\left(\theta^*\right) \geq \Phi\left(\Delta_{\bar{M}^\perp}\right) - \Phi\left(\Delta_{\bar{M}}\right) - 2\Phi\left(\theta^*_{M^\perp}\right).$$

Moreover, for any convex function $\mathcal{L}_n$, conditioned on the event $\mathbb{G}\left(\lambda_n\right)$, we have

$$\mathcal{L}_n\left(\theta^* + \Delta\right) - \mathcal{L}_n\left(\theta^*\right) \geq -\frac{\lambda_n}{2}\left[\Phi\left(\Delta_{\bar{M}}\right) + \Phi\left(\Delta_{\bar{M}^\perp}\right)\right].$$

- **Lemma 9.14 $\Rightarrow$ Proposition 9.13**
- $\Phi\left(\theta^* + \Delta\right) = \Phi\left(\theta^*_M + \theta^*_{M^\perp} + \Delta_{\bar{M}} + \Delta_{\bar{M}^\perp}\right)$, applying the triangle inequality yields

$$\Phi\left(\theta^* + \Delta\right) \geq \Phi\left(\theta^*_M + \Delta_{\bar{M}^\perp}\right) - \Phi\left(\theta^*_{M^\perp} + \Delta_{\bar{M}}\right).$$

-

$$\mathcal{L}_n\left(\theta^* + \Delta\right) - \mathcal{L}_n\left(\theta^*\right) \geq \left\langle \nabla \mathcal{L}_n\left(\theta^*\right), \Delta\right\rangle \geq -\left|\left\langle \nabla \mathcal{L}_n\left(\theta^*\right), \Delta\right\rangle\right|.$$

We begin by describing the notion of restricted strong convexity.Given any differentiable cost function.

$$\mathcal{E}_n(\Delta) := \mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*) - \langle \nabla \mathcal{L}_n(\theta^*), \Delta \rangle .$$

Whenever the function $\theta \mapsto \mathcal{L}_n(\theta)$ is convex.Strong convexity requires that this lower bound holds with a quadratic slack: in particular, for a given norm $\|\cdot\|$, the cost function is locally κ-strongly convex at θ^*

$$\mathcal{E}_n(\Delta) \geq \frac{\kappa}{2}\|\Delta\|^2$$

Definition 9.15 For a given norm $\|\cdot\|$ and regularizer $\Phi(\cdot)$, the cost function satisfies a restricted strong convexity (RSC) condition with radius $R > 0$, curvature $\kappa > 0$ and tolerance τ_n^2 if

$$\mathcal{E}_n(\Delta) \geq \frac{\kappa}{2}\|\Delta\|^2 - \tau_n^2\Phi^2(\Delta) \quad \text{for all } \Delta \in B(R)$$

Definition(9.18) (Subspace Lipschitz constant) For any subspace S of $\mathbb{R}^d$, the subspace Lipschitz constant with respect to the pair $(\Phi, \|\cdot\|)$ is given by

$$\Psi(S) := \sup_{u \in \mathbb{S} \setminus \{0\}} \frac{\Phi(u)}{\|u\|}.$$

- To illustrate its use, let us consider it in the special case when $\theta^* \in \mathbb{M}$. Then for any $\Delta \in \mathbb{C}_{\theta^*}\left(M, \bar{M}^{\perp}\right)$, we have

$$\Phi(\Delta) \leq \Phi\left(\Delta_{\bar{M}}\right) + \Phi\left(\Delta_{\bar{M}^{\perp}}\right) \leq 4\Phi\left(\Delta_{\bar{M}}\right) \leq 4\Psi(\bar{M})\|\Delta\|,$$

- **(***Example***)** M is a subspace of s-sparse vectors, then with regularizer $\Phi(u) = \|u\|_1$ and error norm $\|u\| = \|u\|_2$, we have $\Psi(M) = \sqrt{s}$. In this way, we see the familiar inequality $\|\Delta\|_2 \leq 4\sqrt{s}\|\Delta\|_1$

Thus far, we have discussed the notion of decomposable regularizers, and some related notions of restricted curvature for the cost function. In this section, we state and prove some results on the estimation error, namely, the quantity $\widehat{\theta} - \theta^*$.

- (A1) The cost function is convex, and satisfies the local RSC condition with curvature κ, radius R.
  (A2) There is a pair of subspaces $M \subseteq \bar{M}$ such that the regularizer decomposes over $(M, \bar{M}^{\perp})$.
  (A3) The "good" event $\mathbb{G}(\lambda_n) := \left\{\Phi^*\left(\nabla \mathcal{L}_n(\theta^*)\right) \leq \frac{\lambda_n}{2}\right\}$.
- Our bound involves the quantity

$$\varepsilon_n^2\left(\bar{M}, M^{\perp}\right) := \underbrace{9\frac{\lambda_n^2}{\kappa^2}\Psi^2(\bar{M})}_{\text{estimation error}} + \underbrace{\frac{8}{\kappa}\left\{\lambda_n \Phi\left(\theta^*_{M^{\perp}}\right) + 16\tau_n^2 \Phi^2\left(\theta^*_{M^{\perp}}\right)\right.}_{\text{approximation error}},$$

Theorem(9.19) (Bounds for general models)

(a) Any optimal solution satisfies the bound

$$\Phi\left(\widehat{\theta}-\theta^*\right) \leq 4\left\{\Psi(\bar{M})\left\|\widehat{\theta}-\theta^*\right\| + \Phi\left(\theta^*_{M^\perp}\right)\right\}.$$

(b) For any subspace pair $\left(\bar{M}, M^\perp\right)$ such that $\tau_n^2\Psi^2(\bar{M}) \leq \frac{\kappa}{64}$ and $\varepsilon_n\left(\bar{M}, M^\perp\right) \leq R$, we have

$$\left\|\widehat{\theta}-\theta^*\right\|^2 \leq \varepsilon_n^2\left(\bar{M}, M^\perp\right)$$

- Suppose that,the optimal parameter θ^* belongs to M.

$$\Phi\left(\widehat{\theta}-\theta^*\right) \leq 6\frac{\lambda_n}{\kappa}\Psi^2(\bar{M})$$

$$\left\|\widehat{\theta}-\theta^*\right\|^2 \leq 9\frac{\lambda_n^2}{\kappa^2}\Psi^2(\bar{M})$$

- $\Phi(\widehat{\Delta}) \leq \Phi\left(\widehat{\Delta}_{\bar{M}}\right) + \Phi\left(\widehat{\Delta}_{\bar{M}^{\perp}}\right)$
- $\mathbb{K}(\delta) := \mathbb{C} \cap \{\|\Delta\| = \delta\}$.

Lemma 9.21 If $\mathcal{F}(\Delta) > 0$ for all vectors $\Delta \in \mathbb{K}(\delta)$, then $\|\widehat{\Delta}\| \leq \delta$.

Proof of Lemma:

- If $\|\widehat{\Delta}\| > \delta$, then since $\mathbb{C}$ is star-shaped around the origin , the line joining $\widehat{\Delta}$ to 0 must intersect the set $\mathbb{K}(\delta)$ at some intermediate point of the form $t^*\widehat{\Delta}$ for some $t^* \in [0, 1]$.
- $\mathcal{F}\left(t^*\widehat{\Delta}\right) = \mathcal{F}\left(t^*\widehat{\Delta} + (1 - t^*)\, 0\right) \leq t^*\mathcal{F}(\widehat{\Delta}) + (1 - t^*)\,\mathcal{F}(0)$

Proof

- $\mathcal{F}(\Delta) \geq \langle \nabla \mathcal{L}_n(\theta^*), \Delta \rangle + \frac{\kappa}{2}\|\Delta\|^2 - \tau_n^2 \Phi^2(\Delta) + \lambda_n \left\{ \Phi\left(\Delta_{\bar{M}^\perp}\right) - \Phi\left(\Delta_{\bar{M}}\right) - 2\Phi\left(\theta^*_{M^\perp}\right) \right\}$
- $\mathcal{F}(\Delta) \geq \frac{\kappa}{2}\|\Delta\|^2 - \tau_n^2 \Phi^2(\Delta) - \frac{\lambda_n}{2} \left\{ 3\Phi\left(\Delta_{\bar{M}}\right) + 4\Phi\left(\theta^*_{M^\perp}\right) \right\}$
- $$\begin{aligned} \Phi^2(\Delta) \leq \left\{ 4\Phi\left(\Delta_{\bar{M}}\right) + 4\Phi\left(\theta^*_{M^\perp}\right) \right\}^2 &\leq 32\Phi^2\left(\Delta_{\bar{M}}\right) + 32\Phi^2\left(\theta^*_{M^\perp}\right) \\ &\leq 32\Psi^2(\bar{M})\|\Delta\|^2 + 32\Phi^2\left(\theta^*_{M^\perp}\right) \end{aligned}$$
- $0 \geq \frac{\kappa}{4}\|\Delta\|^2 - \frac{3\lambda_n}{2}\Psi(\bar{M})\|\Delta\| - 32\tau_n^2\Phi^2\left(\theta^*_{M^\perp}\right) - 2\lambda_n \Phi\left(\theta^*_{M^\perp}\right)$

Bounds under curvature

A differentiable function $\mathcal{L}_n$ is locally κ-strongly convex at θ^*,if and only if

$$\langle \nabla \mathcal{L}_n (\theta^* + \Delta)) - \nabla \mathcal{L}_n (\theta^*) , \Delta \rangle \geq \kappa \|\Delta\|^2$$

When the underlying norm $\|\cdot\|$ is the ℓ_2-norm, combined with the Cauchy-Schwarz inequality, implies that

$$\left\| \nabla \mathcal{L}_n (\theta^* + \Delta) - \nabla \mathcal{L}_n (\theta^*) \right\|_2 \geq \kappa \|\Delta\|_2$$

Definition 9.22 The cost function satisfies a Φ^*-norm curvature condition with curvature κ, tolerance τ_n and radius R if

$$\Phi^* (\nabla \mathcal{L}_n (\theta^* + \Delta) - \nabla \mathcal{L}_n (\theta^*)) \geq \kappa \Phi^*(\Delta) - \tau_n \Phi(\Delta)$$

for all $\Delta \in \mathbb{B}_{\Phi^*}(R) := \{\theta \in \Omega \mid \Phi^*(\theta) \leq R\}$.

(A1') The cost satisfies the Φ^*-curvature condition with parameters $(\kappa, \tau_n; R)$.
(A2) The regularizer is decomposable with respect to the subspace pair $\left(M, \bar{M}^{\perp}\right)$ with $M \subseteq \bar{M}$.
Theorem 9.24 Given a target parameter $\theta^* \in M$, consider the regularized M-estimator (9.3) under conditions $(A1')$ and $(A2)$, and suppose that $\tau_n \Psi^2(\bar{M}) < \frac{\kappa}{32}$. Conditioned on the event $\mathbb{G}(\lambda_n) \cap \left\{\Phi^*\left(\widehat{\theta} - \theta^*\right) \leq R\right\}$, any optimal solution $\widehat{\theta}$ satisfies the bound

$$\Phi^*\left(\widehat{\theta} - \theta^*\right) \leq 3\frac{\lambda_n}{\kappa}.$$

By standard optimality conditions for a convex program, for any optimum $\widehat{\theta}$, there must exist a subgradient vector $\widehat{z} \in \partial\Phi(\widehat{\theta})$ such that $\nabla\mathcal{L}_n(\widehat{\theta}) + \lambda_n\widehat{z} = 0$. Introducing the error vector $\widehat{\Delta} := \widehat{\theta} - \theta^*$, some algebra yields

$$\nabla\mathcal{L}_n\left(\theta^* + \widehat{\Delta}\right) - \nabla\mathcal{L}_n\left(\theta^*\right) = -\nabla\mathcal{L}_n\left(\theta^*\right) - \lambda_n\widehat{z}$$

Taking the Φ^*-norm of both sides and applying the triangle inequality yields

$$\Phi^*\left(\nabla\mathcal{L}_n\left(\theta^* + \Delta\right) - \nabla\mathcal{L}_n\left(\theta^*\right)\right) \leq \Phi^*\left(\nabla\mathcal{L}_n\left(\theta^*\right)\right) + \lambda_n\Phi^*(z).$$

On one hand, on the event $\mathbb{G}(\lambda_n)$, we have that $\Phi^*(\nabla \mathcal{L}_n(\theta^*)) \le \lambda_n/2$, whereas, on the other hand, $\Phi^*(z) \le 1$. Putting together the pieces, we find that $\Phi^*(\nabla \mathcal{L}_n(\theta^* + \Delta) - \nabla \mathcal{L}_n(\theta^*)) \le \frac{3\lambda_n}{2}$. Finally, applying the curvature condition, we obtain

$$\kappa \Phi^*(\widehat{\Delta}) \le \frac{3}{2}\lambda_n + \tau_n \Phi(\widehat{\Delta})$$

It remains to bound $\Phi(\widehat{\Delta})$ in terms of the dual norm $\Phi^*(\widehat{\Delta})$. Since this result is useful in other contexts, we state it as a separate lemma here:

Lemma 9.25 If $\theta^* \in M$, then

$$\Phi(\Delta) \le 16\Psi^2(\bar{M})\Phi^*(\Delta) \quad \text{for any } \Delta \in \mathbb{C}_{\theta^*}\left(M, \bar{M}^{\perp}\right)$$

Generalized linear models with sparsity

(G1) The covariates are C-column normalized: $\max_{j=1,\ldots,d} \sqrt{\frac{\sum_{j=1}^{d} x_{ij}^2}{n}} \leq C$.

(G2) Conditionally on x_i , each response y_i is drawn i.i.d. according to a conditional distribution of the form

$$\mathbb{P}_{\theta^*}(y \mid x) = h_\sigma(y) \exp\left\{ \frac{y \langle x, \theta^* \rangle - \psi\left(\langle x, \theta^* \rangle\right)}{c(\sigma)} \right\},$$

where the partition function ψ has a bounded second derivative $\left(\|\psi''\|_\infty \leq B^2\right)$. We analyze the ℓ_1-regularized version of the GLM log-likelihood estimator, namely

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \Big\{ \underbrace{\frac{1}{n} \sum_{i=1}^{n} \left\{ \psi\left(\langle x_i, \theta \rangle\right) - y_i \langle x_i, \theta \rangle \right\}}_{\mathcal{L}_n(\theta)} + \lambda_n \|\theta\|_1 \Big\}. \tag{9.61}$$

Bounds under restricted strong convexity

We begin by proving bounds when the Taylor-series error around θ^* associated with the negative log-likelihood (9.61) satisfies the RSC condition

$$\mathcal{E}_n(\Delta) \geq \frac{\kappa}{2}\|\Delta\|_2^2 - c_1 \frac{\log d}{n}\|\Delta\|_1^2 \quad \text{for all } \|\Delta\|_2 \leq 1. \tag{9.62}$$

The following result applies to any solution $\widehat{\theta}$ of the GLM Lasso (9.61) with regularization parameter $\lambda_n = 4BC\left\{\sqrt{\frac{\log d}{n}} + \delta\right\}$ for some $\delta \in (0, 1)$.

Corollary (9.26)

Consider a GLM satisfying conditions (G1) and (G2), the RSC condition (9.62), and suppose the true regression vector θ^* *is supported on a subset S of cardinality s. Given a sample size n large enough to ensure that* $s\left\{\lambda_n^2 + \frac{\log d}{n}\right\} < \min\left\{\frac{4\kappa^2}{9}, \frac{\kappa}{64c_1}\right\}$, *any GLM Lasso solution* $\widehat{\theta}$ *satisfies the bounds*

$$\left\|\widehat{\theta} - \theta^*\right\|_2^2 \leq \frac{9s\lambda_n^2}{\kappa^2} \quad \text{and} \quad \left\|\widehat{\theta} - \theta^*\right\|_1 \leq \frac{12}{\kappa} s\lambda_n, \tag{9.63}$$

both with probability at least $1 - 2e^{-2n\delta^2}$.

Proof

Both results follow via an application of Corollary 9.20 with the subspaces

$$\mathbb{M}(S) = \overline{\mathbb{M}}(S) = \left\{\theta \in \mathbb{R}^d \mid \theta_j = 0 \quad \text{for all } j \notin S\right\}.$$

With this choice, note that we have $\Psi^2(\mathbb{M}) = s$; moreover, the assumed RSC condition (9.62) is a special case of our general definition with $\tau_n^2 = c_1 \frac{\log d}{n}$. In order to apply Corollary 9.20, we need to ensure that $\tau_n^2 \Psi^2(\mathbb{M}) < \frac{k}{64}$, and since the local RSC holds over a ball with radius R=1 , we also need to ensure that $\frac{9\Psi^2(\mathbb{M})\lambda_n^2}{\kappa^2} < 1$. Both of these conditions are guaranteed by our assumed lower bound on the sample size.

Proof

The only remaining step is to verify that the good event $\mathbb{G}(\lambda_n)$ holds with the probability stated in Corollary 9.26. Given the form (9.61) of the GLM log-likelihood, we can write the score function as the i.i.d. sum $\nabla \mathcal{L}_n(\theta^*) = \frac{1}{n}\sum_{i=1}^{n} V_i$, where $V_i \in \mathbb{R}^d$ is a zero-mean random vector with components

$$V_{ij} = \{\psi'(\langle x_i, \theta^* \rangle) - y_i\} x_{ij}.$$

Let us upper bound the moment generating function of these variables. For any $t \in \mathbb{R}$, we have

$$\begin{aligned}\log \mathbb{E}\left[e^{-tV_{ij}/n}\right] &= \log \mathbb{E}\left[e^{ty_i x_{ij}/n}\right] - tx_{ij}\psi'(\langle x_i, \theta^* \rangle)/n \\ &= \psi(tx_{ij}/n + \langle x_i, \theta^* \rangle) - \psi(\langle x_i, \theta^* \rangle) - tx_{ij}\psi'(\langle x_i, \theta^* \rangle)/n.\end{aligned}$$

Proof

By a Taylor-series expansion, there is some intermediate $\tilde{t}$ such that

$$\log \mathbb{E}\left[e^{-tV_{ij}/n}\right] = \frac{1}{2} t^2 x_{ij}^2 \psi''\left(\tilde{t} x_{ij} + \langle x_i, \theta^* \rangle\right)/n^2 \leq \frac{B^2 t^2 x_{ij}^2}{2n^2},$$

where the final inequality follows from the boundedness condition (G2). Using independence of the samples, we have

$$\log \mathbb{E}\left[e^{-t\sum_{i=1}^n V_{ij}/n}\right] \leq \frac{t^2 B^2}{2n}\left(\frac{1}{n}\sum_{i=1}^n x_{ij}^2\right) \leq \frac{t^2 B^2 C^2}{2n},$$

where the final step uses the column normalization (G1) on the columns of the design matrix $\mathbf{X}$. Since this bound holds for any $t \in \mathbb{R}$, we have shown that $\sum_{i=1}^n V_{ij}/n$ is zero-mean and sub-Gaussian with parameter at most $BC/\sqrt{n}$.

Proof

Thus, sub-Gaussian tail bounds combined with the union bound guarantee that

$$\mathbb{P}\left[\left\|\nabla \mathcal{L}_n\left(\theta^*\right)\right\|_\infty \geq t\right] \leq 2\exp\left(-\frac{nt^2}{2B^2C^2} + \log d\right).$$

Setting $t = 2BC\left\{\sqrt{\frac{\log d}{n}} + \delta\right\}$ completes the proof.

Now we consider the ℓ_∞-curvature condition

$$\left\|\nabla \mathcal{L}_n\left(\theta^* + \Delta\right) - \nabla \mathcal{L}_n\left(\theta^*\right)\right\|_\infty \geq \kappa\|\Delta\|_\infty - \frac{c_0}{32}\sqrt{\frac{\log d}{n}}\|\Delta\|_1, \quad (9.64)$$

for all $\|\Delta\|_\infty \leq 1$.

Bounds under ℓ_∞-curvature conditions

Corollary (9.27)

In addition to the conditions of Corollary 9.26, suppose that the ℓ_∞-curvature condition (9.64) holds, and that $n > c_0^2 s^2 \log d$. Then any optimal solution $\widehat{\theta}$ to the GLM Lasso (9.61) with regularization parameter $\lambda_n = 4BC\left(\sqrt{\frac{\log d}{n}} + \delta\right)$ satisfies

$$\left\|\widehat{\theta} - \theta^*\right\|_\infty \leq 3\frac{\lambda_n}{\kappa} \tag{9.65}$$

with probability at least $1 - 2e^{-2n\delta^2}$.

Proof

We prove this corollary by applying Theorem 9.24 with the familiar subspaces

$$\overline{\mathbb{M}}(S) = \mathbb{M}(S) = \left\{\theta \in \mathbb{R}^d \mid \theta_{S^c} = 0\right\},$$

for which we have $\Psi^2(\overline{\mathbb{M}}(S)) = s$. By assumption (9.64), the ℓ_∞ -curvature condition holds with tolerance $\tau_n = \frac{c_0}{32}\sqrt{\frac{\log d}{n}}$, so that the condition $\tau_n \Psi^2(\mathbb{M}) < \frac{\kappa}{32}$ is equivalent to the lower bound $n > c_0^2 s^2 \log d$ on the sample size.

Since we have assumed the conditions of Corollary 9.26, we are guaranteed that the error vector $\widehat{\Delta} = \widehat{\theta} - \theta^*$ satisfies the bound $\|\widehat{\Delta}\|_\infty \leq \|\widehat{\Delta}\|_2 \leq 1$ with high probability. This localization allows us to apply the local ℓ_∞ -curvature condition to the error vector $\widehat{\Delta} = \widehat{\theta} - \theta^*$.

Proof

Finally, as shown in the proof of Corollary 9.26 , if we choose the regularization parameter $\lambda_n = 4BC\left\{\sqrt{\frac{\log d}{n}} + \delta\right\}$, then the event $\mathbb{G}(\lambda_n)$ holds with probability at least $1 - e^{-2n\delta^2}$. We have thus verified that all the conditions needed to apply Theorem 9.24 are satisfied.

Now we focus on the ℓ_2-version of the group Lasso penalty $\|\theta\|_{\mathcal{G},2} = \sum_{g\in\mathcal{G}} \|\theta_g\|_2$

$$\widehat{\theta} \in \arg\min_{\theta\in\mathbb{R}^d} \left\{ \frac{1}{n}\sum_{i=1}^{n} \{\psi(\langle\theta, x_i\rangle) - y_i\langle\theta, x_i\rangle\} + \lambda_n \sum_{g\in\mathcal{G}} \|\theta_g\|_2 \right\}, \quad (9.66)$$

Letting $\mathbf{X}_g \in \mathbb{R}^{n\times|g|}$ denote the submatrix indexed by g , we also impose the following variant of condition (G1) on the design:
(G1') The covariates satisfy the group normalization condition $\max_{g\in\mathcal{G}} \frac{\|\mathbf{X}_g\|_2}{\sqrt{n}} \le C$.
Moreover, we assume an RSC condition of the form

$$\mathcal{E}_n(\Delta) \ge \kappa\|\Delta\|_2^2 - c_1\left\{\frac{m}{n} + \frac{\log|\mathcal{G}|}{n}\right\}\|\Delta\|_{\mathcal{G},2}^2 \quad \text{for all } \|\Delta\|_2 \le 1, \quad (9.67)$$

where m denotes the maximum size over all groups.

Our bound applies to any solution $\widehat{\theta}$ to the group GLM Lasso (9.66) based on a regularization parameter

$$\lambda_n = 4BC\left\{\sqrt{\frac{m}{n}} + \sqrt{\frac{\log|\mathcal{G}|}{n}} + \delta\right\} \quad \text{for some } \delta \in (0,1).$$

Corollary (9.28)

Given n i.i.d. samples from a GLM satisfying conditions (G1') , (G2), the RSC condition (9.67), suppose that the true regression vector θ^* *has group support* $S_{\mathcal{G}}$ *. As long as* $\left|S_{\mathcal{G}}\right|\left\{\lambda_n^2 + \frac{m}{n} + \frac{\log|\mathcal{G}|}{n}\right\} < \min\left\{\frac{4\kappa^2}{9}, \frac{\kappa}{64c_1}\right\}$ *, the estimate* $\widehat{\theta}$ *satisfies the bound*

$$\left\|\widehat{\theta} - \theta^*\right\|_2^2 \leq \frac{9}{4}\frac{|S_G|\,\lambda_n^2}{\kappa^2} \tag{9.68}$$

with probability at least $1 - 2e^{-2n\delta^2}$ *.*

Proof

First we need to ensure that $\tau_n^2 \Psi^2(\mathbb{M}) < \frac{\kappa}{64}$, and since the local RSC holds over a ball with radius R=1 , we also need to ensure that $\frac{9\Psi^2(M)\lambda_n^2}{\kappa^2} < 1$. Both of these conditions are guaranteed by our assumed lower bound on the sample size.

It remains to verify that, given the specified choice of regularization parameter λ_n , the event $\mathbb{G}\left(\lambda_n\right)$ holds with high probability. Using the form of the dual norm given in Table 9.1, we have $\Phi^*\left(\nabla \mathcal{L}_n\left(\theta^*\right)\right) = \max_{g \in \mathcal{G}} \left\|\left(\nabla \mathcal{L}_n\left(\theta^*\right)\right) g\right\|_2$. Based on the form of the GLM log-likelihood, we have $\nabla \mathcal{L}_n\left(\theta^*\right) = \frac{1}{n} \sum_{i=1}^n V_i$ where the random vector $V_i \in \mathbb{R}^d$ has components $V_{ij} = \left\{\psi'\left(\left\langle x_i, \theta^*\right\rangle\right) - y_i\right\} x_{ij}$. For each group g , we let $V_{i,g} \in \mathbb{R}^{|g|}$ denote the subvector indexed by elements of g . With this notation, we then have

$$\left\|\left(\nabla \mathcal{L}_n\left(\theta^*\right)\right)_g\right\|_2=\left\|\frac{1}{n} \sum_{i=1}^n V_{i, s}\right\|_2=\sup _{u \in \mathcal{S} p-1}\left\langle u, \frac{1}{n} \sum_{i=1}^n V_{i, g}\right\rangle,$$

where $\mathbb{S}^{|g|-1}$ is the Euclidean sphere in $\mathbb{R}^{|g|}$. From Example 5.8, we can find a 1/2-covering of $\mathbb{S}^{|g|-1}$ in the Euclidean norm-say $\left\{u^1, \ldots, u^N\right\}$ -with cardinality at most $N \leq 5^{|g|}$. By the standard discretization arguments from Chapter 5 , we have

$$\left\|\left(\nabla \mathcal{L}_n\left(\theta^*\right)\right)_g\right\|_2 \leq 2 \max _{j=1, \ldots, N}\left\langle u^j, \frac{1}{n} \sum_{i=1}^n V_{i, g}\right\rangle .$$

Using the same proof as Corollary 9.26, the random variable $\left\langle u^j, \frac{1}{n} \sum_{i=1}^n V_{i, g}\right\rangle$ is sub-Gaussian with parameter at most

$$\frac{B}{\sqrt{n}} \sqrt{\frac{1}{n} \sum_{i=1}^n\left\langle u^j, x_{i, g}\right\rangle^2} \leq \frac{B C}{\sqrt{n}},$$

where the inequality follows from condition (G1’). Consequently, from the union bound and standard sub-Gaussian tail bounds, we have

$$\mathbb{P}\left[\left\|\left(\nabla\mathcal{L}_n\left(\theta^*\right)\right)_g\right\|_2 \geq 2t\right] \leq 2\exp\left(-\frac{nt^2}{2B^2C^2} + |g|\log 5\right).$$

Taking the union over all $|\mathcal{G}|$ groups yields

$$\mathbb{P}\left[\max_{g\in\mathcal{G}}\left\|\left(\nabla\mathcal{L}_n\left(\theta^*\right)\right)_g\right\|_2 \geq 2t\right] \leq 2\exp\left(-\frac{nt^2}{2B^2C^2} + m\log 5 + \log|\mathcal{G}|\right),$$

where we have used the maximum group size m as an upper bound on each group size $|g|$. Setting $t^2 = \lambda_n^2$ yields the result.

Let $\theta \in \mathbb{R}^d$ be a vector, and consider the ℓ_1-plus-group overlap norm

$$\Phi_\omega(\theta) := \inf_{\alpha+\beta=\theta} \{\|\alpha\|_1 + \omega\|\beta\|_{\mathcal{G},2}\}, \tag{9.73}$$

where $\mathcal{G}$ is a set of disjoint groups, each of size at most m . We use the weight

$$\omega := \frac{\sqrt{m} + \sqrt{\log |\mathcal{G}|}}{\sqrt{\log d}}. \tag{9.74}$$

With this set-up, the following result applies to the adaptive group GLM Lasso,

$$\widehat{\theta} \in \arg\min_{\theta\in\mathbb{R}^d} \Big\{ \underbrace{\frac{1}{n}\sum_{i=1}^{n} \{\psi(\langle\theta, x_i\rangle) - \langle\theta, x_i y_i\rangle\}}_{\mathcal{L}_n(\theta)} + \lambda_n \Phi_\omega(\theta) \Big\}, \tag{9.75}$$

for which the Taylor-series error satisfies the RSC condition

$$\mathcal{E}_n(\Delta) \geq \frac{\kappa}{2}\|\Delta\|_2^2 - c_1 \frac{\log d}{n} \Phi_\omega^2(\Delta) \quad \text{for all } \|\Delta\|_2 \leq 1. \tag{9.76}$$

With this set-up, the following result applies to any optimal solution $\widehat{\theta}$ of the adaptive group GLM Lasso (9.75) with $\lambda_n = 4BC\left(\sqrt{\frac{\log d}{n}} + \delta\right)$ for some $\delta \in (0,1)$. Moreover, it supposes that the true regression vector can be decomposed as $\theta^* = \alpha^* + \beta^*$, where α is S_{elt} -sparse, and β^* is $S_{\mathcal{G}}$-group-sparse, and with $S_{\mathcal{G}}$ disjoint from S_{elt} .

Corollary (9.31)

Given n i.i.d. samples from a GLM satisfying conditions (G1') and (G2), suppose that the RSC condition (9.76) with curvature $\kappa > 0$ *holds, and that* $\left\{\sqrt{|S_{elt}|} + \omega\sqrt{|S_G|}\right\}^2 \left\{\lambda_n^2 + \frac{\log d}{n}\right\} < \min\left\{\frac{\kappa^2}{36}, \frac{\kappa}{64c_1}\right\}$. *Then the adaptive group GLM Lasso estimate* $\widehat{\theta}$ *satisfies the bounds*

$$\left\|\widehat{\theta} - \theta^*\right\|_2^2 \le \frac{36\lambda_n^2}{\kappa^2}\left\{\sqrt{|S_{elt}|} + \omega\sqrt{|S_{\mathcal{G}}|}\right\}^2 \tag{9.77}$$

with probability at least $1 - 3e^{-8n\delta^2}$.

Proof

The proof is similar to Theorem 9.19. Recall the function $\mathcal{F}$ from equation (9.31), and let $\widehat{\Delta} = \widehat{\theta} - \theta^*$. First we claim that for any vector of the form $\Delta = t\widehat{\Delta}$ for some $t \in [0, 1]$ satisfies the bounds

$$\Phi_\omega(\Delta) \leq 4\left\{\sqrt{|S_{\text{elt}}|} + \omega\sqrt{|S_G|}\right\}\|\Delta\|_2, \tag{9.78a}$$

$$\mathcal{F}(\Delta) \geq \frac{\kappa}{2}\|\Delta\|_2^2 - c_1\frac{\log d}{n}\Phi_\omega^2(\Delta) - \frac{3\lambda_n}{2}\left\{\sqrt{|S_{\text{elt}}|} + \omega\sqrt{|S_G|}\right\}\|\Delta\|_2. \tag{9.78b}$$

Substituting the bound (9.78a) into inequality (9.78b) and rearranging yields

$$\mathcal{F}(\Delta) \geq \|\Delta\|_2\left\{\kappa'\|\Delta\|_2 - \frac{3\lambda_n}{2}\left(\sqrt{|S_{\text{elt}}|} + \omega\sqrt{|S_G|}\right)\right\}$$

where $\kappa' := \frac{\kappa}{2} - 16c_1 \frac{\log d}{n}\left(\sqrt{|S_{\text{elt}}|} + \omega\sqrt{|S_{\mathcal{G}}|}\right)^2$. Under the stated bound on the sample size n, we have $\kappa' \geq \frac{\kappa}{4}$, so that $\mathcal{F}$ is non-negative whenever

$$\|\Delta\|_2 \geq \frac{6\lambda_n}{\kappa}\left(\sqrt{|S_{\text{elt}}|} + \omega\sqrt{|S_G|}\right).$$

Finally, following through the remainder of the proof of Theorem 9.19 yields the claimed bound (9.77).
Let us now return to prove the bounds (9.78a) and (9.78b). To begin, a straightforward calculation shows that the dual norm is given by

$$\Phi^*_\omega(v) = \max\left\{\|v\|_\infty, \frac{1}{\omega}\max_{g\in\mathcal{G}}\left\|v_g\right\|_2\right\}.$$

Consequently, the event $\mathbb{G}(\lambda_n) := \left\{\Phi^*_{\omega}(\nabla\mathcal{L}_n(\theta^*)) \le \frac{\lambda_n}{2}\right\}$ is equivalent to

$$\left\|\nabla\mathcal{L}_n(\theta^*)\right\|_\infty \le \frac{\lambda_n}{2} \quad \text{and} \quad \max_{g\in\mathcal{G}}\left\|(\nabla\mathcal{L}_n(\theta^*))_g\right\|_2 \le \frac{\lambda_n\omega}{2}. \qquad (9.79)$$

We assume that these conditions hold for the moment, returning to verify them at the end of the proof.
Define $\Delta = t\widehat{\Delta}$ for some $t \in [0,1]$. Fix some decomposition $\theta^* = \alpha^* + \beta^*$, where α^* is S_{elt} -sparse and β^* is $S_{\mathcal{G}}$ -group-sparse, and note that

$$\Phi_\omega(\theta^*) \le \left\|\alpha^*\right\|_1 + \omega\left\|\beta^*\right\|_{\mathcal{G},2}.$$

Similarly, let us write $\Delta = \Delta_\alpha + \Delta_\beta$ for some pair such that

$$\Phi_\omega(\theta^* + \Delta) \le \left\|\alpha^* + \Delta_\alpha\right\|_1 + \omega\left\|\beta^* + \Delta_\beta\right\|_{\mathcal{G},2}.$$

Proof of (9.78a): Recalling that

$$\mathcal{F}(\Delta) := \mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*) + \lambda_n\{\Phi_\omega(\theta^* + \Delta) - \Phi_\omega(\theta^*)\}.$$

Consider a vector of the form $\Delta = t\widehat{\Delta}$ for some scalar $t \in [0, 1]$. Noting that $\mathcal{F}$ is convex and minimized at $\widehat{\Delta}$, we have

$$\mathcal{F}(\Delta) = \mathcal{F}(t\widehat{\Delta} + (1-t)0) \leq t\mathcal{F}(\widehat{\Delta}) + (1-t)\mathcal{F}(0) \leq \mathcal{F}(0) = 0.$$

then we have

$$
\begin{aligned}
\mathcal{E}_n(\Delta) &= \mathcal{L}_n\left(\theta^*+\Delta\right)-\mathcal{L}_n\left(\theta^*\right)-\left\langle\nabla \mathcal{L}_n\left(\theta^*\right), \Delta\right\rangle \\
&\leq \lambda_n\left\{\Phi_\omega\left(\theta^*\right)-\Phi_\omega\left(\theta^*+\Delta\right)\right\}+\left|\left\langle\nabla \mathcal{L}_n\left(\theta^*\right), \Delta\right\rangle\right| \\
&\leq \lambda_n\left\{\left\|\alpha^*\right\|_1-\left\|\alpha^*+\Delta_\alpha\right\|_1+\omega\left(\left\|\beta^*\right\|_{\mathcal{G}, 2}-\left\|\beta^*+\Delta_\beta\right\|_{\mathcal{G}, 2}\right)\right\} \\
&\quad +\left|\left\langle\nabla \mathcal{L}_n\left(\theta^*\right), \Delta\right\rangle\right| \\
&\leq \lambda_n\left\{\left(\left\|\left(\Delta_\alpha\right)_{S_{\text{elt}}}\right\|_1-\left\|\left(\Delta_\alpha\right)_{S_{\text{elt}}^c}\right\|_1\right)+\omega\left(\left\|\left(\Delta_\beta\right)_{S_{\mathcal{G}}}\right\|_{\mathcal{G}, 2}-\left\|\left(\Delta_\beta\right)_{S_{\mathcal{G}}^c}\right\|_{\mathcal{G}, 2}\right)\right\} \\
&\quad +\left|\left\langle\nabla \mathcal{L}_n\left(\theta^*\right), \Delta\right\rangle\right| \\
&\leq \frac{\lambda_n}{2}\left\{\left(3\left\|\left(\Delta_\alpha\right)_{S_{\text{elt}}}\right\|_1-\left\|\left(\Delta_\alpha\right)_{S_{\text{elt}}^c}\right\|_1\right)\right. \\
&\quad \left.+\omega\left(3\left\|\left(\Delta_\beta\right)_{S_{\mathcal{G}}}\right\|_{\mathcal{G}, 2}-\left\|\left(\Delta_\beta\right)_{S_{\mathcal{G}}^c}\right\|_{\mathcal{G}, 2}\right)\right\} .
\end{aligned}
$$

Since $\mathcal{E}_n(\Delta) \geq 0$ by convexity, rearranging yields

$$\begin{aligned}
\|\Delta_\alpha\|_1 + \omega \left\|\Delta_\beta\right\|_{G,2} &\leq 4\left\{\left\|(\Delta_\alpha)_{S_{\mathrm{elt}}}\right\|_1 + \omega\left\|\left(\Delta_\beta\right)_{S_{\mathcal{G}}}\right\|_{G,2}\right\} \\
&\leq 4\left\{\sqrt{|S_{\mathrm{elt}}|}\left\|(\Delta_\alpha)_{S_{\mathrm{elt}}}\right\|_2 + \omega\sqrt{|S_{\mathcal{G}}|}\left\|\left(\Delta_\beta\right)_{S_{\mathcal{G}}}\right\|_2\right\} \\
&\leq 4\left\{\sqrt{|S_{\mathrm{elt}}|} + \omega\sqrt{|S_{\mathcal{G}}|}\right\}\left\{\left\|(\Delta_\alpha)_{S_{\mathrm{elt}}}\right\|_2 + \left\|\left(\Delta_\beta\right)_{S_{\mathcal{G}}}\right\|_2\right\},
\end{aligned}$$

The overall vector Δ has the decomposition $\Delta = (\Delta_\alpha)_{S_{\mathrm{elt}}} + \left(\Delta_\beta\right)_{S_{\mathcal{G}}} + \Delta_T$, where T is the complement of the indices in S_{elt} *and* $S_{\mathcal{G}}$. Noting that all three sets are disjoint by construction, we have

$$\left\|(\Delta_\alpha)_{S_{\mathrm{elt}}}\right\|_2 + \left\|\left(\Delta_\beta\right)_{S_{\mathcal{G}}}\right\|_2 = \left\|(\Delta_\alpha)_{S_{\mathrm{elt}}} + \left(\Delta_\beta\right)_{S_{\mathcal{G}}}\right\|_2 \leq \|\Delta\|_2.$$

Proof of inequality (9.78b): From the proof of Theorem 9.19, recall the lower bound (9.50). This inequality, combined with the RSC condition, guarantees that the function value $\mathcal{F}(\Delta)$ is at least

$$\frac{\kappa}{2}\|\Delta\|_2^2 - c_1 \frac{\log d}{n} \Phi_\omega^2(\Delta) - \left|\langle \nabla \mathcal{L}_n(\theta^*), \Delta \rangle\right| \\ + \lambda_n \left\{ \left\|\alpha^* + \Delta_\alpha\right\|_1 - \left\|\alpha^*\right\|_1 \right\} + \lambda_n \omega \left\{ \left\|\beta^* + \Delta_\beta\right\|_{\mathcal{G},2} - \left\|\beta^*\right\|_{\mathcal{G},2} \right\}.$$

Verifying inequalities (9.79): From the proof of Corollary 9.26, we have

$$\mathbb{P}\left[\left\|\nabla \mathcal{L}_n(\theta^*)\right\|_\infty \geq t\right] \leq d e^{-\frac{n^2}{2B^2C^2}}.$$

Similarly, from the proof of Corollary 9.28 , we have

$$\mathbb{P}\left[\frac{1}{\omega} \max_{g \in \mathcal{G}} \left\|(\nabla \mathcal{L}_n(\theta^*))_g\right\|_2 \geq 2t\right] \leq 2 \exp\left(-\frac{n\omega^2 t^2}{2B^2C^2} + m \log 5 + \log |\mathcal{G}|\right).$$

Setting $t = 4BC\left\{\sqrt{\frac{\log d}{n}} + \delta\right\}$ and performing some algebra yields the claimed lower bound $\mathbb{P}\left[\mathbb{G}\left(\lambda_n\right)\right] \geq 1 - 3e^{-8n\sigma^2}$.

We say that $\mathcal{L}$ is locally L-Lipschitz over the ball $\mathbb{B}_2(R)$ if for each sample $Z = (x, y)$

$$|\mathcal{L}(\theta; Z) - \mathcal{L}(\widetilde{\theta}; Z)| \leq L|\langle\theta, x\rangle - \langle\widetilde{\theta}, x\rangle| \quad \text{for all } \theta, \widetilde{\theta} \in \mathbb{B}_2(R) \tag{9.85}$$

Letting $\{\varepsilon_i\}_{i=1}^n$ be an i.i.d. sequence of Rademacher variables, we define the symmetrized random vector $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n \varepsilon_i x_i$, and the random variable

$$\Phi^*(\bar{x}_n) := \sup_{\Phi(\theta)\leq 1} \left\langle \theta, \frac{1}{n}\sum_{i=1}^n \varepsilon_i x_i \right\rangle. \tag{9.88}$$

When $x_i \sim \mathcal{N}(0, \mathbf{I}_d)$, the mean $\mathbb{E}[\Phi^*(\bar{x}_n)]$ is proportional to the Gaussian complexity of the unit ball $\left\{\theta \in \mathbb{R}^d \mid \Phi(\theta) \leq 1\right\}$.

The following theorem applies to any norm Φ that satisfies $\Phi(\Delta) \geq \|\Delta\|_2$ uniformly. Let $M_n(\Phi; R) := 4 \log\left(\frac{R_u}{R_\ell}\right) \log \sup_{\theta \neq 0}\left(\frac{\Phi(\theta)}{\|\theta\|_2}\right)$ and for a pair of radii $0 < R_\ell < R_u$, let

$$\mathbb{B}_2\left(R_\ell, R_u\right) := \left\{\Delta \in \mathbb{R}^d \mid R_\ell \leq \|\Delta\|_2 \leq R_u\right\}. \tag{9.89}$$

Theorem (9.34)

Suppose that the cost function $\mathcal{L}$ is locally L-Lipschitz (9.85), and the population cost $\overline{\mathcal{L}}$ is locally κ-strongly convex (9.82) over the ball $\mathbb{B}_2\left(R_u\right)$. Then for any $\delta > 0$, the first-order Taylor error $\mathcal{E}_n$ satisfies

$$\left|\mathcal{E}_n(\Delta) - \overline{\mathcal{E}}(\Delta)\right| \leq 16 L \Phi(\Delta) \delta \quad \text{for all } \Delta \in \mathbb{B}_2\left(R_\ell, R_u\right) \tag{9.90}$$

with probability at least $1 - M_n(\Phi; R) \inf_{\lambda > 0} \mathbb{E}\left[e^{\lambda(\Phi^*(\bar{x}_n) - \delta)}\right]$.

Proof

Claim: $\mathcal{E}$ is a 2L-Lipschitz function in $\langle \Delta, x_i \rangle$.

We let $\frac{\partial \mathcal{L}}{\partial u}$ denote the derivative of $\mathcal{L}$ with respect to $u = \langle \theta, x \rangle$. For any sample $z_i \in \mathcal{Z}$ and parameters $\Delta, \widetilde{\Delta} \in \mathbb{R}^d$, we have

$$\left|\left\langle \nabla \mathcal{L}(\theta^*; Z_i), \Delta - \widetilde{\Delta} \right\rangle\right| \leq \left|\frac{\partial \mathcal{L}}{\partial u}(\theta^*; Z_i)\right| \left|\langle \Delta, x_i \rangle - \left\langle \widetilde{\Delta}, x_i \right\rangle\right| \\ \leq L \left|\langle \Delta, x_i \rangle - \left\langle \widetilde{\Delta}, x_i \right\rangle\right|. \tag{9.91}$$

Putting together the pieces, for any pair $\Delta, \widetilde{\Delta}$, we have

$$\begin{aligned} &\left|\mathcal{E}(\Delta; Z_i) - \mathcal{E}\left(\widetilde{\Delta}; Z_i\right)\right| \\ &\leq \left|\mathcal{L}(\theta^* + \Delta; Z_i) - \mathcal{L}\left(\theta^* + \widetilde{\Delta}; Z_i\right)\right| + \left|\left\langle \nabla \mathcal{L}(\theta^*; Z_i), \Delta - \widetilde{\Delta} \right\rangle\right| \\ &\leq 2L \left|\langle \Delta, x_i \rangle - \left\langle \widetilde{\Delta}, x_i \right\rangle\right|, \end{aligned} \tag{9.92}$$

For $r_1, r_2 > 0$, let $\mathbb{C}(r_1, r_2) := \mathbb{B}_2(r_2) \cap \{\Phi(\Delta) \le r_1 \|\Delta\|_2\}$ and the random variable $A_n(r_1, r_2) := \frac{1}{4r_1 r_2 L} \sup_{\Delta \in \mathbb{C}(r_1, r_2)} \left| \mathcal{E}_n(\Delta) - \overline{\mathcal{E}}(\Delta) \right|$. Our goal is to control the probability of the event $\{A_n \ge \delta\}$.

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda A_n}\right] &\le \mathbb{E}_{Z,\varepsilon}\left[\exp\left(2\lambda \sup_{\Delta \in \mathbb{C}(r_1, r_2)} \left| \frac{1}{4L r_1 r_2} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \mathcal{E}(\Delta; Z_i) \right| \right)\right] \\
&\le \mathbb{E}\left[\exp\left(\frac{\lambda}{r_1 r_2} \sup_{\Delta \in \mathbb{C}(r_1, r_2)} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \langle \Delta, x_i \rangle \right| \right)\right] \\
&\le \mathbb{E}\left[\exp\left\{\lambda \Phi^*\left(\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i x_i\right)\right\}\right].
\end{aligned}$$

By Markov's inequality,

$$\mathbb{P}\left[A_n(r_1, r_2) \ge \delta\right] \le \inf_{\lambda > 0} \mathbb{E}\left[\exp\left(\lambda\left\{\Phi^*\left(\bar{x}_n\right) - \delta\right\}\right)\right].$$

Let $\mathcal{E}$ be the event that the bound (9.90) is violated. For $k, \ell > 0$, $\mathbb{S}_{k,\ell} := \left\{ \Delta \in \mathbb{R}^d \mid 2^{k-1} \leq \frac{\Phi(\Delta)}{\|\Delta\|_2} \leq 2^k \text{ and } 2^{\ell-1} R_\ell \leq \|\Delta\|_2 \leq 2^\ell R_\ell \right\}$. By construction, any vector that can possibly violate the bound (9.90) is contained in the union $\bigcup_{k=1}^{N_1} \bigcup_{\ell=1}^{N_2} \mathbb{S}_{k,\ell}$, where $N_1 := \left\lceil \log \sup_{\theta \neq 0} \frac{\Phi(\theta)}{\|\theta\|} \right\rceil$ and $N_2 := \left\lceil \log \frac{R_u}{R_\ell} \right\rceil$. Suppose that the bound (9.90) is violated by some $\widehat{\Delta} \in \mathbb{S}_{k,\ell}$. In this case, we have

$$\left| \mathcal{E}_n(\widehat{\Delta}) - \overline{\mathcal{E}}(\widehat{\Delta}) \right| \geq 16L \frac{\Phi(\widehat{\Delta})}{\|\widehat{\Delta}\|_2} \|\widehat{\Delta}\|_2 \delta \geq 16L 2^{k-1} 2^{\ell-1} R_\ell \delta = 4L 2^k 2^\ell R_\ell \delta,$$

which implies that $A_n\left(2^k, 2^\ell R_\ell\right) \geq \delta$. Consequently, we have shown that

$$\mathbb{P}[\mathcal{E}] \leq \sum_{k=1}^{N_1} \sum_{\ell=1}^{N_2} \mathbb{P}\left[A_n\left(2^k, 2^\ell R_\ell\right) \geq \delta\right] \leq N_1 N_2 \inf_{\lambda > 0} \mathbb{E}\left[e^{\lambda(\Phi^*(\bar{x}_n) - \delta)}\right].$$

Let $\{\varepsilon_i\}_{i=1}^n$ denote an i.i.d. sequence of Rademacher variables,

$$\mu_n(\Phi^*) := \mathbb{E}_{x,\varepsilon}\left[\Phi^*\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i x_i\right)\right] = \mathbb{E}\left[\sup_{\Phi(\Delta)\le 1} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle \Delta, x_i\rangle\right].$$

This is simply the Rademacher complexity of the linear function class $x \mapsto \langle \Delta, x\rangle$ as Δ ranges over the unit ball of the norm Φ . Our theory applies to covariates $\{x_i\}_{i=1}^n$ drawn i.i.d. from a zero-mean distribution such that, for some positive constants (α, β), we have $\mathbb{E}\left[\langle \Delta, x\rangle^2\right] \ge \alpha$ and $\mathbb{E}\left[\langle \Delta, x\rangle^4\right] \le \beta$ for all vectors $\Delta \in \mathbb{R}^d$ with $\|\Delta\|_2 = 1$.

Theorem (9.36)

Consider any generalized linear model with covariates drawn from a zero-mean distribution satisfying the condition (9.94). Then the Taylor-series error $\mathcal{E}_n$ in the log-likelihood is lower bounded as

$$\mathcal{E}_n(\Delta) \geq \frac{\kappa}{2}\|\Delta\|_2^2 - c_0\mu_n^2\left(\Phi^*\right)\Phi^2(\Delta) \quad \text{for all } \Delta \in \mathbb{R}^d \text{ with } \|\Delta\|_2 \leq 1 \tag{9.95}$$

with probability at least $1 - c_1 e^{-c_2 n}$.

Proof

For some $t \in [0,1]$. Fix some vector $\Delta \in \mathbb{R}^d$ with $\|\Delta\|_2 = \delta \in (0,1]$, and set $\tau = K\delta$ for a constant $K > 0$ to be chosen. $\varphi_\tau(u) = u^2 \mathbb{I}[|u| \le 2\tau]$. $\gamma := \min_{|u| \le T+2K} \psi''(u)$.

$$
\begin{aligned}
\mathcal{E}_n(\Delta) &= \frac{1}{n}\sum_{i=1}^{n} \psi''\left(\langle \theta^*, x_i\rangle + t\langle \Delta, x_i\rangle\right)\langle \Delta, x_i\rangle^2 \\
&\ge \frac{1}{n}\sum_{i=1}^{n} \psi''\left(\langle \theta^*, x_i\rangle + t\langle \Delta, x_i\rangle\right)\varphi_\tau\left(\langle \Delta, x_i\rangle\right)\mathbb{I}\left[\left|\langle \theta^*, x_i\rangle\right| \le T\right] \\
&\ge \gamma\frac{1}{n}\sum_{i=1}^{n} \varphi_\tau\left(\langle \Delta, x_i\rangle\right)\mathbb{I}\left[\left|\langle \theta^*, x_i\rangle\right| \le T\right].
\end{aligned}
$$

It suffices to show that

$$\frac{1}{n}\sum_{i=1}^{n}\varphi_{\tau(\delta)}\left(\langle\Delta, x_i\rangle\right)\mathbb{I}\left[\left|\langle\theta^*, x_i\rangle\right| \le T\right] \ge c_3\delta^2 - c_4\mu_n\left(\Phi^*\right)\Phi(\Delta)\delta.$$

$$\widetilde{\varphi}_\tau(u) := u^2\mathbb{I}[|u| \le \tau] + (u-2\tau)^2\mathbb{I}[\tau < u \le 2\tau] + (u+2\tau)^2\mathbb{I}[-2\tau \le u < -\tau].$$

Note that it is Lipschitz with parameter 2τ and it lower bounds φ_τ, it suffices to show that for all unit-norm vectors Δ , we have

$$\frac{1}{n}\sum_{i=1}^{n}\widetilde{\varphi}_{\tau}\left(\langle\Delta, x_i\rangle\right)\mathbb{I}\left[\left|\langle\theta^*, x_i\rangle\right| \le T\right] \ge c_3 - c_4\mu_n\left(\Phi^*\right)\Phi(\Delta).$$

For a given radius $r \geq 1$, define the random variable

$$Z_n(r) := \sup_{\substack{\|\Delta\|_2=1 \\ \Phi(\Delta)\leq r}} \left| \frac{1}{n} \sum_{i=1}^{n} \widetilde{\varphi}_\tau \left(\langle \Delta, x_i \rangle\right) \mathbb{I}\left[\left|\langle \theta^*, x_i \rangle\right| \leq T\right] - \mathbb{E}\left[\widetilde{\varphi}_\tau(\langle \Delta, x \rangle)\mathbb{I}\left[\left|\langle \theta^*, x \rangle\right| \leq T\right]\right] \right| .$$

Suppose that we can prove that

$$\mathbb{E}\left[\widetilde{\varphi}_\tau(\langle \Delta, x \rangle)[\left[\left|\langle \theta^*, x \rangle\right| \leq T\right]\right] \geq \frac{3}{4}\alpha, \tag{9.99a}$$

and

$$\mathbb{P}\left[Z_n(r) > \alpha/2 + c_4 r \mu_n\left(\Phi^*\right)\right] \leq \exp\left(-c_2 \frac{n r^2 \mu_n^2\left(\Phi^*\right)}{\sigma^2} - c_2 n\right). \tag{9.99b}$$