

A Preliminaries

A.1 Contraction Lemmas

Lemma 1 (e.g. [25]). Let \mathcal{F} be any scalar-valued function class and ϕ_1, \dots, ϕ_n be any sequence of functions where $\phi_t : \mathbb{R} \rightarrow \mathbb{R}$ is L -Lipschitz. Then

$$\mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t \phi_t(f(x_t)) \leq L \cdot \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t f(x_t). \quad (9)$$

The following is a weighted generalization of the vector-valued Lipschitz contraction inequality.

Lemma 2. Let $\mathcal{F} \subseteq (\mathcal{X} \rightarrow \mathbb{R}^K)$, and let $h_t : \mathbb{R}^K \rightarrow \mathbb{R}$ be a sequence of functions for $t \in [n]$. Suppose each h_t is 1-Lipschitz with respect to $\|z\|_{A_t} := \sqrt{\langle z, A_t z \rangle}$, where $A_t \in \mathbb{R}^{K \times K}$ is positive semidefinite. Then

$$\mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t h_t(f(x_t)) \leq \sqrt{2} \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \sum_{t=1}^n \langle \epsilon_t, A_t^{1/2} f(x_t) \rangle. \quad (10)$$

Proof sketch for Lemma 2. Same proof as Theorem 3 in [30], with the additional observation that $\|z\|_{A_t} = \|A_t^{1/2} z\|_2$. \square

Lemma 3. Let \mathcal{G} be a class of vector-valued functions whose output space forms M blocks of vectors, i.e. each $g \in \mathcal{G}$ has the form $g : \mathcal{Z} \rightarrow \mathbb{R}^{d_1+d_2+\dots+d_M}$, where $g(z)_i \in \mathbb{R}^{d_i}$ denotes the i th block. Let $h_t : \mathbb{R}^{d_1+d_2+\dots+d_M} \rightarrow \mathbb{R}$, be a sequence of functions for $t \in [n]$ that satisfy the following block-wise Lipschitz property: For any assignment a_1, \dots, a_M with each $a_i \in \mathbb{R}^{d_i}$, $h_t(a_1, \dots, a_M)$ is L_i -Lipschitz with respect to a_i in the ℓ_2 norm. Then

$$\mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \sum_{t=1}^n h_t(g_1(z_t), \dots, g_M(z_t)) \leq \sqrt{2M} \sum_{i=1}^M L_i \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \sum_{t=1}^n \langle \epsilon_t, g_i(z_t) \rangle.$$

Proof. Immediate consequence of Lemma 2, along with sub-additivity of the supremum. \square

A.2 Bound for Vector-Valued Random Variables

Definition 5. For any vector space \mathcal{V} , a convex function $\Psi : \mathcal{V} \rightarrow \mathbb{R}$ is β -smooth with respect to a norm $\|\cdot\|$ if

$$\Psi(x) \leq \Psi(y) + \langle \nabla \Psi(y), x - y \rangle + \frac{\beta}{2} \|x - y\|^2 \quad \forall x, y \in \mathcal{V}.$$

A norm $\|\cdot\|$ is said to be β -smooth if the function $\Psi(x) = \frac{1}{2} \|x\|^2$ is β -smooth with respect to $\|\cdot\|$.

Theorem 7. Let $\|\cdot\|$ be any norm for which there exists Ψ such that $\Psi(x) \geq \frac{1}{2} \|x\|^2$, $\Psi(0) = 0$, and Ψ is β -smooth with respect to $\|\cdot\|$. Then

$$\mathbb{E}_\epsilon \left\| \sum_{t=1}^n \epsilon_t x_t \right\| \leq \sqrt{\beta \sum_{t=1}^n \|x_t\|^2}.$$

The reader may consult [34] for a high-probability version of this theorem.

Fact 1. The following spaces and norms satisfy the preconditions of Theorem 7:

- (\mathbb{R}^d, ℓ_p) for $p \geq 2$, with $\beta = p - 1$ [19].
- $(\mathbb{R}^d, \ell_\infty)$, with $\beta = O(\log d)$ [19].
- $(\mathbb{R}^{d_1 \times d_2}, \|\cdot\|_\sigma)$, with $\beta = O(\log(d_1 + d_2))$ [22].

Proof of Theorem 7. Using Jensen's inequality and the upper bound property of Ψ we have

$$\mathbb{E}_\epsilon \left\| \sum_{t=1}^n \epsilon_t x_t \right\| \leq \sqrt{\mathbb{E}_\epsilon \left\| \sum_{t=1}^n \epsilon_t x_t \right\|^2} \leq \sqrt{2} \cdot \sqrt{\mathbb{E}_\epsilon \Psi \left(\sum_{t=1}^n \epsilon_t x_t \right)}.$$

Applying the smoothness property at time n , and using that ϵ_n is independent of $\epsilon_1, \dots, \epsilon_{n-1}$:

$$\sqrt{\mathbb{E}_\epsilon \Psi \left(\sum_{t=1}^n \epsilon_t x_t \right)} \leq \sqrt{\mathbb{E}_\epsilon \left[\Psi \left(\sum_{t=1}^{n-1} \epsilon_t x_t \right) + \left\langle \Psi \left(\sum_{t=1}^{n-1} \epsilon_t x_t \right), \epsilon_n x_n \right\rangle + \frac{\beta}{2} \|x_n\|^2 \right]} = \sqrt{\mathbb{E}_\epsilon \Psi \left(\sum_{t=1}^{n-1} \epsilon_t x_t \right) + \frac{\beta}{2} \|x_n\|^2}.$$

The result follows by repeating this argument from time $t = n - 1$ to $t = 1$. \square

B Proofs from Section 2

Theorem 8 ([5], Theorem A.2/Lemma A.5). *Let $\mathcal{F} \subseteq (Z \rightarrow \mathbb{R})$ be a class of functions. Let $Z_1, \dots, Z_n \sim \mathcal{D}$ i.i.d. for some distribution \mathcal{D} . Then with probability at least $1 - \delta$ over the draw of $Z_{1:n}$,*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \mathbb{E}_Z f(Z) - \frac{1}{n} \sum_{t=1}^n f(Z_t) \right| \leq 4 \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n \epsilon_t f(Z_t) + 4 \sup_{f \in \mathcal{F}} \sup_{z \in Z} |f(z)| \cdot \frac{\log\left(\frac{2}{\delta}\right)}{n}. \quad (11)$$

Lemma 4 (Uniform convergence for vector-valued functions). *Let $\mathcal{G} \subseteq \{g : Z \rightarrow \mathfrak{B}\}$ for arbitrary set Z and vector space \mathfrak{B} . Let $Z_1, \dots, Z_n \sim \mathcal{D}$ i.i.d. for some distribution \mathcal{D} . Let a norm $\|\cdot\|$ over \mathfrak{B} be fixed. Then with probability at least $1 - \delta$ over the draw of $Z_{1:n}$,*

$$\mathbb{E} \sup_{g \in \mathcal{G}} \left\| \mathbb{E}_Z g(Z) - \frac{1}{n} \sum_{t=1}^n g(Z_t) \right\| \leq 4 \mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \left\| \frac{1}{n} \sum_{t=1}^n \epsilon_t g(Z_t) \right\| + 4 \sup_{g \in \mathcal{G}} \sup_{Z \in Z} \|g(Z)\| \cdot \frac{\log\left(\frac{2}{\delta}\right)}{n} \quad (12)$$

for some absolute constant $c > 0$.

Proof of Lemma 4. This follows immediately by applying [Theorem 8](#) to the expanded function class $\mathcal{F} := \{Z \mapsto \langle g(Z), v \rangle \mid g \in \mathcal{G}, \|v\|_* \leq 1\}$. \square

Proof of Proposition 1. This is a direct consequence of McDiarmid's inequality. Consider any vector-valued function class of functions \mathcal{G} . Let $Z_1, \dots, Z_n \sim \mathcal{D}$ i.i.d. for some distribution \mathcal{D} . Then McDiarmid's inequality implies that with probability at least $1 - \delta$ over the draw of $Z_{1:n}$,

$$\sup_{g \in \mathcal{G}} \left\| \mathbb{E}_Z g(Z) - \frac{1}{n} \sum_{t=1}^n g(Z_t) \right\| \leq \mathbb{E} \sup_{g \in \mathcal{G}} \left\| \mathbb{E}_Z g(Z) - \frac{1}{n} \sum_{t=1}^n g(Z_t) \right\| + c \cdot \sup_{g \in \mathcal{G}} \sup_{Z \in Z} \|g(Z)\| \cdot \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{n}}. \quad (13)$$

Proof of Proposition 2. This follows by applying the uniform convergence lemma, [Lemma 4](#), to the class $\mathcal{G} = \{(x, y) \mapsto \nabla \ell(w; x, y) \mid w \in \mathcal{W}\}$. \square

Proof of Theorem 1. We write

$$\mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \left\| \sum_{t=1}^n \epsilon_t \nabla (G_t(F_t(w))) \right\| = \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sup_{v \in \mathfrak{B}^*: \|v\|_* \leq 1} \sum_{t=1}^n \epsilon_t \langle \nabla (G_t(F_t(w))), v \rangle,$$

Using the chain rule for differentiation we have

$$\langle \nabla (G_t(F_t(w))), v \rangle = \langle (\nabla G_t)(F_t(w)), (\langle \nabla F_{t,k}(w), v \rangle)_{k \in [K]} \rangle.$$

We now introduce new functions that relabel the quantities in this expression. Let $h : \mathbb{R}^{2K} \rightarrow \mathbb{R}$ be given by $h(a, b) = \langle a, b \rangle$, let $f_1 : \mathcal{W} \rightarrow \mathbb{R}^K$ be given by $f_1(w) = (\nabla G_t)(F_t(w))$ and f_2 be given by $f_2(w, v) = (\langle \nabla F_{t,k}(w), v \rangle)_{k \in [K]}$. We apply the block-wise contraction lemma [Lemma 3](#) with one block for f_1 and one block for f_2 to conclude

$$\begin{aligned} & \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sup_{v \in \mathfrak{B}^*: \|v\|_* \leq 1} \sum_{t=1}^n \epsilon_t h(f_1(w), f_2(w, v)) \\ & \leq 2L_F \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sup_{v \in \mathfrak{B}^*: \|v\|_* \leq 1} \sum_{t=1}^n \langle \epsilon_t, f_1(w) \rangle + 2L_G \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sup_{v \in \mathfrak{B}^*: \|v\|_* \leq 1} \sum_{t=1}^n \langle \epsilon_t, f_2(w, v) \rangle, \end{aligned}$$

which establishes the result after expanding terms. All that must be verified is that the assumptions on the norm bounds for ∇G_t and ∇F_t in the theorem statement ensure the Lipschitz requirement in the statement of [Lemma 3](#) is met. \square

C Proofs from Section 3

For all proofs in this section we adopt the notation $s := \|w^*\|_0$, and use $c > 0$ to denote an absolute constant whose precise value depends on context.

C.1 Generalized Linear Models

Proof of Theorem 3. To begin, we apply Proposition 1 and Proposition 2 to conclude that whenever (α, μ) -PL holds, with probability at least $1 - \delta$ over the examples $\{(x_t, y_t)\}_{t=1}^n$, any learning algorithm $\widehat{w}^{\text{alg}} \in \mathcal{W}$ satisfies

$$L_{\mathcal{D}}(\widehat{w}^{\text{alg}}) - L^* \leq c \cdot \mu \left(\|\nabla \widehat{L}_n(\widehat{w}^{\text{alg}})\|^\alpha + \left(\frac{\Re_{\|\cdot\|}(\nabla \ell \circ \mathcal{W}; x_{1:n}, y_{1:n})}{n} + 2C_\sigma R \sqrt{\frac{\log(1/\delta)}{n}} \right)^\alpha \right). \quad (14)$$

Here $c > 0$ is an absolute constant and we have used that $\|\nabla \ell(w; x_t, y_t)\| \leq 2C_\sigma R$.

Smooth high-dimensional setup For the general smooth norm pair setup in (14), Lemma 5 and Lemma 7 imply

$$\begin{aligned} L_{\mathcal{D}}(\widehat{w}^{\text{alg}}) - L^* &\leq c \cdot \frac{BC_\sigma}{c_\sigma} \left(\|\nabla \widehat{L}_n(\widehat{w}^{\text{alg}})\| + \left(BR^2 C_\sigma^2 \sqrt{\frac{\beta}{n}} + 2C_\sigma R \sqrt{\frac{\log(1/\delta)}{n}} \right) \right) \\ &= \mu_h \cdot \|\nabla \widehat{L}_n(\widehat{w}^{\text{alg}})\| + \frac{C_h}{\sqrt{n}}, \end{aligned}$$

where we recall $C_h = c \cdot \frac{B^2 R^2 C_\sigma^3 \sqrt{\beta} + 2C_\sigma^2 BR \sqrt{\log(1/\delta)}}{c_\sigma}$ and $\mu_h = c \cdot \frac{BC_\sigma}{c_\sigma}$.

Low-dimensional ℓ_2/ℓ_2 setup For the low-dimension ℓ_2/ℓ_2 pair setup in (14), Lemma 5 and Lemma 7 imply

$$\begin{aligned} L_{\mathcal{D}}(\widehat{w}^{\text{alg}}) - L^* &\leq c \cdot \frac{C_\sigma}{4c_\sigma^3 \lambda_{\min}(\Sigma)} \left(\|\nabla \widehat{L}_n(\widehat{w}^{\text{alg}})\|^2 + \left(BR^2 C_\sigma^2 \sqrt{\frac{1}{n}} + 2C_\sigma R \sqrt{\frac{\log(1/\delta)}{n}} \right)^2 \right) \\ &= \frac{\mu_1}{\lambda_{\min}(\Sigma)} \cdot \|\nabla \widehat{L}_n(\widehat{w}^{\text{alg}})\|^2 + \frac{C_1}{n \cdot \lambda_{\min}(\Sigma)}, \end{aligned}$$

where we have used that the ℓ_2 norm is 1-smooth in Lemma 7. Recall that $C_1 = c \cdot \frac{2C_\sigma^5 R^4 B^2 + 8C_\sigma^3 R^2 \log(1/\delta)}{4c_\sigma^3}$ and $\mu_1 = c \cdot \frac{C_\sigma}{4c_\sigma^3}$.

Sparse ℓ_∞/ℓ_1 setup For the sparse ℓ_∞/ℓ_1 pair setup in (14), Lemma 5 and Lemma 7 imply

$$\begin{aligned} L_{\mathcal{D}}(\widehat{w}^{\text{alg}}) - L^* &\leq c \cdot \frac{C_\sigma s}{c_\sigma^3 \psi_{\min}(\Sigma)} \left(\|\nabla \widehat{L}_n(\widehat{w}^{\text{alg}})\|^2 + \left(BR^2 C_\sigma^2 \sqrt{\frac{\log d}{n}} + 2C_\sigma R \sqrt{\frac{\log(1/\delta)}{n}} \right)^2 \right) \\ &= \frac{\mu_s \cdot s}{\psi_{\min}(\Sigma)} \cdot \|\nabla \widehat{L}_n(\widehat{w}^{\text{alg}})\|^2 + \frac{s}{n} \cdot \frac{C_s}{\psi_{\min}(\Sigma)}, \end{aligned}$$

where we have used that the ℓ_∞ norm has the smoothness property with $\beta = O(\log(d))$ in Lemma 7. Recall that $C_s = c \cdot \frac{2C_\sigma^5 R^4 B^2 \log(d) + 8C_\sigma^3 R^2 \log(1/\delta)}{c_\sigma^3}$ and $\mu_s = c \cdot \frac{C_\sigma}{c_\sigma^3}$.

□

Lemma 5 (GD condition for the GLM). Consider the generalized linear model setup of Section 3.

- When $\|\cdot\|/\|\cdot\|_*$ are any dual norm pair, we have $(1, \frac{BC_\sigma}{c_\sigma})$ -GD:

$$L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*) \leq \frac{BC_\sigma}{c_\sigma} \|\nabla L_{\mathcal{D}}(w)\| \quad \forall w \in \mathcal{W}. \quad (15)$$

- In the ℓ_2/ℓ_2 setup, we have $\left(2, \frac{C_\sigma}{4c_\sigma^3 \lambda_{\min}(\Sigma)}\right)$ -GD:

$$L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*) \leq \frac{C_\sigma}{4c_\sigma^3 \lambda_{\min}(\Sigma)} \|\nabla L_{\mathcal{D}}(w)\|_2^2 \quad \forall w \in \mathcal{W}. \quad (16)$$

- In the sparse ℓ_∞/ℓ_1 setup, where $\|w^*\|_0 \leq s$, we have $\left(2, \frac{C_\sigma s}{c_\sigma^3 \psi_{\min}(\Sigma)}\right)$ -GD:

$$L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*) \leq \frac{C_\sigma s}{c_\sigma^3 \psi_{\min}(\Sigma)} \|\nabla L_{\mathcal{D}}(w)\|_\infty^2 \quad \forall w \in \mathcal{W}. \quad (17)$$

Proof of Lemma 5.

Upper bound for excess risk. We first prove the following intermediate upper bound:

$$L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*) \leq \frac{C_\sigma}{2c_\sigma} \langle \nabla L_{\mathcal{D}}(w), w - w^* \rangle. \quad (18)$$

Letting $w \in \mathcal{W}$ be fixed, we have

$$\langle \nabla L_{\mathcal{D}}(w), w - w^* \rangle = 2 \mathbb{E}_{(x,y)} [(\sigma(\langle w, x \rangle) - y) \sigma'(\langle w, x \rangle) \langle w - w^*, x \rangle].$$

Using the well-specified assumption:

$$= 2 \mathbb{E}_x [(\sigma(\langle w, x \rangle) - \sigma(\langle w^*, x \rangle)) \sigma'(\langle w, x \rangle) \langle w - w^*, x \rangle].$$

We now consider the term inside the expectation. Since σ is increasing we have

$$\sigma(\langle w, x \rangle) - \sigma(\langle w^*, x \rangle) \sigma'(\langle w, x \rangle) \langle w - w^*, x \rangle = |\sigma(\langle w, x \rangle) - \sigma(\langle w^*, x \rangle)| \cdot |\langle w - w^*, x \rangle| \cdot \sigma'(\langle w, x \rangle)$$

point-wise. We apply two lower bounds. First, $\sigma'(\langle w, x \rangle) > c_\sigma$ by assumption. Second, Lipschitzness of σ implies

$$|\sigma(\langle w, x \rangle) - \sigma(\langle w^*, x \rangle)| \leq C_\sigma |\langle w - w^*, x \rangle|.$$

Combining these inequalities, we also obtain the following inequality in expectation over x :

$$\mathbb{E}_x (\sigma(\langle w, x \rangle) - \sigma(\langle w^*, x \rangle))^2 \leq \frac{C_\sigma}{2c_\sigma} \langle \nabla L_{\mathcal{D}}(w), w - w^* \rangle.$$

Lastly, since the model is well-specified we have

$$L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*) = \mathbb{E}_x (\sigma(\langle w, x \rangle) - \sigma(\langle w^*, x \rangle))^2,$$

by a standard argument:

$$\begin{aligned} L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*) &= \mathbb{E}_{x,y} [\sigma^2(\langle w, x \rangle) + y^2 - 2\sigma(\langle w, x \rangle)y - \sigma^2(\langle w^*, x \rangle) - y^2 + 2\sigma(\langle w^*, x \rangle)y] \\ &= \mathbb{E}_x [\sigma^2(\langle w, x \rangle) - 2\sigma(\langle w, x \rangle)\sigma(\langle w^*, x \rangle) + \sigma^2(\langle w^*, x \rangle)] \\ &= \mathbb{E}_x (\sigma(\langle w, x \rangle) - \sigma(\langle w^*, x \rangle))^2. \end{aligned}$$

Proving the GD conditions. With the inequality (18) established the various GD inequalities follow in quick succession.

- $\left(1, \frac{BC_\sigma}{c_\sigma}\right)$ -GD:

To prove this inequality, simply use Hölder's inequality to obtain the upper bound,

$$\langle \nabla L_{\mathcal{D}}(w), w - w^* \rangle \leq 2B \|\nabla L_{\mathcal{D}}(w)\|.$$

- $\left(2, \frac{C_\sigma}{4c_\sigma^3 \lambda_{\min}(\mathbb{E}[xx^\top])}\right)$ -GD:

Resuming from (18) we have

$$L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*) \leq \frac{C_\sigma}{2c_\sigma} \langle \nabla L_{\mathcal{D}}(w), w - w^* \rangle.$$

Let $P_{\mathcal{X}}$ denote the orthogonal projection onto $\text{span}(\mathbb{E}[xx^\top])$. Note that $\nabla \ell(w; x, y)$ is parallel to x , we can thus introduce the projection matrix $P_{\mathcal{X}}$ while preserving the inner product

$$= \frac{C_\sigma}{2c_\sigma} \langle P_{\mathcal{X}} \nabla L_{\mathcal{D}}(w), P_{\mathcal{X}}(w - w^*) \rangle$$

Applying Cauchy-Schwarz:

$$\leq \frac{C_\sigma}{2c_\sigma} \|\nabla L_{\mathcal{D}}(w)\|_2 \cdot \|P_{\mathcal{X}}(w - w^*)\|_2. \quad (19)$$

What remains is to relate the gradient norm to the term $\|P_{\mathcal{X}}(w - w^*)\|_2$. We proceed with another lower bound argument similar to the one used to establish (18),

$$\langle \nabla L_{\mathcal{D}}(w), w - w^* \rangle = 2 \mathbb{E}_{(x,y)} [(\sigma(\langle w, x \rangle) - y) \sigma'(\langle w, x \rangle) \langle w - w^*, x \rangle].$$

Using the well-specified assumption once more:

$$= 2 \mathbb{E}_x [(\sigma(\langle w, x \rangle) - \sigma(\langle w^*, x \rangle)) \sigma'(\langle w, x \rangle) \langle w - w^*, x \rangle].$$

Monotonicity of σ , implies the argument to the expectation is non-negative pointwise, so we have the lower bound,

$$\geq 2c_\sigma \mathbb{E}_x [(\sigma(\langle w, x \rangle) - \sigma(\langle w^*, x \rangle)) \langle w - w^*, x \rangle].$$

Consider a particular draw of x and assume $\langle w, x \rangle \geq \langle w^*, x \rangle$ without loss of generality. Using the mean value theorem, there is some $s \in [\langle w^*, x \rangle, \langle w, x \rangle]$ such that

$$(\sigma(\langle w, x \rangle) - \sigma(\langle w^*, x \rangle)) \langle w - w^*, x \rangle = \langle w - w^*, x \rangle^2 \sigma'(s) \geq \langle w - w^*, x \rangle^2 c_\sigma.$$

Grouping terms, we have shown

$$\begin{aligned} \langle P_{\mathcal{X}} \nabla L_{\mathcal{D}}(w), P_{\mathcal{X}}(w - w^*) \rangle &= \langle \nabla L_{\mathcal{D}}(w), w - w^* \rangle \geq 2c_\sigma^2 \mathbb{E} \langle w - w^*, x \rangle^2 \\ &= 2c_\sigma^2 \langle w - w^*, \mathbb{E}[xx^\top](w - w^*) \rangle \quad (20) \\ &\geq 2c_\sigma^2 \lambda_{\min}(\mathbb{E}[xx^\top]) \|P_{\mathcal{X}}(w - w^*)\|_2^2. \end{aligned}$$

In other words, by rearranging and applying Cauchy-Schwarz we have

$$\|P_{\mathcal{X}}(w - w^*)\|_2 \leq \frac{1}{2c_\sigma^2 \lambda_{\min}(\mathbb{E}[xx^\top])} \cdot \|\nabla L_{\mathcal{D}}(w)\|_2.$$

Combining this inequality with (19), we have

$$L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*) \leq \frac{C_\sigma}{4c_\sigma^3 \lambda_{\min}(\mathbb{E}[xx^\top])} \cdot \|\nabla L_{\mathcal{D}}(w)\|_2^2.$$

• $\left(2, \frac{C_\sigma s}{c_\sigma^3 \psi_{\min}(\mathbb{E}[xx^\top])}\right)$ -GD:

Using the inequality (20) from the preceeding GD proof, we have

$$\langle \nabla L_{\mathcal{D}}(w), w - w^* \rangle \geq 2c_\sigma^2 \langle w - w^*, \mathbb{E}[xx^\top](w - w^*) \rangle.$$

By the assumption that $\|w\|_1 \leq \|w^*\|_1$, we apply Lemma 6 to conclude that 1) $w - w^* \in \mathcal{C}(S(w^*), 1)$ and 2) $\|w - w^*\|_1 \leq 2\sqrt{s}\|w - w^*\|_2$. The first fact implies that

$$\langle w - w^*, \mathbb{E}[xx^\top](w - w^*) \rangle \geq \psi_{\min}(\mathbb{E}[xx^\top]) \|w - w^*\|_2^2.$$

Rearranging, we have

$$\begin{aligned} \|w - w^*\|_2 &\leq \frac{1}{2c_\sigma^2 \psi_{\min}(\mathbb{E}[xx^\top])} \frac{\langle \nabla L_{\mathcal{D}}(w), w - w^* \rangle}{\|w - w^*\|_2} \\ &\leq \frac{1}{2c_\sigma^2 \psi_{\min}(\mathbb{E}[xx^\top])} \frac{\|\nabla L_{\mathcal{D}}(w)\|_\infty \|w - w^*\|_1}{\|w - w^*\|_2} \\ &\leq \frac{\sqrt{s}}{c_\sigma^2 \psi_{\min}(\mathbb{E}[xx^\top])} \|\nabla L_{\mathcal{D}}(w)\|_\infty. \end{aligned}$$

On the other hand, from (18) we have

$$\begin{aligned}
L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*) &\leq \frac{C_{\sigma}}{2c_{\sigma}} \langle \nabla L_{\mathcal{D}}(w), w - w^* \rangle \\
&\leq \frac{C_{\sigma}}{2c_{\sigma}} \|\nabla L_{\mathcal{D}}(w)\|_{\infty} \|w - w^*\|_1 \\
&\leq \frac{C_{\sigma}\sqrt{s}}{c_{\sigma}} \|\nabla L_{\mathcal{D}}(w)\|_{\infty} \|w - w^*\|_2.
\end{aligned}$$

Combining this with the preceding inequality yields the result. \square

The following utility lemma is a standard result in high-dimensional statistics (e.g. [41]).

Lemma 6. Let $w, w^* \in \mathbb{R}^d$. If $\|w\|_1 \leq \|w^*\|_1$ then $w - w^* =: \nu \in \mathcal{C}(S(w^*), 1)$. Furthermore, $\|\nu\|_1 \leq 2\sqrt{|S(w^*)|} \|\nu\|_2$.

Proof of Lemma 6. Let $S := S(w^*)$. Then the constraint that $\|w\|_1 \leq \|w^*\|_1$ implies

$$\|w^*\|_1 \geq \|w\|_1 = \|w^* + \nu\|_1 = \|w^*\|_1 + \|\nu_S\|_1 + \|\nu_{S^c}\|_1 \geq \|w^*\|_1 - \|\nu_S\|_1 + \|\nu_{S^c}\|_1.$$

Rearranging, this implies $\|\nu_{S^c}\|_1 \leq \|\nu_S\|_1$, so the first result is established.

For the second result, $\nu \in \mathcal{C}(S, 1)$ implies $\|\nu\|_1 = \|\nu_S\|_1 + \|\nu_{S^c}\|_1 \leq 2\|\nu_S\|_1 \leq 2\sqrt{|S|} \|\nu_S\|_2 \leq 2\sqrt{|S|} \|\nu\|_2$. \square

Lemma 7. Let the norm $\|\cdot\|$ satisfy the smoothness property of Theorem 7 with constant β . Then the empirical loss gradient for the generalized linear model setting enjoys the normed Rademacher complexity bound,

$$\mathbb{E}_{\epsilon} \sup_{w \in \mathcal{W}} \left\| \sum_{t=1}^n \epsilon_t \nabla \ell(w; x_t, y_t) \right\| \leq O\left(BR^2 C_{\sigma}^2 \sqrt{\beta n}\right). \quad (21)$$

Proof of Lemma 7. Let $G_t(s) = (\sigma(s) - y_t)^2$ and $F_t(w) = \langle w, x_t \rangle$, so that $\ell(w; x_t, y_t) = G_t(F_t(w))$.

Observe that $G'_t(s) = 2(\sigma(s) - y_t)\sigma'(s)$ and $\nabla F_t(w) = x_t$, so our assumptions imply that $|G'_t(s)| \leq 2C_{\sigma}$ and $\|\nabla F_t(w)\| \leq R$. We can thus apply Theorem 1 to conclude

$$\mathbb{E}_{\epsilon} \sup_{w \in \mathcal{W}} \left\| \sum_{t=1}^n \epsilon_t \nabla \ell(w; x_t, y_t) \right\| \leq 2R \mathbb{E}_{\epsilon} \sup_{w \in \mathcal{W}} \sum_{t=1}^n \epsilon_t G'_t(\langle w, x_t \rangle) + 4C_{\sigma} \mathbb{E}_{\epsilon} \left\| \sum_{t=1}^n \epsilon_t x_t \right\|.$$

For the first term on the left-hand side, observe that for any s , $|G''_t(s)| \leq 2|\sigma''(s)| + 2|\sigma'(s)|^2 \leq 4C_{\sigma}^2$, so G'_t is $4C_{\sigma}^2$ -Lipschitz. The classical scalar Lipschitz contraction inequality for Rademacher complexity (Lemma 1) therefore implies

$$\mathbb{E}_{\epsilon} \sup_{w \in \mathcal{W}} \sum_{t=1}^n \epsilon_t G'_t(\langle w, x_t \rangle) \leq 4C_{\sigma}^2 \mathbb{E}_{\epsilon} \sup_{w \in \mathcal{W}} \sum_{t=1}^n \epsilon_t \langle w, x_t \rangle = 4C_{\sigma}^2 B \mathbb{E}_{\epsilon} \left\| \sum_{t=1}^n \epsilon_t x_t \right\|.$$

Finally, by our smoothness assumption on the norm, Theorem 7 implies

$$\mathbb{E}_{\epsilon} \left\| \sum_{t=1}^n \epsilon_t x_t \right\| \leq \sqrt{2\beta R^2 n}.$$

\square

C.2 Robust Regression

Proof of Theorem 4. This proof follows the same template as Theorem 3. We use Proposition 1 and Proposition 2 to conclude that whenever (α, μ) -PL holds, with probability at least $1 - \delta$ over the examples $\{(x_t, y_t)\}_{t=1}^n$, any learning algorithm \widehat{w}^{alg} satisfies

$$L_{\mathcal{D}}(\widehat{w}^{\text{alg}}) - L^* \leq c \cdot \mu \left(\left\| \nabla \widehat{L}_n(\widehat{w}^{\text{alg}}) \right\|^\alpha + \left(\frac{\mathfrak{R}_{\|\cdot\|}(\nabla \ell \circ \mathcal{W}; x_{1:n}, y_{1:n})}{n} + C_\rho R \sqrt{\frac{\log(1/\delta)}{n}} \right)^\alpha \right), \quad (22)$$

where $c > 0$ is an absolute constant and we have used that $\|\nabla \ell(w; x_t, y_t)\| \leq C_\rho R$ with probability 1.

Smooth high-dimensional setup For the general smooth norm pair setup in (22), Lemma 8 and Lemma 9 imply

$$\begin{aligned} L_{\mathcal{D}}(\widehat{w}^{\text{alg}}) - L^* &\leq c \cdot \frac{BC_\rho}{c_\rho} \left(\left\| \nabla \widehat{L}_n(\widehat{w}^{\text{alg}}) \right\| + \left(BR^2 C_\rho \sqrt{\frac{\beta}{n}} + C_\rho R \sqrt{\frac{\log(1/\delta)}{n}} \right) \right) \\ &= \mu_h \cdot \left\| \nabla \widehat{L}_n(\widehat{w}^{\text{alg}}) \right\| + \frac{C_h}{\sqrt{n}}. \end{aligned}$$

Where we recall $C_h = c \cdot \frac{B^2 R^2 C_\rho^2 \sqrt{\beta} + C_\rho^2 BR \sqrt{\log(1/\delta)}}{c_\rho}$ and $\mu_h = c \cdot \frac{BC_\rho}{c_\rho}$.

Low-dimensional ℓ_2/ℓ_2 setup For the low-dimension ℓ_2/ℓ_2 pair setup in (22), Lemma 8 and Lemma 9 imply

$$\begin{aligned} L_{\mathcal{D}}(\widehat{w}^{\text{alg}}) - L^* &\leq c \cdot \frac{C_\rho}{2c_\rho^2 \lambda_{\min}(\Sigma)} \left(\left\| \nabla \widehat{L}_n(\widehat{w}^{\text{alg}}) \right\|^2 + \left(BR^2 C_\rho \sqrt{\frac{1}{n}} + C_\rho R \sqrt{\frac{\log(1/\delta)}{n}} \right)^2 \right) \\ &= \frac{\mu_1}{\lambda_{\min}(\Sigma)} \cdot \left\| \nabla \widehat{L}_n(\widehat{w}^{\text{alg}}) \right\|^2 + \frac{C_1}{n \cdot \lambda_{\min}(\Sigma)}, \end{aligned}$$

where we have used that the ℓ_2 norm is 1-smooth in Lemma 7. Recall that $C_1 = c \cdot \frac{C_\rho^3 R^4 B^2 + C_\rho^3 R^2 \log(1/\delta)}{c_\rho^2}$ and $\mu_1 = c \cdot \frac{C_\rho}{2c_\rho^2}$.

Sparse ℓ_∞/ℓ_1 setup For the sparse ℓ_∞/ℓ_1 pair setup in (22), Lemma 8 and Lemma 9 imply

$$\begin{aligned} L_{\mathcal{D}}(\widehat{w}^{\text{alg}}) - L^* &\leq c \cdot \frac{2C_\rho s}{c_\rho^2 \psi_{\min}(\Sigma)} \left(\left\| \nabla \widehat{L}_n(\widehat{w}^{\text{alg}}) \right\|^2 + \left(BR^2 C_\rho \sqrt{\frac{\log d}{n}} + C_\rho R \sqrt{\frac{\log(1/\delta)}{n}} \right)^2 \right) \\ &= \frac{\mu_s \cdot s}{\psi_{\min}(\Sigma)} \cdot \left\| \nabla \widehat{L}_n(\widehat{w}^{\text{alg}}) \right\|^2 + \frac{s}{n} \cdot \frac{C_s}{\psi_{\min}(\Sigma)}, \end{aligned}$$

where we have used that the ℓ_∞ norm has the smoothness property with $\beta = O(\log(d))$ in Lemma 7. Recall that $C_s = c \cdot \frac{4C_\rho^3 R^4 B^2 \log(d) + 4C_\rho^3 R^2 \log(1/\delta)}{c_\rho^2}$ and $\mu_s = c \cdot \frac{2C_\rho}{c_\rho^2}$.

□

Lemma 8 (GD condition for robust regression). Consider the robust regression setup of Section 3.

- When $\|\cdot\|/\|\cdot\|_*$ are any dual norm pair, we have $\left(1, \frac{BC_\rho}{c_\rho}\right)$ -GD:

$$L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*) \leq \frac{BC_\rho}{c_\rho} \left\| \nabla L_{\mathcal{D}}(w) \right\| \quad \forall w \in \mathcal{W}. \quad (23)$$

- In the ℓ_2/ℓ_2 setup, we have $\left(2, \frac{C_\rho}{2c_\rho^2\lambda_{\min}(\Sigma)}\right)$ -GD:

$$L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*) \leq \frac{C_\rho}{2c_\rho^2\lambda_{\min}(\Sigma)} \|\nabla L_{\mathcal{D}}(w)\|_2^2 \quad \forall w \in \mathcal{W}. \quad (24)$$

- In the sparse ℓ_∞/ℓ_1 setup, where $\|w^*\|_0 \leq s$, we have $\left(2, \frac{2C_\rho s}{c_\rho^2\psi_{\min}(\Sigma)}\right)$ -GD:

$$L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*) \leq \frac{2C_\rho s}{c_\rho^2\psi_{\min}(\Sigma)} \|\nabla L_{\mathcal{D}}(w)\|_\infty^2 \quad \forall w \in \mathcal{W}. \quad (25)$$

Proof of Lemma 8.

Excess risk upper bound. To begin, smoothness of ρ implies that for any $s, s^* \in \mathcal{S}$ we have

$$\rho(s) - \rho(s^*) \leq \rho'(s^*)(s - s^*) + \frac{C_\rho}{2}(s - s^*)^2.$$

Since this holds point-wise, we use it to derive the following in-expectation bound

$$\begin{aligned} L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*) &\leq \mathbb{E}_{x,y}[\rho(\langle w^*, x \rangle - y)\langle w - w^*, x \rangle] + \frac{C_\rho}{2} \mathbb{E}\langle w - w^*, x \rangle^2 \\ &= \langle \nabla L_{\mathcal{D}}(w^*), w - w^* \rangle + \frac{C_\rho}{2} \mathbb{E}\langle w - w^*, x \rangle^2. \end{aligned}$$

Note however that

$$\nabla L_{\mathcal{D}}(w^*) = \mathbb{E}_{x,\zeta}[\rho'(-\zeta)x] = 0,$$

since ζ is conditionally symmetric and ρ' is odd. We therefore have

$$L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*) \leq \frac{C_\rho}{2} \mathbb{E}\langle w - w^*, x \rangle^2.$$

On the other hand, using the form of the gradient we have

$$\begin{aligned} \langle L_{\mathcal{D}}(w), w - w^* \rangle &= \mathbb{E}_x[\mathbb{E}_\zeta \rho'(\langle w - w^*, x \rangle - \zeta)\langle w - w^*, x \rangle] \\ &= \mathbb{E}_x[h(\langle w - w^*, x \rangle)\langle w - w^*, x \rangle]. \end{aligned}$$

To lower bound the term inside the expectation, consider a particular draw of x and assume $\langle w - w^*, x \rangle \geq 0$; this is admissible because h , like ρ' , is odd. Then we have

$$h(\langle w - w^*, x \rangle)\langle w - w^*, x \rangle = \frac{h(\langle w - w^*, x \rangle)}{\langle w - w^*, x \rangle} \langle w - w^*, x \rangle^2 \geq c_\rho \langle w - w^*, x \rangle^2,$$

where the last line follows because $h(0) = 0$ and $h'(0) > c_\rho$. Since this holds pointwise, we simply take the expectation to show that

$$\langle \nabla L_{\mathcal{D}}(w), w - w^* \rangle \geq c_\rho \mathbb{E}\langle w - w^*, x \rangle^2, \quad (26)$$

and consequently the excess risk is bounded by

$$L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*) \leq \frac{C_\rho}{2c_\rho} \langle \nabla L_{\mathcal{D}}(w), w - w^* \rangle. \quad (27)$$

Proving the GD conditions. We now use (27) to establish the GD condition variants.

- $\left(1, \frac{BC_\rho}{c_\rho}\right)$ -GD:

Use Hölder's inequality to obtain the upper bound,

$$\langle \nabla L_{\mathcal{D}}(w), w - w^* \rangle \leq 2B \|\nabla L_{\mathcal{D}}(w)\|.$$

- $\left(2, \frac{C_\rho}{2c_\rho^2\lambda_{\min}(\Sigma)}\right)$ -GD:

Begin with

$$L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*) \leq \frac{C_\rho}{2c_\rho} \langle \nabla L_{\mathcal{D}}(w), w - w^* \rangle.$$

Using the same reasoning as in [Lemma 5](#), this is upper bounded by

$$\leq \frac{C_\rho}{2c_\rho} \|\nabla L_{\mathcal{D}}(w)\|_2 \cdot \|P_{\mathcal{X}}(w - w^*)\|_2, \quad (28)$$

where $P_{\mathcal{X}}$ denotes the orthogonal projection onto $\text{span}(\Sigma)$.

Recalling [\(26\)](#), it also holds that

$$\begin{aligned} \langle P_{\mathcal{X}} \nabla L_{\mathcal{D}}(w), P_{\mathcal{X}}(w - w^*) \rangle &= \langle \nabla L_{\mathcal{D}}(w), w - w^* \rangle \geq c_\rho \mathbb{E}_x \langle w - w^*, x \rangle^2 \\ &= c_\rho \langle w - w^*, \mathbb{E}_x [xx^T] (w - w^*) \rangle \\ &= c_\rho \langle w - w^*, \Sigma(w - w^*) \rangle \\ &\geq c_\rho \lambda_{\min}(\Sigma) \|P_{\mathcal{X}}(w - w^*)\|_2^2. \end{aligned} \quad (29)$$

Rearranging and applying Cauchy-Schwarz, we have

$$\|P_{\mathcal{X}}(w - w^*)\|_2 \leq \frac{1}{c_\rho \lambda_{\min}(\Sigma)} \cdot \|\nabla L_{\mathcal{D}}(w)\|_2.$$

Combining this inequality with [\(28\)](#), we have

$$L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*) \leq \frac{C_\rho}{2c_\rho^2 \lambda_{\min}(\Sigma)} \cdot \|\nabla L_{\mathcal{D}}(w)\|_2^2.$$

- $\left(2, \frac{2C_\rho s}{c_\rho^2 \psi_{\min}(\Sigma)}\right)$ -GD:

Using the inequality [\(29\)](#) from the ℓ_2/ℓ_2 GD condition proof above

$$\langle \nabla L_{\mathcal{D}}(w), w - w^* \rangle \geq c_\rho \langle w - w^*, \Sigma(w - w^*) \rangle.$$

By the assumption that $\|w\|_1 \leq \|w^*\|_1$, we apply [Lemma 6](#) to conclude that 1) $w - w^* \in \mathcal{C}(S(w^*), 1)$ and 2) $\|w - w^*\|_1 \leq 2\sqrt{s}\|w - w^*\|_2$, and so

$$\langle w - w^*, \Sigma(w - w^*) \rangle \geq \psi_{\min}(\Sigma) \|w - w^*\|_2^2.$$

Rearranging, and applying the $\|w - w^*\|_1 \leq 2\sqrt{s}\|w - w^*\|_2$ inequality:

$$\begin{aligned} \|w - w^*\|_2 &\leq \frac{1}{c_\rho \psi_{\min}(\Sigma)} \frac{\langle \nabla L_{\mathcal{D}}(w), w - w^* \rangle}{\|w - w^*\|_2} \\ &\leq \frac{1}{c_\rho \psi_{\min}(\Sigma)} \frac{\|\nabla L_{\mathcal{D}}(w)\|_\infty \|w - w^*\|_1}{\|w - w^*\|_2} \\ &\leq \frac{2\sqrt{s}}{c_\rho \psi_{\min}(\Sigma)} \|\nabla L_{\mathcal{D}}(w)\|_\infty. \end{aligned}$$

Finally, from [\(27\)](#) we have

$$\begin{aligned} L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*) &\leq \frac{C_\rho}{2c_\rho} \langle \nabla L_{\mathcal{D}}(w), w - w^* \rangle \\ &\leq \frac{C_\rho}{2c_\rho} \|\nabla L_{\mathcal{D}}(w)\|_\infty \|w - w^*\|_1 \\ &\leq \frac{C_\rho \sqrt{s}}{c_\rho} \|\nabla L_{\mathcal{D}}(w)\|_\infty \|w - w^*\|_2. \end{aligned}$$

Combining the two inequalities gives the final result.

□

Lemma 9. Let the norm $\|\cdot\|$ satisfy the smoothness property (see [Theorem 7](#)) with constant β . Then the gradient for robust regression satisfies the following normed Rademacher complexity bound:

$$\mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \left\| \sum_{t=1}^n \epsilon_t \nabla \ell(w; x_t, y_t) \right\| \leq O\left(BR^2 C_\rho \sqrt{\beta n}\right). \quad (30)$$

Proof of Lemma 9. Let $G_t(s) = \rho(s - y_t)$ and $F_t(w) = \langle w, x_t \rangle$, so that $\ell(w; x_t, y_t) = G_t(F_t(w))$. Then $G'_t(s) = \rho'(s - y_t)$ and $\nabla F_t(w) = x_t$, so our assumptions imply that $|G'_t(s)| \leq C_\rho$ and $\|\nabla F_t(w)\| \leq R$. We apply [Theorem 1](#) to conclude

$$\mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \left\| \sum_{t=1}^n \epsilon_t \nabla \ell(w; x_t, y_t) \right\| \leq 2R \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sum_{t=1}^n \epsilon_t G'_t(\langle w, x_t \rangle) + 2C_\rho \mathbb{E}_\epsilon \left\| \sum_{t=1}^n \epsilon_t x_t \right\|.$$

For the first term on the left-hand side, we have that for any s , $|G''_t(s)| = 2|\rho''(s - y_t)| \leq 2C_\rho$, so G'_t is $2C_\rho$ -Lipschitz. Then the by scalar contraction for Rademacher complexity ([Lemma 1](#)),

$$\mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sum_{t=1}^n \epsilon_t G'_t(\langle w, x_t \rangle) \leq 2C_\rho \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sum_{t=1}^n \epsilon_t \langle w, x_t \rangle = 2C_\rho B \mathbb{E}_\epsilon \left\| \sum_{t=1}^n \epsilon_t x_t \right\|.$$

Finally, the smoothness assumption on the norm (via [Theorem 7](#)) implies

$$\mathbb{E}_\epsilon \left\| \sum_{t=1}^n \epsilon_t x_t \right\| \leq \sqrt{2\beta R^2 n}.$$

□

Proof of Proposition 3. Observe that [Assumption 1](#) and [Assumption 2](#) respectively imply that $\|\nabla L_{\mathcal{D}}(w^*)\| = 0$ for the GLM and RR settings. Begin by invoking [Theorem 3](#). It is immediate that any algorithm that guarantees $\mathbb{E} \|\nabla \widehat{L}_n(\widehat{w}^{\text{alg}})\| \leq 1/\sqrt{n}$ will obtain the claimed sample complexity bound (the high-probability statement [Theorem 3](#) immediately yields an in-expectation statement due to boundedness), so all we must do is verify that such a point exists. [Proposition 2](#) along with [Lemma 7](#) and [Lemma 9](#) respectively indeed imply that $\|\nabla \widehat{L}_n(w^*)\|_2 \leq C/\sqrt{n}$ for both settings.

For completeness, we show below that both models indeed have Lipschitz gradients, and so standard smooth optimizers can be applied to the empirical loss.

Generalized Linear Model. Observe that for any (x, y) pair we have

$$\|\nabla \ell(w; x, y) - \nabla \ell(w'; x, y)\|_2 = 2\|x\|_2 |(\sigma(\langle w, x \rangle) - y)\sigma'(\langle w, x \rangle) - (\sigma(\langle w', x \rangle) - y)\sigma'(\langle w', x \rangle)|.$$

Letting $f(s) = (\sigma(s) - y)\sigma'(s)$, we see that the assumption on the loss guarantees $|f'(s)| \leq 3C_\sigma^2$, so we have

$$\|\nabla \ell(w; x, y) - \nabla \ell(w'; x, y)\|_2 \leq 6C_\sigma^2 R |\langle w - w', x \rangle| \leq 6C_\sigma^2 R^2 \|w - w'\|_2,$$

so smoothness is established.

Robust Regression. Following a similar calculation to the GLM case, we have

$$\begin{aligned} \|\nabla \ell(w; x, y) - \nabla \ell(w'; x, y)\|_2 &= \|x\|_2 |\rho'(\langle w, x \rangle - y) - \rho'(\langle w', x \rangle - y)| \\ &\leq C_\rho \|x\|_2 |\langle w - w', x \rangle| \\ &\leq C_\rho \|x\|_2^2 \|w - w'\|_2 \\ &\leq C_\rho R^2 \|w - w'\|_2. \end{aligned}$$

Now let $f(s) = (\sigma(s) - y)\sigma'(s)$, and observe that $|f'(s)| \leq 3C_\sigma^2$, so we have

$$\|\nabla \ell(w; x, y) - \nabla \ell(w'; x, y)\|_2 \leq 6C_\sigma^2 R |\langle w - w', x \rangle| \leq 6C_\sigma^2 R^2 \|w - w'\|_2.$$

□

C.3 Further Discussion

Detailed comparison with [31] We now sketch in more detail the relation between the rates of Theorem 3 and Theorem 4 and those of [31]. We focus on the fast rate regime, and on the case $R = \sqrt{d}$ (e.g., when $x \sim \mathcal{N}(0, I_{d \times d})$).

- *Uniform convergence.* Their uniform convergence bounds scale as $O(\tau\sqrt{d/n})$, where τ is the subgaussian parameter for the data x , whereas our uniform convergence bounds scale as $O(R^2\sqrt{1/n})$. When $R = \sqrt{d}$ both bounds scale as $O(d\sqrt{1/n})$, but our bounds do not depend on d when R is constant, whereas their bound always pays \sqrt{d} .
- *Parameter convergence.* The final result of [31] is a parameter convergence bound of the form $\|\widehat{w}^{\text{alg}} - w^*\|_2 \leq O\left(\frac{\tau}{\gamma\tau^2} \sqrt{\frac{d}{n}}\right)$ (see Theorem 4/6; Eqs. (106) and (96)). Our main result for the “low-dimensional” setup in Theorem 3 and Theorem 4 is an excess risk bound of the form $L_{\mathcal{D}}(\widehat{w}^{\text{alg}}) - L_{\mathcal{D}}(w^*) \leq O\left(\frac{R^4}{\lambda_{\min}(\Sigma)n}\right)$ which implies a parameter convergence bound of $\|\widehat{w}^{\text{alg}} - w^*\|_2 \leq \frac{R^2}{\lambda_{\min}(\Sigma)\sqrt{n}}$ (using similar reasoning as in the proof of Lemma 5 and Lemma 8). With $\tau = R = \sqrt{d}$ and Assumptions 6 and 9 in [31], we have $\lambda_{\min}(\Sigma) = \underline{\gamma}\tau^2$, and so again both the bounds resolve to $O\left(\frac{d}{\lambda_{\min}(\Sigma)\sqrt{n}}\right)$.

Analysis of regularized stationary point finding for high-dimensional setting Here we show that any algorithm that finds a stationary point of the regularized empirical loss generically succeeds obtains optimal sample complexity in the high-dimensional/norm-based setting. We focus on the generalized linear model in the Euclidean setting.

Let $r(w) = \frac{\lambda}{2}\|w\|_2^2$. Define $L_{\mathcal{D}}^{\lambda}(w) = L_{\mathcal{D}}(w) + r(w)$ and $\widehat{L}_n^{\lambda}(w) = \widehat{L}_n(w) + r(w)$. We consider any algorithm that returns a point \widehat{w} with $\nabla \widehat{L}_n^{\lambda}(\widehat{w}) = 0$, i.e. any stationary point of the regularized empirical risk.

Theorem 9. *Consider the generalized linear model setting. Let \widehat{w} be any point with $\nabla \widehat{L}_n^{\lambda}(\widehat{w}) = 0$. Suppose that $\|w^*\|_2 = 1$ and $C_{\sigma}, R > 1$. Then there is some absolute constant $c > 0$ such that for any fixed $\delta > 0$, if the regularization parameter λ satisfies*

$$\lambda > c \cdot \sqrt{\frac{R^4 C_{\sigma}^6}{c_{\sigma}^2} \cdot \frac{\log(\log(C_{\sigma} R n)/\delta)}{n}},$$

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

$$L_{\mathcal{D}}(\widehat{w}) - L_{\mathcal{D}}(w^*) \leq O\left(\frac{R^2 C_{\sigma}^4}{c_{\sigma}^2} \cdot \sqrt{\frac{\log(\log(C_{\sigma} R n)/\delta)}{n}}\right).$$

Theorem 9 easily extends to the robust regression setting by replacing invocations of Lemma 7 with Lemma 9 and use of (18) with (27).

Proof of Theorem 9. Recall that w^* minimizes the *unregularized* population risk, and that $\|w^*\|_2 = 1$. The technical challenge is to apply Lemma 7 even though we lack a good a-priori upper bound on the norm of \widehat{w} . We proceed by splitting the analysis into two cases. The idea is that if $\|\widehat{w}\|_2 \leq \|w^*\|_2$ we can apply Lemma 7 directly with no additional difficulty. On other hand, when $\|\widehat{w}\|_2 \geq \|w^*\|_2$ the regularized population risk satisfies the $(2, O(1/\lambda))$ -GD inequality, which is enough to show that excess risk is small even though $\|\widehat{w}\|_2$ could be larger than $\|w^*\|_2$.

Case 1: $\|\widehat{w}\|_2 \geq \|w^*\|_2$.

Let $\widetilde{W} = \{w \in \mathbb{R}^d \mid \|w\|_2 \geq \|w^*\|_2\}$, so that $\widehat{w} \in \widetilde{W}$. Observe that since $r(w)$ is λ -strongly convex it satisfies $r(w) - r(w^*) \leq \langle \nabla r(w), w - w^* \rangle - \frac{\lambda}{2}\|w - w^*\|_2^2$ for all w . Moreover, if $w \in \widetilde{W}$, we have

$$\langle \nabla r(w), w - w^* \rangle \geq r(w) - r(w^*) + \frac{\lambda}{2}\|w - w^*\|_2^2 \geq 0.$$

Using (18) and the definition of w^* , along with the strong convexity of r , we get

$$L_{\mathcal{D}}^{\lambda}(w) - L_{\mathcal{D}}^{\lambda}(w^*) \leq \frac{C_{\sigma}}{2c_{\sigma}} \langle \nabla L_{\mathcal{D}}(w), w - w^* \rangle + \langle \nabla r(w), w - w^* \rangle - \frac{\lambda}{2} \|w - w^*\|_2^2.$$

Since $\langle \nabla L_{\mathcal{D}}(w), w - w^* \rangle \geq 0$, this is upper bounded by

$$L_{\mathcal{D}}^{\lambda}(w) - L_{\mathcal{D}}^{\lambda}(w^*) \leq \frac{C_{\sigma}}{c_{\sigma}} \langle \nabla L_{\mathcal{D}}(w), w - w^* \rangle + \langle \nabla r(w), w - w^* \rangle - \frac{\lambda}{2} \|w - w^*\|_2^2.$$

Using the non-negativity of $\langle \nabla r(w), w - w^* \rangle$ over \widetilde{W} , and that $C_{\sigma}/c_{\sigma} > 1$, this implies

$$L_{\mathcal{D}}^{\lambda}(w) - L_{\mathcal{D}}^{\lambda}(w^*) \leq \frac{C_{\sigma}}{c_{\sigma}} \langle \nabla L_{\mathcal{D}}^{\lambda}(w), w - w^* \rangle - \frac{\lambda}{2} \|w - w^*\|_2^2 \quad \forall w \in \widetilde{W}.$$

Applying Cauchy-Schwarz:

$$\leq \frac{C_{\sigma}}{c_{\sigma}} \|\nabla L_{\mathcal{D}}^{\lambda}(w)\|_2 \|w - w^*\|_2 - \frac{\lambda}{2} \|w - w^*\|_2^2 \quad \forall w \in \widetilde{W}.$$

Using the AM-GM inequality:

$$\leq \frac{C_{\sigma}^2}{c_{\sigma}^2 \lambda} \|\nabla L_{\mathcal{D}}^{\lambda}(w)\|_2^2 \quad \forall w \in \widetilde{W}.$$

Using that $\widehat{w} \in \widetilde{W}$, and that $\nabla \widehat{L}_n^{\lambda}(\widehat{w}) = 0$, we have

$$L_{\mathcal{D}}^{\lambda}(\widehat{w}) - L_{\mathcal{D}}^{\lambda}(w^*) \leq \frac{C_{\sigma}^2}{c_{\sigma}^2 \lambda} \|\nabla L_{\mathcal{D}}^{\lambda}(\widehat{w}) - \nabla \widehat{L}_n^{\lambda}(\widehat{w})\|_2^2. \quad (31)$$

Observe that since \widehat{w} is a stationary point of the empirical risk, $\nabla \widehat{L}_n(\widehat{w}) = -\lambda \widehat{w}$, and so $\|\widehat{w}\|_2 \leq \frac{1}{\lambda} \|\nabla \widehat{L}_n(\widehat{w})\|_2 \leq \frac{2C_{\sigma}R}{\lambda}$ with probability 1. Thus, if we apply Lemma 10 with $B_{\max} = \frac{2C_{\sigma}R}{\lambda}$, we get that with probability at least $1 - \delta$,

$$\|\nabla L_{\mathcal{D}}^{\lambda}(\widehat{w}) - \nabla \widehat{L}_n^{\lambda}(\widehat{w})\|_2 \leq O\left(\|\widehat{w}\|_2 R^2 C_{\sigma}^2 \sqrt{\frac{1}{n}} + C_{\sigma} R \sqrt{\frac{\log(\log(C_{\sigma}R/\lambda)/\delta)}{n}}\right),$$

where we have used additionally that the regularization term does not depend on data. Combining this bound with (31), and using that $\widehat{w} \in \widetilde{W}$ and the elementary inequality $(a + b)^2 \leq 2(a^2 + b^2)$, we see that there exist constants $c, c' > 0$ such that

$$L_{\mathcal{D}}^{\lambda}(\widehat{w}) - L_{\mathcal{D}}^{\lambda}(w^*) \leq c \cdot \|\widehat{w}\|_2^2 \cdot \frac{R^4 C_{\sigma}^6}{\lambda c_{\sigma}^2} \cdot \frac{1}{n} + c' \cdot \frac{R^2 C_{\sigma}^4}{\lambda c_{\sigma}^2} \cdot \frac{\log(\log(C_{\sigma}R/\lambda)/\delta)}{n}.$$

Expanding the definition of the regularized excess risk, this is equivalent to

$$L_{\mathcal{D}}(\widehat{w}) - L_{\mathcal{D}}(w^*) \leq \lambda + \|\widehat{w}\|_2^2 \cdot \left(c \cdot \frac{R^4 C_{\sigma}^6}{\lambda c_{\sigma}^2} \cdot \frac{1}{n} - \lambda\right) + c' \cdot \frac{R^2 C_{\sigma}^4}{\lambda c_{\sigma}^2} \cdot \frac{\log(\log(C_{\sigma}R/\lambda)/\delta)}{n}.$$

Observe that if $\lambda > \sqrt{c \cdot \frac{R^4 C_{\sigma}^6}{c_{\sigma}^2} \cdot \frac{1}{n}}$ the middle term in this expression is at most zero. We choose

$$\lambda > \sqrt{c \cdot \frac{R^4 C_{\sigma}^6}{c_{\sigma}^2} \cdot \frac{\log(\log(C_{\sigma}Rn)/\delta)}{n}}.$$

Substituting choice this into the expression above leads to a final bound of

$$L_{\mathcal{D}}(\widehat{w}) - L_{\mathcal{D}}(w^*) \leq O\left(\frac{R^2 C_{\sigma}^3}{c_{\sigma}} \cdot \sqrt{\frac{\log(\log(C_{\sigma}Rn)/\delta)}{n}}\right).$$

Case 2: $\|\hat{w}\|_2 \leq \|w^*\|_2$.

Recall that $\nabla \widehat{L}_n^\lambda(\hat{w}) = 0$. This implies $\nabla \widehat{L}_n(\hat{w}) = -\lambda \hat{w}$, and so $\|\nabla \widehat{L}_n(\hat{w})\|_2 \leq \lambda \|\hat{w}\|_2 \leq \lambda$. Using (18) we have

$$L_{\mathcal{D}}(\hat{w}) - L_{\mathcal{D}}(w^*) \leq \frac{C_\sigma}{2c_\sigma} \langle \nabla L_{\mathcal{D}}(\hat{w}), \hat{w} - w^* \rangle \leq \frac{C_\sigma}{c_\sigma} \|\nabla L_{\mathcal{D}}(\hat{w})\|_2.$$

Using the bound on the empirical gradient above, we get

$$\|\nabla L_{\mathcal{D}}(\hat{w})\|_2 \leq \lambda + \|\nabla L_{\mathcal{D}}(\hat{w}) - \nabla \widehat{L}_n(\hat{w})\|_2.$$

Using (12), (13), and Lemma 7, applied with $B = 1$, we have that with probability at least $1 - \delta$,

$$\|\nabla L_{\mathcal{D}}(\hat{w}) - \nabla \widehat{L}_n(\hat{w})\|_2 \leq O\left(R^2 C_\sigma^2 \sqrt{\frac{\log(1/\delta)}{n}}\right),$$

and so

$$L_{\mathcal{D}}(\hat{w}) - L_{\mathcal{D}}(w^*) \leq O\left(\lambda \frac{C_\sigma}{c_\sigma} + \frac{R^2 C_\sigma^3}{c_\sigma} \sqrt{\frac{\log(1/\delta)}{n}}\right).$$

Substituting in the choice for λ :

$$\leq O\left(\frac{R^2 C_\sigma^4}{c_\sigma^2} \cdot \sqrt{\frac{\log(\log(C_\sigma R n)/\delta)}{n}}\right).$$

□

Lemma 10. Let $L_{\mathcal{D}}$ and \widehat{L}_n be the population and empirical risk for the generalized linear model setting. Let a parameter $B_{\max} \geq 1$ be given. Then with probability at least $1 - \delta$, for all $w \in \mathbb{R}^d$ with $1 \leq \|w\|_2 \leq B_{\max}$,

$$\|\nabla L_{\mathcal{D}}(w) - \nabla \widehat{L}_n(w)\|_2 \leq O\left(\|w\|_2 R^2 C_\sigma^2 \sqrt{\frac{1}{n}} + C_\sigma R \sqrt{\frac{\log(\log(B_{\max})/\delta)}{n}}\right),$$

where all constants are as in Assumption 1.

Proof. (12), (13), and Lemma 7 imply that for any fixed B , with probability at least $1 - \delta$,

$$\sup_{w: \|w\|_2 \leq B} \|\nabla L_{\mathcal{D}}(w) - \nabla \widehat{L}_n(w)\| \leq O\left(B R^2 C_\sigma^2 \sqrt{\frac{1}{n}} + C_\sigma R \sqrt{\frac{\log(\frac{1}{\delta})}{n}}\right).$$

Define $B_i = e^{i-1}$ for $1 \leq i \leq \lceil \log(B_{\max}) \rceil + 1$. The via a union bound, we have that for all i simultaneously,

$$\sup_{w: \|w\|_2 \leq B_i} \|\nabla L_{\mathcal{D}}(w) - \nabla \widehat{L}_n(w)\| \leq O\left(B_i R^2 C_\sigma^2 \sqrt{\frac{1}{n}} + C_\sigma R \sqrt{\frac{\log(\log(B_{\max})/\delta)}{n}}\right).$$

In particular, for any fixed w with $1 \leq \|w\|_2 \leq B_{\max}$, if we take i to be the smallest index for which $\|w\|_2 \leq B_i$, the expression above implies

$$\|\nabla L_{\mathcal{D}}(w) - \nabla \widehat{L}_n(w)\|_2 \leq O\left(\|w\|_2 R^2 C_\sigma^2 \sqrt{\frac{1}{n}} + C_\sigma R \sqrt{\frac{\log(\log(B_{\max})/\delta)}{n}}\right),$$

since $B_i \leq e\|w\|_2$.

□

Analysis of mirror descent for high-dimensional setting. Here we show that mirror descent obtains optimal excess risk for the norm-based/high-dimensional regime in [Theorem 3](#) and [Theorem 4](#).

Our approach is to run mirror descent with Ψ^* as the regularizer. Observe that Ψ^* is $\frac{1}{\beta}$ -strongly convex with respect to the dual norm $\|\cdot\|_*$, and that we have $\|\nabla \ell(w; x, y)\| \leq 2C_\sigma R$ for the GLM setting and $\|\nabla \ell(w; x, y)\| \leq C_\rho R$ for the RR setting.

Focusing on the GLM, if we take a single pass over the entire dataset $\{(x_t, y_t)\}_{t=1}^n$ in order, the standard analysis for mirror descent starting at $w_1 = 0$ with optimal learning rate tuning [\[15\]](#) guarantees that the following inequality holds deterministically:

$$\frac{1}{n} \sum_{t=1}^n \langle \nabla \ell(w_t; x_t, y_t), w_t - w^* \rangle \leq O\left(RBC_\sigma \sqrt{\frac{\beta}{n}}\right).$$

Since each point is visited a single time, this leads to the following guarantee on the population loss in expectation

$$\mathbb{E}\left[\frac{1}{n} \sum_{t=1}^n \langle \nabla L_{\mathcal{D}}(w_t), w_t - w^* \rangle\right] \leq O\left(RBC_\sigma \sqrt{\frac{\beta}{n}}\right).$$

Consequently, if we define \widehat{w} to be the result of choosing a single time $t \in [n]$ uniformly at random and returning w_t , this implies that

$$\mathbb{E}[\langle \nabla L_{\mathcal{D}}(\widehat{w}), \widehat{w} - w^* \rangle] \leq O\left(RBC_\sigma \sqrt{\frac{\beta}{n}}\right).$$

Combining this inequality with [\(18\)](#), we have

$$\mathbb{E}[L_{\mathcal{D}}(\widehat{w}) - L_{\mathcal{D}}(w^*)] \leq O\left(RB \frac{C_\sigma^2}{c_\sigma} \sqrt{\frac{\beta}{n}}\right).$$

Likewise, combining the mirror descent upper bound with [\(27\)](#) leads to a rate of $O\left(RB \frac{C_\rho^2}{c_\rho} \sqrt{\frac{\beta}{n}}\right)$ for robust regression. Thus, when all parameters involved are constant, it suffices to take $n = \frac{1}{\varepsilon^2}$ to obtain $O(\varepsilon)$ excess risk in both settings.

D Proofs from [Section 4](#)

Proof of [Theorem 5](#). Let $B \in \mathbb{R}^{d \times d}$ be a matrix for which the i th row B_i is given by $B_i = \frac{1}{\sqrt{d}}(\mathbf{1} - e_i)$.

We first focus on the more technical case where $n \geq d$.

Let $n = N \cdot d$ for some odd $N \in \mathbb{N}$. We partition time into d consecutive segments: $S_1 = \{1, \dots, N\}$, $S_2 = \{N+1, \dots, 2N\}$ and on. The sequence of instances $x_{1:n}$ we will use will be to set $x_t = B_i$ for $t \in S_i$. Note that $\|B_i\|_2 \leq 1$, so this choice indeed satisfies the boundedness constraint.

For simplicity, assume that $y_t = -1$ for all $t \in [n]$. Then it holds that

$$\begin{aligned} \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \left\| \sum_{t=1}^n \epsilon_t \nabla \ell(w; x_t, y_t) \right\|_2 &= \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \left\| \sum_{t=1}^n \epsilon_t \mathbb{1}\{\langle w, x_t \rangle \geq 0\} x_t \right\|_2 \\ &= \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \left\| \sum_{i=1}^d \mathbb{1}\{\langle w, B_i \rangle \geq 0\} \sum_{t \in S_i} \epsilon_t x_t \right\|_2 \end{aligned}$$

We introduce the notation $\varphi_i = \sum_{t \in S_i} \epsilon_t$.

$$\begin{aligned} &= \mathbb{E}_\varphi \sup_{w \in \mathcal{W}} \left\| \sum_{i=1}^d \mathbb{1}\{\langle w, B_i \rangle \geq 0\} \varphi_i B_i \right\|_2 \\ &= \mathbb{E}_\varphi \sup_{w \in \mathcal{W}} \left\| \sum_{i=1}^d \mathbb{1}\{\langle w, B_i \rangle \geq 0\} \varphi_i \frac{1}{\sqrt{d}}(\mathbf{1} - e_i) \right\|_2 \end{aligned}$$

Using triangle inequality:

$$\begin{aligned}
&\geq \mathbb{E}_\varphi \sup_{w \in \mathcal{W}} \left\| \sum_{i=1}^d \mathbb{1}\{\langle w, B_i \rangle \geq 0\} \varphi_i \frac{1}{\sqrt{d}} \mathbf{1} \right\|_2 - \frac{1}{\sqrt{d}} \mathbb{E}_\varphi \sum_{i=1}^d |\varphi_i| \\
&= \mathbb{E}_\varphi \sup_{w \in \mathcal{W}} \left| \sum_{i=1}^d \mathbb{1}\{\langle w, B_i \rangle \geq 0\} \varphi_i \right| - \frac{1}{\sqrt{d}} \mathbb{E}_\varphi \sum_{i=1}^d |\varphi_i| \\
&\geq \mathbb{E}_\varphi \sup_{w \in \mathcal{W}} \left| \sum_{i=1}^d \mathbb{1}\{\langle w, B_i \rangle \geq 0\} \varphi_i \right| - O(\sqrt{n}).
\end{aligned}$$

Now, for a given draw of φ , we choose $w \in \mathcal{W}$ such that $\text{sgn}(\langle w, B_i \rangle) = \text{sgn}(\varphi_i)$. The key trick here is that B is invertible, so for a given sign pattern, say $\sigma \in \{\pm 1\}^d$, we can set $\tilde{w} = B^{-1}\sigma$ and then $w = \tilde{w}/\|\tilde{w}\|_2$ to achieve this pattern. To see that B is invertible, observe that we can write it as $B = \frac{1}{\sqrt{d}}(\mathbf{1}\mathbf{1}^\top - I)$. The identity matrix can itself be written as $\frac{1}{d}\mathbf{1}\mathbf{1}^\top + A_\perp$, where $\mathbf{1} \notin \text{span}(A_\perp)$, so it can be seen that $B = \frac{1}{\sqrt{d}}((1 - \frac{1}{d})\mathbf{1}\mathbf{1}^\top - A_\perp)$, and that the $\mathbf{1}\mathbf{1}^\top$ component is preserved by this addition.

We have now arrived at a lower bound of $\mathbb{E}_\varphi \left| \sum_{i=1}^d \mathbb{1}\{\text{sgn}(\varphi_i) \geq 0\} \varphi_i \right|$. This value is lower bounded by

$$\begin{aligned}
&\mathbb{E}_\varphi \left| \sum_{i=1}^d \mathbb{1}\{\text{sgn}(\varphi_i) \geq 0\} \varphi_i \right| \\
&= \mathbb{E}_\varphi \sum_{i=1}^d \mathbb{1}\{\text{sgn}(\varphi_i) \geq 0\} |\varphi_i|
\end{aligned}$$

Now, observe that since N is odd we have $\text{sgn}(\varphi_i) \in \{\pm 1\}$, and so $\mathbb{1}\{\text{sgn}(\varphi_i) \geq 0\} = (1 + \text{sgn}(\varphi_i))/2$. Furthermore, since φ_i is symmetric, we may replace $\text{sgn}(\varphi_i)$ with an independent Rademacher random variable σ_i

$$\begin{aligned}
&= \mathbb{E}_\varphi \mathbb{E}_\sigma \frac{1}{2} \sum_{i=1}^d (1 + \sigma_i) |\varphi_i| \\
&= \mathbb{E}_\varphi \frac{1}{2} \sum_{i=1}^d |\varphi_i|.
\end{aligned}$$

Lastly, the Khintchine inequality implies that $\mathbb{E}_{\varphi_i} |\varphi_i| \geq \sqrt{N/2}$, so the final lower bound is $\Omega(d\sqrt{N}) = \Omega(\sqrt{dn})$.

In the case where $d \geq n$, the argument above easily yields that $\mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \left\| \sum_{t=1}^n \epsilon_t \nabla \ell(w; x_t, y_t) \right\|_2 = \Omega(n)$.

□

D.1 Proof of Theorem 6

Before proceeding to the proof, let us introduce some auxiliary definitions and results. The following functions will be used throughout the proof. They are related by Lemma 11.

$$\begin{aligned}
\xi_{\mathcal{D}}(w, \gamma) &= \mathbb{E}_{x \sim \mathcal{D}} \mathbb{1}\left\{ \frac{|\langle w, x \rangle|}{\|w\| \|x\|} \leq \gamma \right\}, \\
\widehat{\xi}_n(w, \gamma) &= \frac{1}{n} \sum_{t=1}^n \mathbb{1}\left\{ \frac{|\langle w, x_t \rangle|}{\|w\| \|x_t\|} \leq \gamma \right\}.
\end{aligned}$$

Lemma 11. With probability at least $1 - \delta$, simultaneously for all $w \in \mathcal{W}$ and all $\gamma > 0$,

$$\begin{aligned}
\xi_{\mathcal{D}}(w, \gamma) &\leq \widehat{\xi}_n(w, 2\gamma) + \frac{4}{\gamma\sqrt{n}} + \sqrt{\frac{2 \log(\log_2(4/\gamma)/\delta)}{n}}, \\
\widehat{\xi}_n(w, \gamma) &\leq \xi_{\mathcal{D}}(w, 2\gamma) + \frac{4}{\gamma\sqrt{n}} + \sqrt{\frac{2 \log(\log_2(4/\gamma)/\delta)}{n}}.
\end{aligned}$$

Proof sketch for Lemma 11. We only sketch the proof here as it follows standard analysis (see Theorem 5 of [20]). The key technique is to introduce a Lipschitz function $\zeta_\gamma(t)$:

$$\zeta_\gamma(t) = \begin{cases} 1 & |t| \leq \gamma \\ 2 - |t|/\gamma & \gamma < |t| < 2\gamma \\ 0 & |t| \geq 2\gamma \end{cases}.$$

Observe that ζ_γ satisfies $\mathbb{1}\{|t| > \gamma\} \leq \zeta_\gamma(t) \leq \mathbb{1}\{|t| > 2\gamma\}$ for all t . This sandwiching allows us to bound $\sup_{w \in \mathcal{W}} \{\xi_{\mathcal{D}}(w, \gamma) - \widehat{\xi}_n(w, 2\gamma)\}$ (and $\sup_{w \in \mathcal{W}} \{\widehat{\xi}_n(w, \gamma) - \xi_{\mathcal{D}}(w, 2\gamma)\}$) by instead bounding the difference between the empirical and population averages of the surrogate ζ_γ . This is achieved easily using the Lipschitz contraction lemma for Rademacher complexity, and by noting that the Rademacher complexity of the class $\{x \mapsto \langle w, x \rangle \mid \|w\|_2 \leq 1\}$ is at most \sqrt{n} whenever data satisfies $\|x_t\|_2 \leq 1$ for all t . Finally, a union bound over values of γ in the range $[0, 1]$ yields the statement. \square

Proof of Theorem 6. Let the margin function ϕ and $\delta > 0$ be fixed. Define functions $\psi(\cdot)$, $\phi_1(\cdot)$, and $\phi_2(\cdot)$ as follows:

$$\begin{aligned} \psi(\gamma) &= \frac{4}{\gamma\sqrt{n}} + \sqrt{\frac{2 \log(\log_2(4/\gamma)/\delta)}{n}} \\ \phi_1(\gamma) &= \phi(2\gamma) + \psi(\gamma) \\ \phi_2(\gamma) &= \phi(4\gamma) + 2\psi(2\gamma). \end{aligned}$$

Now, conditioning on the events of Lemma 11, we have that with probability at least $1 - \delta$,

$$\mathcal{W}(\phi, \widehat{\mathcal{D}}_n) \subseteq \mathcal{W}(\phi_1, \mathcal{D}) \subseteq \mathcal{W}(\phi_2, \widehat{\mathcal{D}}_n). \quad (32)$$

Consequently, we have the upper bound

$$\begin{aligned} \sup_{w \in \mathcal{W}(\phi, \widehat{\mathcal{D}}_n)} \|\nabla L_{\mathcal{D}}(w) - \nabla \widehat{L}_n(w)\|_2 &\leq \sup_{w \in \mathcal{W}(\phi_1, \mathcal{D})} \|\nabla L_{\mathcal{D}}(w) - \nabla \widehat{L}_n(w)\|_2 \\ &\leq 4 \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}(\phi_1, \mathcal{D})} \left\| \frac{1}{n} \sum_{t=1}^n \epsilon_t \nabla \ell(w; x_t, y_t) \right\| + 4 \sqrt{\frac{\log(2/\delta)}{n}} \\ &\leq 4 \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}(\phi_2, \widehat{\mathcal{D}}_n)} \left\| \frac{1}{n} \sum_{t=1}^n \epsilon_t \nabla \ell(w; x_t, y_t) \right\| + 4 \sqrt{\frac{\log(2/\delta)}{n}}, \end{aligned} \quad (33)$$

where the second inequality holds with probability at least $1 - \delta$ using Lemma 4. The key here is that we are able to apply the standard symmetrization result because we have replaced $\mathcal{W}(\phi, \widehat{\mathcal{D}}_n)$ with a set that does not depend on data. Next, invoking the chain rule (Theorem 1), we split the Rademacher complexity term above as:

$$\mathbb{E}_\epsilon \sup_{w \in \mathcal{W}(\phi_2, \widehat{\mathcal{D}}_n)} \left\| \frac{1}{n} \sum_{t=1}^n \epsilon_t \nabla \ell(w; x_t, y_t) \right\| \leq \underbrace{2 \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}(\phi_2, \widehat{\mathcal{D}}_n)} \frac{1}{n} \sum_{t=1}^n \epsilon_t \mathbb{1}\{y_t \langle w, x_t \rangle \leq 0\}}_{(\diamond)} + \frac{2}{n} \mathbb{E}_\epsilon \left\| \sum_{t=1}^n \epsilon_t x_t \right\|_2. \quad (34)$$

The second term is controlled by Theorem 7, which gives $\frac{1}{n} \mathbb{E}_\epsilon \left\| \sum_{t=1}^n \epsilon_t x_t \right\|_2 \leq \frac{1}{\sqrt{n}}$. For the first term, we appeal to the fat-shattering dimension and the ϕ_2 -soft-margin assumption.

Controlling (\diamond) . Observe that for any fixed $\tilde{\gamma} > 0$, we can split (\diamond) as

$$\begin{aligned} \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}(\phi_2, \widehat{\mathcal{D}}_n)} \frac{1}{n} \sum_{t=1}^n \epsilon_t \mathbb{1}\{y_t \langle w, x_t \rangle \leq 0\} &\leq \underbrace{\mathbb{E}_\epsilon \sup_{w \in \mathcal{W}(\phi_2, \widehat{\mathcal{D}}_n)} \frac{1}{n} \sum_{t=1}^n \epsilon_t \mathbb{1}\{y_t \langle w, x_t \rangle \leq 0 \wedge \frac{|\langle w, x_t \rangle|}{\|w\|_2 \|x_t\|_2} \geq \tilde{\gamma}\}}_{(*)} \\ &\quad + \underbrace{\mathbb{E}_\epsilon \sup_{w \in \mathcal{W}(\phi_2, \widehat{\mathcal{D}}_n)} \frac{1}{n} \sum_{t=1}^n \epsilon_t \mathbb{1}\{y_t \langle w, x_t \rangle \leq 0 \wedge \frac{|\langle w, x_t \rangle|}{\|w\|_2 \|x_t\|_2} < \tilde{\gamma}\}}_{(**)}. \end{aligned}$$

For $(\star\star)$, the definition of $\mathcal{W}(\phi_2, \widehat{\mathcal{D}}_n)$ implies

$$\mathbb{E}_\epsilon \sup_{w \in \mathcal{W}(\phi_2, \widehat{\mathcal{D}}_n)} \frac{1}{n} \sum_{t=1}^n \epsilon_t \mathbb{1} \left\{ y_t \langle w, x_t \rangle \leq 0 \wedge \frac{|\langle w, x_t \rangle|}{\|w\|_2 \|x_t\|_2} < \tilde{\gamma} \right\} \leq \sup_{w \in \mathcal{W}(\phi_2, \widehat{\mathcal{D}}_n)} \frac{1}{n} \sum_{t=1}^n \mathbb{1} \left\{ \frac{|\langle w, x_t \rangle|}{\|w\|_2 \|x_t\|_2} < \tilde{\gamma} \right\} \leq \phi_2(\tilde{\gamma}). \quad (35)$$

The quantity (\star) can be bounded by writing it as

$$\mathbb{E}_\epsilon \sup_{v \in V} \frac{1}{n} \sum_{t=1}^n \epsilon_t v_t,$$

where V is a boolean concept class defined as $V = \{(v_1(w), \dots, v_n(w)) \in \{0, 1\}^n \mid w \in \mathcal{W}(\phi_2, \widehat{\mathcal{D}}_n)\}$, where $v_i(w) := \mathbb{1} \left\{ y_i \frac{\langle w, x_i \rangle}{\|w\|_2 \|x_i\|_2} \leq 0 \right\}$. $\mathbb{1} \left\{ \frac{|\langle w, x_i \rangle|}{\|w\|_2 \|x_i\|_2} \geq \tilde{\gamma} \right\}$. The standard Massart finite class lemma (e.g. [32]) implies

$$\mathbb{E}_\epsilon \sup_{v \in V} \frac{1}{n} \sum_{t=1}^n \epsilon_t v_t \leq \sqrt{\frac{2 \log |V|}{n}}.$$

All that remains is to bound the cardinality of V . To this end, note that we can bound $|V|$ by first counting the number of realizations of $\left(\mathbb{1} \left\{ \frac{|\langle w, x_1 \rangle|}{\|w\|_2 \|x_1\|_2} \geq \tilde{\gamma} \right\}, \dots, \mathbb{1} \left\{ \frac{|\langle w, x_n \rangle|}{\|w\|_2 \|x_n\|_2} \geq \tilde{\gamma} \right\} \right)$ as we vary $w \in \mathcal{W}(\phi_2, \widehat{\mathcal{D}}_n)$. This is at most $\binom{n}{n\phi_2(\tilde{\gamma})} \leq n^{n\phi_2(\tilde{\gamma})}$, since the number of points with margin smaller than $\tilde{\gamma}$ is bounded by $n\phi_2(\tilde{\gamma})$ via (32).

Next, we consider only the points x_t for which $\mathbb{1} \left\{ \frac{|\langle w, x_t \rangle|}{\|w\|_2 \|x_t\|_2} \geq \tilde{\gamma} \right\} = 1$. On these points, on which we are guaranteed to have a margin at least $\tilde{\gamma}$, we count the number of realizations of $\mathbb{1} \left\{ y_t \frac{\langle w, x_t \rangle}{\|w\|_2 \|x_t\|_2} \leq 0 \right\}$.

This is bounded by $n^{O(\frac{1}{\tilde{\gamma}^2})}$ due to the Sauer-Shelah lemma (e.g. [39]). The fat-shattering dimension at margin $\tilde{\gamma}$ coincides with the notion of shattering on these points, and [4] bound the fat-shattering dimension at scale $\tilde{\gamma}$ by $O(\frac{1}{\tilde{\gamma}^2})$. Hence, the cardinality of V is bounded by

$$|V| \leq n^{n\phi_2(\tilde{\gamma})} n^{O(\frac{1}{\tilde{\gamma}^2})}. \quad (36)$$

Final bound. Assembling equations (33), (34), (35), and (36) yields

$$\begin{aligned} \sup_{w \in \mathcal{W}(\phi, \widehat{\mathcal{D}}_n)} \|\nabla L_{\mathcal{D}}(w) - \nabla \widehat{L}_n(w)\|_2 &\leq O \left(\phi_2(\tilde{\gamma}) + \sqrt{\frac{\log |V|}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right) \\ &\leq O \left(\phi_2(\tilde{\gamma}) + \sqrt{\left(\phi_2(\tilde{\gamma}) + \frac{1}{\tilde{\gamma}^2 n} \right) \log(n)} + \sqrt{\frac{\log(1/\delta)}{n}} \right) \\ &\leq \tilde{O} \left(\sqrt{\phi_2(\tilde{\gamma})} + \frac{1}{\tilde{\gamma} \sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right) \\ &\leq \tilde{O} \left(\sqrt{\phi(4\tilde{\gamma})} + \frac{1}{\tilde{\gamma}} \sqrt{\frac{\log(1/\delta)}{n}} + \frac{1}{\sqrt{\tilde{\gamma}} n^{1/4}} \right). \end{aligned}$$

The chain of inequalities above follows by observing that $\phi_2(\tilde{\gamma}) = \phi(4\tilde{\gamma}) + 2\psi(2\tilde{\gamma})$ is bounded and thus $\phi_2(\gamma) \leq c\sqrt{\phi_2(\gamma)}$ for some constant c independent of $\tilde{\gamma}$. We get the desired result by optimizing over $\tilde{\gamma}$. \square

E Additional Results

Theorem 10 (Second-order chain rule for Rademacher complexity). *Let two sequences of twice-differentiable functions $G_t : \mathbb{R}^K \rightarrow \mathbb{R}$ and $F_t : \mathbb{R}^d \rightarrow \mathbb{R}^K$ be given, and let $F_{t,i}(w)$ denote the i th of*

coordinate of $F_t(w)$. Suppose there are constants $L_{F,1}$, $L_{F,2}$, $L_{G,1}$, $L_{G,2}$ such that for all $1 \leq t \leq n$, $\|\nabla G_t\|_2 \leq L_{G,1}$, $\sqrt{\sum_{i,j} \|(\nabla F_{t,i})(\nabla F_{t,j})^\top\|_\sigma^2} \leq L_{F,1}$, $\|\nabla^2 G_t\|_2 \leq L_{G,2}$ and $\sqrt{\sum_{k=1}^K \|\nabla^2 F_{t,k}\|_\sigma^2} \leq L_{F,2}$. Then,

$$\begin{aligned} \frac{1}{2} \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \left\| \sum_{t=1}^n \epsilon_t \nabla^2 (G_t(F_t(w))) \right\|_\sigma &\leq L_{F,1} \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sum_{t=1}^n \langle \epsilon_t, \nabla G_t(F_t(w)) \rangle \\ &\quad + L_{G,1} \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \left\| \sum_{t=1}^n \sum_{i=1}^K \epsilon_{t,i} \nabla^2 F_{t,i}(w) \right\|_\sigma \\ &\quad + L_{F,2} \mathbb{E}_{\tilde{\epsilon}} \sup_{w \in \mathcal{W}} \sum_{t=1}^n \langle \tilde{\epsilon}_t, \nabla^2 G_t(F_t(w)) \rangle \\ &\quad + L_{G,2} \mathbb{E}_{\tilde{\epsilon}} \sup_{w \in \mathcal{W}} \left\| \sum_{t=1}^n \sum_{i,j=1}^K \tilde{\epsilon}_{t,i,j} \nabla F_{t,i}(w) \nabla F_{t,j}(w)^\top \right\|_\sigma, \end{aligned}$$

where for all $i \in [K]$, $\nabla F_{t,i}(w)$ denotes the i^{th} column of the Jacobian matrix $\nabla F_t \in \mathbb{R}^{d \times K}$, $\nabla^2 F_{t,i} \in \mathbb{R}^{d \times d}$ denotes the i th slice of the Hessian operator $\nabla^2 F_t \in \mathbb{R}^{d \times d \times K}$, and $\epsilon \in \{\pm 1\}^{n,K}$ and $\tilde{\epsilon} \in \{\pm 1\}^{n \times K \times K}$ are matrices of Rademacher random variables.

As an application of [Theorem 10](#), we give a simple proof of dimension-independent Rademacher bound for the generalized linear model setting.

Lemma 12. Assume in addition to [Assumption 1](#) assume that $|\sigma'''(s)| \leq C_\sigma$ for all $s \in \mathcal{S}$, and suppose $\|\cdot\|$ is any β -smooth norm. Then the empirical loss Hessian for the generalized linear model setting enjoys the normed Rademacher complexity bound,

$$\mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \left\| \sum_{t=1}^n \epsilon_t \nabla^2 \ell(w; x_t, y_t) \right\|_\sigma \leq O\left((BR^3 C_\sigma^2 \sqrt{\beta} + C_\sigma^2 R^2 \sqrt{\log(d)}) \sqrt{n}\right). \quad (37)$$

It is easy to see that the same approach leads to a normed Rademacher complexity bound for the Hessian in the robust regression setting as well. We leave the proof as an exercise.

Lemma 13. Assume in addition to [Assumption 2](#) that $|\rho'''(s)| \leq C_\rho$ for all $s \in \mathcal{S}$, and suppose $\|\cdot\|$ is any β -smooth norm. Then the empirical loss Hessian for the robust regression setting enjoys the normed Rademacher complexity bound:

$$\mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \left\| \sum_{t=1}^n \epsilon_t \nabla^2 \ell(w; x_t, y_t) \right\|_\sigma \leq O\left((BR^3 C_\rho \sqrt{\beta} + C_\rho R^2 \sqrt{\log(d)}) \sqrt{n}\right). \quad (38)$$

Proof of Theorem 10. We start by writing

$$\mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \left\| \sum_{t=1}^n \epsilon_t \nabla^2 (G_t(F_t(w))) \right\|_\sigma = \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sup_{\substack{u \in \mathbb{R}^d \\ \|u\|_2 \leq 1}} \sum_{t=1}^n \epsilon_t u^\top \nabla^2 (G_t(F_t(w))) u. \quad (39)$$

Using the chain rule for differentiation, we have for any $u \in \mathbb{R}^n$

$$u^\top \nabla^2 (G_t(F_t(w))) u = \langle \nabla F_t(w), u \rangle^\top \nabla^2 G_t(F_t(w)) \langle \nabla F_t(w), u \rangle + \langle \nabla G_t(F_t(w)), \nabla^2 F_t(w)[u, u] \rangle,$$

where $\nabla G_t(F_t(w))$ and $\nabla^2 G_t(F_t(w))$ denote the gradient and Hessian of G_t at $F_t(w)$, and $\nabla^2 F_t(w)[u, u] \in \mathbb{R}^K$ is a vector for which the i th coordinate is the evaluation of the Hessian operator for $F_{t,i}$ at (u, u) . Using this identity along with (39), we get

$$\begin{aligned} \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \left\| \sum_{t=1}^n \epsilon_t \nabla^2 (G_t(F_t(w))) \right\|_\sigma &\leq \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sup_{\substack{u \in \mathbb{R}^d \\ \|u\|_2 \leq 1}} \sum_{t=1}^n \epsilon_t \langle \nabla G_t(F_t(w)), \nabla^2 F_t(w)[u, u] \rangle \\ &\quad + \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sup_{\substack{u \in \mathbb{R}^d \\ \|u\|_2 \leq 1}} \sum_{t=1}^n \epsilon_t \langle \nabla F_t(w), u \rangle^\top \nabla^2 G_t(F_t(w)) \langle \nabla F_t(w), u \rangle. \end{aligned}$$

We bound the two terms separately.

1. *First Term:* We introduce a new function that relabels the quantities in the expression. Let $h_1 : \mathbb{R}^{2K} \rightarrow \mathbb{R}$ be defined as $h_1(a, b) = \langle a, b \rangle$, let $f_1 : \mathcal{W} \times \mathbb{R}^d \rightarrow \mathbb{R}^K$ be given by $f_1(w, u) = \nabla G_t(F_t(w))$ and $f_2 : \mathcal{W} \times \mathbb{R}^d \rightarrow \mathbb{R}^K$ be given by $f_2(w, u) = (\nabla^2 F_{t,k}(w)[u, u])_{k \in [K]}$. We apply the block-wise contraction lemma [Lemma 3](#) with one block for f_1 and one block for f_2 to conclude

$$\begin{aligned}
& \frac{1}{2} \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sup_{\substack{u \in \mathbb{R}^d \\ \|u\|_2 \leq 1}} \sum_{t=1}^n \epsilon_t h_1(f_1(w, u), f_2(w, u)) \\
& \leq L_{F,1} \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sup_{\substack{u \in \mathbb{R}^d \\ \|u\|_2 \leq 1}} \sum_{t=1}^n \langle \epsilon_t, f_1(w, u) \rangle + L_{G,1} \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sup_{\substack{u \in \mathbb{R}^d \\ \|u\|_2 \leq 1}} \sum_{t=1}^n \langle \epsilon_t, f_2(w, u) \rangle \\
& \leq L_{F,1} \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sup_{\substack{u \in \mathbb{R}^d \\ \|u\|_2 \leq 1}} \sum_{t=1}^n \langle \epsilon_t, \nabla G_t(F_t(w)) \rangle + L_{G,1} \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sup_{\substack{u \in \mathbb{R}^d \\ \|u\|_2 \leq 1}} \sum_{t=1}^n \langle \epsilon_t, \nabla^2 F_t(w)[u, u] \rangle \\
& \leq L_{F,1} \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sum_{t=1}^n \langle \epsilon_t, \nabla G_t(F_t(w)) \rangle + L_{G,1} \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sup_{\substack{u \in \mathbb{R}^d \\ \|u\|_2 \leq 1}} \left(\sum_{t=1}^n \nabla^2 F_t(w) \epsilon_t \right) [u, u] \\
& = L_{F,1} \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sum_{t=1}^n \langle \epsilon_t, \nabla G_t(F_t(w)) \rangle + L_{G,1} \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \left\| \sum_{t=1}^n \nabla^2 F_t(w) \epsilon_t \right\|_\sigma.
\end{aligned}$$

2. *Second Term:* Let us first simplify as

$$\begin{aligned}
\langle \nabla F_t(w), u \rangle^\top \nabla^2 G_t(F_t(w)) \langle \nabla F_t(w), u \rangle &= \sum_{i,j=1}^K \langle \nabla F_t(w), u \rangle_i \nabla^2 G_t(F_t(w))_{i,j} \langle \nabla F_t(w), u \rangle_j \\
&= \sum_{i,j=1}^K (u^\top \nabla F_{t,i}(w)) \times \frac{\partial^2 G_t}{\partial z_i \partial z_j} \times (\nabla F_{t,j}(w)^\top u) \\
&= \sum_{i,j=1}^K \frac{\partial^2 G_t}{\partial z_i \partial z_j} (u^\top \nabla F_{t,i}(w) \nabla F_{t,j}(w)^\top u),
\end{aligned}$$

where $\nabla F_{t,j}(w) := \nabla F_t(w)[\cdot, j] \in \mathbb{R}^d$, and the last equality follows by observing that $\frac{\partial^2 G_t}{\partial z_i \partial z_j}$ is scalar. We thus have

$$\begin{aligned}
& \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sup_{\substack{u \in \mathbb{R}^d \\ \|u\|_2 \leq 1}} \sum_{t=1}^n \epsilon_t \langle \nabla F_t(w), u \rangle^\top \nabla^2 G_t(F_t(w)) \langle \nabla F_t(w), u \rangle \\
& = \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sup_{\substack{u \in \mathbb{R}^d \\ \|u\|_2 \leq 1}} \sum_{t=1}^n \epsilon_t \sum_{i,j=1}^K \frac{\partial^2 G_t}{\partial z_i \partial z_j} (u^\top \nabla F_{t,i}(w) \nabla F_{t,j}(w)^\top u).
\end{aligned}$$

Similar to the first part, we introduce a new function that relabels the quantities in the expression. Let $h_2 : \mathbb{R}^{2K^2} \rightarrow \mathbb{R}$ be defined as $h_2(a, b) = \sum_{i,j=1}^K a_{i,j} b_{i,j}$. Let $f_1 : \mathcal{W} \times \mathbb{R}^d \rightarrow \mathbb{R}^{K^2}$ be given by $f_1(w, u) = (\nabla^2 G_t)(F_t(w))$ and $f_2 : \mathcal{W} \times \mathbb{R}^d \rightarrow \mathbb{R}^{K^2}$ be given by $f_2(w, u) = (u^\top \nabla F_{t,i}(w) \nabla F_{t,j}(w)^\top u)_{i,j \in [K]}$. We apply block-wise contraction ([Lemma 3](#))

with one block for f_1 and one block for f_2 to conclude

$$\begin{aligned}
& \frac{1}{2} \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sup_{\substack{u \in \mathbb{R}^d \\ \|u\|_2 \leq 1}} \sum_{t=1}^n \epsilon_t h_2(f_1(w, u), f_2(w, u)) \\
& \leq L_{F,2} \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sup_{\substack{u \in \mathbb{R}^d \\ \|u\|_2 \leq 1}} \sum_{t=1}^n \langle \epsilon_t, f_1(w, u) \rangle + L_{G,2} \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sup_{\substack{u \in \mathbb{R}^d \\ \|u\|_2 \leq 1}} \sum_{t=1}^n \langle \epsilon_t, f_2(w, u) \rangle \\
& \leq L_{F,2} \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sum_{t=1}^n \langle \epsilon_t, \nabla^2 G_t(F_t(w)) \rangle + L_{G,2} \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sup_{\substack{u \in \mathbb{R}^d \\ \|u\|_2 \leq 1}} \sum_{t=1}^n \sum_{i,j=1}^K \epsilon_{t,i,j} u^\top \nabla F_{t,i}(w) \nabla F_{t,j}(w)^\top u \\
& = L_{F,2} \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sum_{t=1}^n \langle \epsilon_t, \nabla^2 G_t(F_t(w)) \rangle + L_{G,2} \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \left\| \sum_{t=1}^n \sum_{i,j=1}^K \epsilon_{t,i,j} \nabla F_{t,i}(w) \nabla F_{t,j}(w)^\top \right\|_\sigma.
\end{aligned}$$

Combining the two terms gives the desired chain rule. \square

Proof of Lemma 12. As in Lemma 7, let $G_t(s) = (\sigma(s) - y_t)^2$ and $F_t(w) = \langle w, x_t \rangle$, so that $\ell(w; x_t, y_t) = G_t(F_t(w))$.

Observe that $G'_t(s) = 2(\sigma(s) - y_t)\sigma'(s)$, $\nabla F_t(w) = x_t$, $\nabla^2 F_t = \mathbf{0}$, $G''_t(s)(s) = 2(\sigma'(s))^2 + 2y_t\sigma''(s)$, and $G'''_t(s) = 4\sigma'(s)\sigma''(s) + 2y_t\sigma'''(s)$, which implies that $|G'''_t(s)| \leq 6C_\sigma^2$. Using Theorem 10 with constants $L_{F,1} = R^2$, $L_{F,2} = 0$, $L_{G,1} = 2C_\sigma^2$ and $L_{G,2} = 4C_\sigma^2$, we get

$$\mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \left\| \sum_{t=1}^n \epsilon_t \nabla^2 \ell(w; x_t, y_t) \right\|_\sigma \leq 2R^2 \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sum_{t=1}^n \epsilon_t G''_t(\langle w, x_t \rangle) + 8C_\sigma^2 \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \left\| \sum_{t=1}^n \epsilon_t x_t x_t^\top \right\|_\sigma,$$

applying Lemma 3,

$$\begin{aligned}
& \leq 12R^2 C_\sigma^2 \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \sum_{t=1}^n \epsilon_t \langle w, x_t \rangle + 8C_\sigma^2 \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}} \left\| \sum_{t=1}^n \epsilon_t x_t x_t^\top \right\|_\sigma \\
& = 12R^2 C_\sigma^2 B \mathbb{E}_\epsilon \left\| \sum_{t=1}^n \epsilon_t x_t \right\| + 8C_\sigma^2 \mathbb{E}_\epsilon \left\| \sum_{t=1}^n \epsilon_t x_t x_t^\top \right\|_\sigma.
\end{aligned}$$

Invoking Theorem 7 and Fact 1, we have $\mathbb{E}_\epsilon \left\| \sum_{t=1}^n \epsilon_t x_t \right\| \leq \sqrt{2\beta R^2 n}$ and $\mathbb{E}_\epsilon \left\| \sum_{t=1}^n \epsilon_t x_t x_t^\top \right\|_\sigma \leq \sqrt{2\log(d) R^4 n}$. \square