

A. Proof of theorems and technical lemmas

A.1. Proof of Proposition 3.2

From the definition of robust surrogate in (9) for the setting of Proposition 3.2 we have

$$\phi_\gamma(\theta; z_0) := \sup_x \left\{ (y_0 - x^T \theta)^2 - \gamma \|x - x_0\|_{\ell_2}^2 \right\} ,$$

by introducing $g_\gamma(x) := (y_0 - x^T \theta)^2 - \gamma \|x - x_0\|_{\ell_2}^2$, for every scalar $\epsilon$ we get

$$g_\gamma(x_0 + \epsilon \theta) = g_\gamma(x_0) + 2\epsilon (x_0^T \theta - y_0) \|\theta\|_{\ell_2}^2 + \epsilon^2 \|\theta\|_{\ell_2}^2 + \gamma \epsilon^2 > -\epsilon \gamma \|\theta\|_{\ell_2}^2 \geq 0 ,$$

this implies if $\gamma < \|\theta\|_{\ell_2}^2$, then $\phi_\gamma(\theta; z_0) = +\infty$. Consider $\gamma \geq \|\theta\|_{\ell_2}^2$, then from relation $\nabla^2 g_\gamma(x) = 2(\theta \theta^T - \gamma I)$ we realize that $g_\gamma$ is concave. Writing the first order optimal condition we have

$$\begin{align*}
(y_0 - x^T \theta) + \gamma (x - x_0) = 0 .
\end{align*}$$

Multiplying by $\theta$ and solving for $x^T \theta$, we get

$$x^T \theta = \frac{\gamma x^T \theta - y_0 \|\theta\|_{\ell_2}^2}{\gamma - \|\theta\|_{\ell_2}^2} .$$

Substituting for $x^T \theta$ in the stationary condition (34) implies

$$x^* = x_0 + \frac{x_0^T \theta - y_0 \theta}{\gamma - \|\theta\|_{\ell_2}^2} .$$

Replacing $x^*$ in $g_\gamma$ yields

$$\phi_\gamma(\theta; z) = \begin{cases} +\infty & \text{if } \gamma < \|\theta\|_{\ell_2}^2 , \\ \frac{(y_0 - x^T \theta)^2}{\gamma - \|\theta\|_{\ell_2}^2} & \text{if } \gamma \geq \|\theta\|_{\ell_2}^2 . \end{cases}$$

Then, we use dual formulation (8) to compute the Wasserstein adversarial risk:

$$\begin{align*}
\text{AR}(\theta) &= \sup_{Q \in \mathcal{U}_v(\mathcal{P}_x)} \mathbb{E}_z \mathbb{Q}[\ell(\theta; z)] \\
&= \inf_{\gamma \geq 0} \{ \gamma \mathbb{E}_Q[\phi_\gamma(\theta; z)] \} \\
&= \inf_{\gamma \geq \|\theta\|_{\ell_2}^2} \{ \gamma \mathbb{E}_Q[\phi_\gamma(\theta; z)] \} \\
&= \inf_{\gamma \geq \|\theta\|_{\ell_2}^2} \{ \gamma \mathbb{E}_Q[\ell(\theta(\theta; z))] \frac{\|\theta\|_{\ell_2}^2}{\gamma - \|\theta\|_{\ell_2}^2} .
\end{align*}$$

the infimum is achieved at

$$\gamma^* = \frac{1}{\epsilon} \sqrt{\mathbb{E}_Q[\ell(\theta; z)] \|\theta\|_{\ell_2}^2 + \|\theta\|_{\ell_2}^2} ,$$

Finally, this gives us

$$\text{AR}(\theta) = \left( \sqrt{\mathbb{E}_Q[\ell(\theta; z)] + \epsilon \|\theta\|_{\ell_2}^2} \right)^2 .$$

A.2. Proof of Theorem 3.3

Define $\mathcal{R}(\theta) := \lambda \text{SR}(\theta) + \text{AR}(\theta)$. Proposition 3.2 implies $\text{AR}(\theta) = \text{SR}(\theta) + 2\epsilon \|\theta\|_{\ell_2} \sqrt{\text{SR}(\theta)} + \epsilon^2 \|\theta\|_{\ell_2}^2$, then by expanding adversarial risk relation $\text{AR}(\theta)$ in $\mathcal{R}(\theta)$ we get

$$\mathcal{R}(\theta) = (1 + \lambda) \text{SR}(\theta) + \epsilon^2 \|\theta\|_{\ell_2}^2 + 2\epsilon \|\theta\|_{\ell_2} \sqrt{\text{SR}(\theta)} .$$

(36)

It is easy to see $\text{SR}(\theta) = \sigma_*^2 + \theta^T \Sigma \theta = 2v^T \theta$. Replace $\nabla \theta \text{SR}(\theta)$ as $2(\Sigma \theta - v)$ in (36) to get

$$\nabla \theta \mathcal{R}(\theta) = 2(1 + \lambda)(\Sigma \theta - v) + 2\epsilon\theta$$

$$+ 2\epsilon \left( \theta \frac{\theta}{\|\theta\|_{\ell_2}^2} \sqrt{\text{SR}(\theta)} + (\Sigma \theta - v) \frac{\|\theta\|_{\ell_2}}{\sqrt{\text{SR}(\theta)}} \right) .$$

(37)

therefore stationary points (solutions of $\nabla \theta \mathcal{R}(\theta) = 0$) and a critical point $\theta = 0$ are candidates for global minimizers. From equation $\text{SR}(\theta) = \sigma_*^2 + \theta^T \Sigma \theta = 2v^T \theta$ and adversarial risk relation in Proposition 3.2 it is clear that for $\theta = 0$ we have $\text{SR}(\theta) = \text{AR}(\theta) = \sigma_*^2$. Next, we focus on characterizing stationary minimizers of $\mathcal{R}(\theta)$ and their corresponding standard and adversarial risk values. If $\theta_*$ is a stationary point, then putting (37) to be zero yields

$$\begin{align*}
&\left( \left( 1 + \frac{\epsilon}{\lambda} \frac{\|\theta_*\|_{\ell_2}}{\sqrt{\text{SR}(\theta_*)}} \right) \Sigma + \left( \frac{\epsilon^2 \|\theta_\sigma\|_{\ell_2}}{\sqrt{\text{SR}(\theta_\sigma)}} \right) I \right) \theta_* \\
&= \left( 1 + \frac{\epsilon}{\lambda} \frac{\|\theta_*\|_{\ell_2}}{\sqrt{\text{SR}(\theta_*)}} \right) v .
\end{align*}$$

(38)

Introduce $A_* := \frac{\sqrt{\text{SR}(\theta_*)}}{\|\theta_*\|_{\ell_2}}$ and $\gamma_* := \frac{\epsilon^2 + \epsilon A_*}{1 + \lambda \frac{A_*}{2}}$, then (38) can be simplified to $\theta_* = (\Sigma + \gamma_2 v) \gamma_*^{-1} v$. By replacing $\theta_* = (\Sigma + \gamma_* v) \gamma_*^{-1} v$ in $A_\sigma$, along with equation $\text{SR}(\theta) = \sigma_*^2 + \theta^T \Sigma \theta = 2v^T \theta$ we get

$$\begin{align*}
A_* &= \sqrt{\text{SR}(\Sigma + \gamma_* v) \gamma_*^{-1} v} \\
&= \frac{1}{\|\Sigma + \gamma_* v\|_{\ell_2}^2} \left( \sigma_*^2 + \|\Sigma^{1/2} \Sigma + \gamma_* v\|_{\ell_2}^2 \right) \gamma_*^{-1} v \\
&= \frac{1}{\|\Sigma + \gamma_* v\|_{\ell_2}^2} \left( \sigma_*^2 + \|\Sigma^{1/2} \Sigma + \gamma_* v\|_{\ell_2}^2 \right) \gamma_*^{-1} v \\
&= \gamma_*
\end{align*}$$

therefore $\gamma_*$ is a fixed point solution of two equations (15) and (16). Moreover, definition of $A_*$ gives us $\text{SR}(\gamma_*) = A_*^2 \|\Sigma + \gamma_* v\|_{\ell_2}^{-1} v ||_{\ell_2}$. Next, from adversarial risk relation in Proposition A.1 we know that $\text{AR}(\theta_*) = (\sqrt{\text{SR}(\theta_*)} + \epsilon ||\theta_*||_3^2)^2$. This implies $\text{AR}(\theta_*) = (A_* + \epsilon^2 ||\Sigma + \gamma_* v||_{\ell_2}^{-1} v ||_{\ell_2}^2)$.
A.3. Proof of Corollary 3.4

For linear data model \( y = x^T \theta_0 + w \) with isotropic features \( \mathbb{E}[xx^T] = I_d \) and Gaussian noise \( w \sim \mathcal{N}(0, \sigma^2) \) we have \( \mathbb{E}[y] = \theta_0 \). In addition, we have \( \mathbb{E}[y^2] = \sigma^2 + \| \theta_0 \|^2_2 \).

This gives us \( \sigma_y^2 = \sigma^2 + \| \theta_0 \|^2_2 \). Use Theorem 3.3 with \( v = \theta_0, \Sigma = I, \) and \( \sigma_y^2 = \sigma^2 + \| \theta_0 \|^2_2 \) to get Corollary 3.4.

A.4. Proof of Proposition 3.5

We start by proving the expression for standard risk. By definition, for adversarial risk we use the dual form (8). Our next lemma characterizes the function \( \phi_\gamma \) given by (9) for the binary problem under the Gaussian mixture model.

**Lemma A.1.** Consider the binary classification problem under the Gaussian mixture model with 0-1 loss. Then, the robust surrogate for the loss function \( \phi_\gamma \) given by (9) for the binary problem under the Gaussian mixture model.

**Proof (Lemma A.1).** By definition of the \( \phi_\gamma \) function, for the setting of Lemma A.1 we have

\[
\phi_\gamma(\theta; z) = \sup_{x} \left\{ \mathbb{I}(y_0 x^T \theta \leq 0) - \frac{\gamma}{2} \|x - x_0\|^2_{\ell_2} \right\}.
\]

We let \( v_0 := y_0 x_0 \) and \( v = y_0 x \). Given that \( y_0 \in \{ \pm 1 \} \), the function \( \phi_\gamma \) can be written as

\[
\phi_\gamma(\theta; z_0) = \sup_{\nu} \left\{ \mathbb{I}(v^T \theta \leq 0) - \frac{\gamma}{2} \|v - v_0\|^2_{\ell_2} \right\}.
\]

First observe that by choosing \( x = x_0 \), we obtain \( \phi_\gamma(\theta; z_0) \geq 0 \). It is also clear that \( \phi_\gamma(\theta; z_0) \leq 1 \). We consider two cases.

**Case 1:** \( (v_0^T \theta \leq 0) \). By choosing \( v = v_0 \) we obtain that \( \phi_\gamma(\theta; z_0) \geq 1 \) and hence \( \phi_\gamma(\theta; z_0) = 1 \).

**Case 2:** \( (v_0^T \theta > 0) \). Let \( v_* \) be the maximizer in definition of \( \phi_\gamma(\theta; z_0) \). If \( v_*^T \theta > 0 \), then we have

\[
\phi_\gamma(\theta; z_0) = \mathbb{I}(v_0^T \theta \leq 0) - \frac{\gamma}{2} \|v_* - v_0\|^2_{\ell_2} = -\frac{\gamma}{2} \|v_* - v_0\|^2_{\ell_2} \leq 0.
\]

Therefore, \( \phi_\gamma(\theta; z_0) = 0 \) in this case. We next focus on the case that \( v_0^T \theta \leq 0 \). It is easy to see that in this case, \( v_* \) is the solution of the following optimization:

\[
\min_{v \in \mathbb{R}^d} \|v - v_0\|_{\ell_q} \quad \text{subject to} \quad v^T \theta \leq 0
\]

with \( u \sim \mathcal{N}(0, I_d) \) and \( v \sim \mathcal{N}(0, 1) \). To prove the expression for adversarial risk we use the dual form (8). Our next lemma characterizes the function \( \phi_\gamma \) given by (9) for the binary problem under the Gaussian mixture model.

\[
\mathbb{E}_{\mathcal{D}_2}[\phi_\gamma(\theta; z)] = \Phi\left(\frac{2}{b_0 \gamma} - a\right) + b_0 \gamma \left\{ \left(a_0 + \sqrt{\frac{2}{b_0 \gamma}}\right) \Phi\left(a_0 - \sqrt{\frac{2}{b_0 \gamma}}\right) - a_0 \Phi\left(a_0\right) + \left(a_0 + 1\right) \Phi\left(a_0 - \sqrt{\frac{2}{b_0 \gamma}}\right) - \Phi\left(a_0\right) \right\}.
\]

with \( a_0 = \frac{\mu^T \theta}{\|\Sigma^{1/2} \theta\|_{\ell_2}} \) and \( b_0 = \frac{\|\Sigma^{1/2} \theta\|_{\ell_2}}{\|\theta\|_{\ell_q}} \).

The function \( \phi_\gamma \) is then given by \( \phi_\gamma(\theta; z_0) = 1 - \frac{\gamma}{2} \left(\frac{v_0^T \theta}{\|\theta\|_{\ell_q}}\right)^2 \). Putting the two conditions \( v_0^T \theta \leq 0 \) and \( v_0^T \theta > 0 \) together, we obtain

\[
\phi_\gamma(\theta; z_0) = \max\left\{ 1 - \frac{\gamma}{2} \left(\frac{v_0^T \theta}{\|\theta\|_{\ell_q}}\right)^2, 0 \right\},
\]

in this case. Combining case 1 and case 2 we arrive at

\[
\phi_\gamma(\theta; z_0) = \mathbb{I}(v_0^T \theta \leq 0) + \max\left(1 - \frac{\gamma}{2} \left(\frac{v_0^T \theta}{\|\theta\|_{\ell_q}}\right)^2, 0\right) \mathbb{I}(v_0^T \theta > 0).
\]
Letting \( a_\theta := \frac{\mu^T \theta}{\|\Sigma^{1/2} \theta\|_2} \), (41) can be written as
\[
\phi_\gamma(\theta; z_0) = \mathbb{I}(\nu \leq -a_\theta) + \max \left( 1 - \frac{\gamma}{2} \frac{\|\Sigma^{1/2} \theta\|_2^2}{\|\theta\|_2^2}, 0 \right) \mathbb{I}(\nu > -a_\theta).
\]
and therefore the maximum of \( \varphi(t - a) - \varphi(a) + a (\Phi(t - a) - \Phi(a)) \) is achieved at \( t = 2a \). As a result \( d\mathcal{R}(a)/da \leq -\lambda \varphi(-a) < 0 \), which implies that the objective (25) is decreasing in \( a \). Since \( |a| \leq \|\mu\|_{\ell^2} \), its infimum is achieved at \( a = \|\mu\|_{\ell^2} \).

Equations (26) follows from (24) by substituting for \( a_\theta = \|\mu\|_{\ell^2} \) and \( b_\theta = 1 \).

### A.6. Proof of Corollary 3.8

Recall the distance \( d(\cdot, \cdot) \) on the space \( Z = \{ z = (x, y), x \in \mathbb{R}^d, y \in \mathbb{R} \} \) given by \( d(z, \tilde{z}) = \|x - \tilde{x}\|_2 + \infty \cdot \mathbb{I}(y < \tilde{y}) \). This metric is induced from norm \( ||z|| = \|x\|_{\ell^2} + \infty \cdot \mathbb{I}(y = 0) \) with corresponding conjugate norm \( ||z||_* = \|x\|_{\ell^2} \). We will use Proposition 2.3 to find the variation of loss \( \ell \) and derive the first-order approximation for the Wasserstein adversarial risk. Denoting by \( u_j \in \mathbb{R}^d \) be the \( j \)th row of matrix \( U \), for \( j = 1, 2, ..., N \), we have
\[
\nabla_x \ell(\theta; Z) = \nabla_x (y - \theta^T \sigma(Ux))^2 = 2(\theta^T \sigma(Ux) - y) \sum_{j=1}^N \theta_j \sigma'(u_j^T x)u_j = 2(\theta^T \sigma(Ux) - y)U^T \text{diag}(\sigma'(Ux))\theta .
\]

As we work with Wasserstein of order \( \rho = 2 \), we have conjugate order \( q = 2 \). Therefore, Proposition 2.3 gives us
\[
\mathcal{V}_{\rho, q}(\ell) = \left( \mathbb{E} \left[ ||\nabla_x \ell(\theta; Z)||_2^2 \right] \right)^{1/2}. 
\]
By using (46) we get
\[
\mathcal{V}_{\rho, q}(\ell) = 2 \left( \mathbb{E} \left[ (\theta^T \sigma(Ux) - y)^2 \left\| U^T \text{diag}(\sigma'(Ux))\theta \right\|_{\ell^2}^2 \right] \right)^{1/2}.
\]

Finally, relation AR(\( \theta \)) = SR(\( \theta \)) + \( \epsilon \mathcal{V}_{\rho, q}(\ell) + O(\epsilon^2) \) from Proposition 2.3 completes the proof. We just need to verify that the necessary condition in Proposition 2.3 holds for the loss \( \ell(\theta; z) = (y - \theta^T \sigma(Wx))^2 \). By the setting of the problem, we have \( x \in \mathbb{S}^{d-1} \cdot (\sqrt{d}) \) and \( u_j \in \mathbb{S}^{d-1}(1) \).

Therefore \( ||x||_{\ell^2} \leq \sqrt{d} \) and \( ||U||_{op} \leq \sqrt{\max(N, d)} \).

In the following lemma we show that the solution \( \theta_\lambda \) to (14) is bounded as \( \lambda \) varies in \( (0, \infty) \).

### Lemma A.2. Under the setting of Corollary 3.8, and for \( \theta_\lambda \) given by (14), there exist constants \( c_0 \) and \( c_1 \), independent of \( \lambda \), such that with probability at least \( 1 - e^{-c_0d} \) we have \( ||\theta_\lambda||_{\ell^2} \leq c_1 \).

Using Lemma A.2 we can restrict ourselves to the ball of \( \ell^2 \) radius \( c_1 \). More specifically, we can define a ‘surrogate’ loss for (14) where \( \theta \) is constrained to be in ball of radius \( c_1 \), without changing its solution. We can then apply Proposition 2.3 to establish a relation between SR and AR. In the following part we show that the conditions of Proposition 2.3 are satisfied.

We adopt the shorthands \( D = \text{diag}(\sigma'(Ux)) \), \( \tilde{D} = \)
\[ \frac{1}{2} \left\| \nabla_{\theta} \ell (\theta; z) - \nabla_{\theta} \ell (\theta; \tilde{z}) \right\|_2^2 \]

\[ = \frac{1}{2} \left\| \nabla_{\theta} \ell (\theta; z) - \nabla_{\theta} \ell (\theta; \tilde{z}) \right\|_2^2 \]

\[ \leq \left\| \theta^T s - y \right\|_{\{2\}}^2 + \left\| \theta^T (s - \tilde{s}) \right\|_{\{2\}}^2 + \left\| y \theta - \tilde{y} \theta \right\|_{\{2\}}^2 \]

\[ \leq Nc_1^2 + \sqrt{N}c_1^2 \left\| s - \tilde{s} \right\|_{\{2\}}^2 + \sqrt{N}c_1^2 + \sqrt{N}c_1 \left\| y - \tilde{y} \right\|_{\{2\}}^2 \]

\[ \leq \left\| y \theta - \tilde{y} \theta \right\|_{\{2\}}^2 + \left\| y \theta - \tilde{y} \theta \right\|_{\{2\}}^2 + \left\| Nc_1 \left\{ \| x - \tilde{x} \|_{\{2\}}^2 + \infty \right\} \right\|_{\{2\}}^2 \]

\[ \leq M + L \left\| z - \tilde{z} \right\|_{\{2\}}^2 \]

where (a) comes from (46), in (b) we used triangle inequality, (c) is a direct result of Cauchy inequality and the fact that \( \sigma(u) \leq u \), (d) comes from Lipschitz continuity of \( \sigma \), and in (e) we used \( C = (N + \sqrt{N})c_1^2 \) and \( L = Nc_1^2 \). Therefore the necessary condition in Proposition 2.3 is satisfied.

A.6.1. PROOF OF LEMMA A.2

By comparing the objective value (14) at \( \theta_\lambda \) and 0 and using the optimality of \( \theta_\lambda \), we get

\[ (1 + \lambda)SR(\theta_\lambda) \leq (1 + \lambda)SR(0) + 2\varepsilon \mathbb{E}_x \left[ (f_d(x) - \theta_\lambda^T \sigma(Ux))^2 \right] \]

\[ \times \left\| U \right\|_{\{2\}}^2 \left\| \text{diag}(\sigma'(U\tilde{x})) \right\|_{\{2\}}^2 \]

\[ \leq (1 + \lambda)SR(0) . \]

Therefore by invoking (31) we get

\[ \mathbb{E}_x \left[ (f_d(x) - \theta_\lambda^T \sigma(Ux))^2 \right] \leq \mathbb{E}_x \left[ f_d(x)^2 \right] \]

Using the inequality \( (a - b)^2 \geq \frac{a^2}{2} - b^2 \), we get

\[ \mathbb{E}[(\theta_\lambda^T \sigma(Ux))^2] \leq 4\mathbb{E}_x[f_d(x)^2] < c_2 , \]

with probability at least \( 1 - e^{-c_3 d} \) for some constants \( c_2, c_3 > 0 \). We next lower bound the eigenvalues of \( \mathbb{E}[\sigma(Ux) \sigma(Ux)^T] \) from which we can upper bound \( \| \theta_\lambda \|_{\{2\}}^2 \).

Define the dual activation of \( \sigma \) as

\[ \tilde{\sigma}(\rho) = \mathbb{E}_{(v, w) \sim N_\rho \mid \sigma(v) \sigma(w)} \]

where \( N_\rho \) denotes the two dimensional Gaussian with mean zero and covariance \( \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \). With this definition, we have

\[ \mathbb{E}[(\sigma(Ux) \sigma(Ux)^T)_{ij}] = \tilde{\sigma}(u_i^T u_j) \text{ for } i, j = 1, \ldots, N. \]

Let \( \{ a_r \}_{r=0}^\infty \) denote the Hermite coefficients defined by

\[ a_r := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sigma(g) h_r(g) e^{-\frac{g^2}{2}} dg , \]

where \( h_r(g) \) is the normalized Hermite polynomial defined by

\[ h_r(x) := \frac{1}{\sqrt{\pi r!}} \frac{d^r}{dx^r} e^{-\frac{x^2}{2}} . \]

Using the properties of normalized Hermite polynomials we have

\[ \tilde{\sigma}(\rho) = \mathbb{E}_{(v, w) \sim N_\rho} \left[ \sum_{r=0}^{\infty} a_r h_r(v) \right] \left[ \sum_{r=0}^{\infty} a_r h_r(w) \right] = \sum_{r=0}^{\infty} a_r^2 \rho^r . \]

Writing in matrix form we obtain

\[ \mathbb{E}[(\sigma(Ux) \sigma(Ux)^T)] = \tilde{\sigma}(U U^T) = \sum_{r=0}^{\infty} a_r^2 (UU^T)^{\circ r} , \]

(50)

where for a matrix \( A^{\circ r} = A \odot (A^{(r-1)}) \) with \( \odot \) denoting the Hadamard product (entrywise product).

We next use the identity \( (AA^T) \odot (BB^T) = (A \ast B)(A \ast B)^T \), with * indicating the Khatri-Rao product. By using this identity and applying induction on r it is straightforward to get the following relation for any matrix \( A \):

\[ (AA^T)^{\circ r} = (A^{*)r})(A^{*)r)^T , \]

(51)

with \( A^{*)r} = A \ast (A^{(*r-1)}) \). By using the above identity in Equation (50) we obtain

\[ \mathbb{E}[(\sigma(Ux) \sigma(Ux)^T)] = \sum_{r=0}^{\infty} a_r^2 (UU^T)^{\circ r} \]

\[ = \sum_{r=0}^{\infty} (a_r U^{*)r})(a_r U^{*)r)^T \]

\[ \geq a_r^2 (U^{*)r})(U^{*)r)^T , \]

for any \( r \geq 0 \). Using this bound with \( r = 2 \) and the fact that \( a_2 = \frac{1}{2\sqrt{\pi}} \) for ReLU activation, we get

\[ \mathbb{E}[(\sigma(Ux) \sigma(Ux)^T)] \geq \frac{1}{4\pi} (U \ast U) \geq c_4 , \]

(52)

where the last step holds with probability at least \( 1 - e^{-c_5 d} \) for some constants \( c_4 \) and \( c_5 \) using the result of (7)Corollary 7.5)soltanolkotabi2018theoretical.

Combining Equations (48) and (52) gives us \( \| \theta_\lambda \|_{\{2\}} \leq \sqrt{c_2 / c_4} \), which completes the proof.