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

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

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

$$\begin{aligned} g_\gamma(x_0 + c\theta) &= g_\gamma(x_0) + 2c(x_0^\mathsf{T}\theta - y_0) \|\theta\|_{\ell_2}^2 \\ &\quad + c^2 \|\theta\|_{\ell_2}^2 (\|\theta\|_{\ell_2}^2 - \gamma), \end{aligned}$$

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 $\nabla^2 g_\gamma(x) = 2(\theta\theta^\mathsf{T} - \gamma I)$ we realize that g_γ is concave. Writing the first order optimal condition we have

$$(y_0 - x^\mathsf{T}\theta)\theta + \gamma(x - x_0) = 0. \tag{34}$$

Multiplying by θ and solving for $x^\mathsf{T}\theta$, we get

$$x^\mathsf{T}\theta = \frac{\gamma x_0^\mathsf{T}\theta - y_0\|\theta\|^2}{\gamma - \|\theta\|^2}.$$

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

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

Replacing x^* in g_γ yields

$$\phi_\gamma(\theta; z) = \begin{cases} +\infty & \text{if } \gamma < \|\theta\|_{\ell_2}^2, \\ \frac{\gamma(y_0 - x^\mathsf{T}\theta)^2}{(\gamma - \|\theta\|_{\ell_2}^2)} & \text{if } \gamma \geq \|\theta\|_{\ell_2}^2. \end{cases} \tag{35}$$

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

$$\begin{aligned} \mathsf{AR}(\theta) &:= \sup_{\mathbb{Q} \in \mathcal{U}_\varepsilon(\mathbb{P}_Z)} \mathbb{E}_{z\sim\mathbb{Q}}[\ell(\theta; z)] \\ &= \inf_{\gamma \geq 0} \{\gamma\varepsilon^2 + \mathbb{E}_{\mathbb{P}_Z}[\phi_\gamma(\theta; z)]\} \\ &= \inf_{\gamma \geq \|\theta\|_{\ell_2}^2} \{\gamma\varepsilon^2 + \mathbb{E}_{\mathbb{P}_Z}[\phi_\gamma(\theta; z)]\} \\ &= \inf_{\gamma \geq \|\theta\|_{\ell_2}^2} \{\gamma\varepsilon^2 + \frac{\gamma \mathbb{E}_{\mathbb{P}_Z}[\ell(\theta; z)]}{\gamma - \|\theta\|_{\ell_2}^2}\}, \end{aligned}$$

the infimum is achieved at

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

Finally, this gives us

$$\mathsf{AR}(\theta) = \left(\sqrt{\mathbb{E}_{\mathbb{P}_Z}[\ell(\theta; z)]} + \varepsilon \|\theta\|_{\ell_2} \right)^2.$$

A.2. Proof of Theorem 3.3

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

$$\mathcal{R}(\theta) = (1+\lambda)\mathsf{SR}(\theta) + \varepsilon^2\|\theta\|_{\ell_2}^2 + 2\varepsilon\|\theta\|_{\ell_2}\sqrt{\mathsf{SR}(\theta)}. \tag{36}$$

It is easy to see $\mathsf{SR}(\theta) = \sigma_y^2 + \theta^\mathsf{T}\Sigma\theta - 2v^\mathsf{T}\theta$. Replace $\nabla_\theta \mathsf{SR}(\theta) = 2(\Sigma\theta - v)$ in (36) to get

$$\begin{aligned} \nabla_\theta \mathcal{R}(\theta) &= 2(1+\lambda)(\Sigma\theta - v) + 2\varepsilon^2\theta \\ &\quad + 2\varepsilon \left(\frac{\theta}{\|\theta\|_{\ell_2}}\sqrt{\mathsf{SR}(\theta)} + (\Sigma\theta - v)\frac{\|\theta\|_{\ell_2}}{\sqrt{\mathsf{SR}(\theta)}} \right), \end{aligned} \tag{37}$$

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

$$\begin{aligned} &\left(\left(1 + \lambda + \frac{\varepsilon\|\theta_*\|_{\ell_2}}{\sqrt{\mathsf{SR}(\theta_*)}} \right)\Sigma + \left(\varepsilon^2 + \frac{\varepsilon\sqrt{\mathsf{SR}(\theta_*)}}{\|\theta_*\|_{\ell_2}} \right) I \right)\theta_* \\ &= \left(1 + \lambda + \frac{\varepsilon\|\theta_*\|_{\ell_2}}{\sqrt{\mathsf{SR}(\theta_*)}} \right) v. \end{aligned} \tag{38}$$

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

$$\begin{aligned} A_* &= \frac{\sqrt{\mathsf{SR}((\Sigma + \gamma_* I)^{-1}v)}}{\|(\Sigma + \gamma_* I)^{-1}v\|_{\ell_2}} \\ &= \frac{1}{\|(\Sigma + \gamma_* I)^{-1}v\|_{\ell_2}} \Big(\sigma_y^2 + \left\|\Sigma^{1/2}(\Sigma + \gamma_* I)^{-1}v\right\|_{\ell_2}^2 \\ &\qquad - 2v^\mathsf{T}(\Sigma + \gamma_* I)^{-1}v \Big)^{1/2}, \end{aligned}$$

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

A.3. Proof of Corollary 3.4

For linear data model $y = x^\mathsf{T}\theta_0 + w$ with isotropic features $\mathbb{E}[xx^T] = I_d$ and Gaussian noise $w \sim \mathsf{N}(0, \sigma^2)$ we have $\mathbb{E}[xy] = \theta_0$. In addition, we have $\mathbb{E}[y^2] = \sigma^2 + \|\theta_0\|_{\ell_2}^2$. This gives us $\sigma_y^2 = \sigma^2 + \|\theta_o\|_{\ell_2}^2$. Use Theorem 3.3 with $v = \theta_0$, $\Sigma = I$, and $\sigma_y^2 = \sigma^2 + \|\theta_0\|_{\ell_2}^2$ to get Corollary 3.4.

A.4. Proof of Proposition 3.5

We start by proving the expression for standard risk. By definition we have

$$
\begin{aligned}
\mathsf{SR}(\theta) &:= \mathbb{E}[\mathbb{I}(y \neq \hat{y})] = \mathbb{P}(yx^\mathsf{T}\theta \le 0) \\
&= \mathbb{P}\left(y(y\mu + \Sigma^{1/2}u)^\mathsf{T}\theta \le 0\right) \\
&= \mathbb{P}\left((\mu + \Sigma^{1/2}u)^\mathsf{T}\theta \le 0\right) \\
&= \mathbb{P}\left(\mu^\mathsf{T}\theta + \left\|\Sigma^{1/2}\theta\right\|_{\ell_2} \nu \le 0\right) \\
&= \Phi\left(-\frac{\mu^\mathsf{T}\theta}{\left\|\Sigma^{1/2}\theta\right\|_{\ell_2}}\right), \qquad (39)
\end{aligned}
$$

with $u \sim \mathsf{N}(0, I_d)$ and $\nu \sim \mathsf{N}(0,1)$. To prove the expression for adversarial risk we use the dual form (8). Our next lemma characterizes the function ϕ_γ 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 ϕ_γ given by (9) with distance $d(\cdot,\cdot)$ (12) satisfies*

$$
\begin{aligned}
&\mathbb{E}_{\mathbb{P}_Z}[\phi_\gamma(\theta; z)] = \Phi\Big(\sqrt{\frac{2}{b_\theta\gamma}} - a\Big) \\
&+ \frac{b_\theta\gamma}{2}\Big\{\Big(a_\theta + \sqrt{\frac{2}{b_\theta\gamma}}\Big)\varphi\Big(a_\theta - \sqrt{\frac{2}{b_\theta\gamma}}\Big) \\
&\qquad - a_\theta\varphi(a_\theta) + (a_\theta^2 + 1)\Big[\Phi\Big(a_\theta - \sqrt{\frac{2}{b_\theta\gamma}}\Big) - \Phi(a_\theta)\Big]\Big\},
\end{aligned}
$$

with $a_\theta = \frac{\mu^\mathsf{T}\theta}{\left\|\Sigma^{1/2}\theta\right\|_{\ell_2}}$ and $b_\theta = \frac{\left\|\Sigma^{1/2}\theta\right\|_{\ell_2}^2}{\|\theta\|_{\ell_q}^2}$.

Proof (Lemma A.1). By definition of the ϕ_γ function, for the setting of Lemma A.1 we have

$$
\phi_\gamma(\theta; z_0) = \sup_x \{\mathbb{I}(y_0 x^\mathsf{T}\theta \le 0) - \frac{\gamma}{2}\|x - x_0\|_{\ell_r}^2\}.
$$

We let $v_0 := y_0x_0$ and $v = y_0x$. Given that $y_0 \in \{\pm 1\}$, the function ϕ_γ can be written as

$$
\phi_\gamma(\theta; z_0) = \sup_v \{\mathbb{I}(v^\mathsf{T}\theta \le 0) - \frac{\gamma}{2}\|v - v_0\|_{\ell_r}^2\}.
$$

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

Case 1: $(v_0^\mathsf{T}\theta \le 0)$. By choosing $v = v_0$ we obtain that $\phi_\gamma(\theta; z_0) \ge 1$ and hence $\phi_\gamma(\theta; z_0) = 1$.

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

$$
\begin{aligned}
\phi_\gamma(\theta; z_0) &= \mathbb{I}(v_*^\mathsf{T}\theta \le 0) - \frac{\gamma}{2}\|v_* - v_0\|_{\ell_r}^2 \\
&= -\frac{\gamma}{2}\|v_* - v_0\|_{\ell_r}^2 \le 0\,.
\end{aligned}
$$

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

$$
\begin{aligned}
&\min_{v \in \mathbb{R}^d} \quad \|v - v_0\|_{\ell_r} \\
&\text{subject to} \quad v^\mathsf{T}\theta \le 0 \qquad (40)
\end{aligned}
$$

Given that $v_0^\mathsf{T}\theta > 0$ by assumption, using the Holder inequality it is straightforward to see that the optimal value is given by $\|v - v_0\|_{\ell_r} = \frac{v_0^\mathsf{T}\theta}{\|\theta\|_{\ell_q}}$, with $\frac{1}{r} + \frac{1}{q} = 1$.

The function ϕ_γ is then given by $\phi_\gamma(\theta; z_0) = 1 - \frac{\gamma}{2}\Big(\frac{v_0^\mathsf{T}\theta}{\|\theta\|_{\ell_q}}\Big)^2$. Putting the two conditions $v_*^\mathsf{T}\theta \le 0$ and $v_0^\mathsf{T}\theta > 0$ together, we obtain

$$
\phi_\gamma(\theta; z_0) = \max\Big\{1 - \frac{\gamma}{2}\Big(\frac{v_0^\mathsf{T}\theta}{\|\theta\|_{\ell_q}}\Big)^2, 0\Big\},
$$

in this case.

Combining case 1 and case 2 we arrive at

$$
\begin{aligned}
\phi_\gamma(\theta; z_0) &= \mathbb{I}(v_0^\mathsf{T}\theta \le 0) \qquad (41) \\
&\quad + \max\left(1 - \frac{\gamma}{2}\Big(\frac{v_0^\mathsf{T}\theta}{\|\theta\|_{\ell_q}}\Big)^2, 0\right)\mathbb{I}(v_0^\mathsf{T}\theta > 0)\,. \qquad (42)
\end{aligned}
$$

For (x_0, y_0) generated according to the Gaussian mixture model, we have $v_0^\mathsf{T}\theta = y_0x_0^\mathsf{T}\theta = \mu^\mathsf{T}\theta + \left\|\Sigma^{1/2}\theta\right\|_{\ell_2}\nu$ with $\nu \sim \mathsf{N}(0,1)$. Hence,

$$
\Big|\frac{v_0^\mathsf{T}\theta}{\|\theta\|_{\ell_q}}\Big| = \Big|\frac{\mu^\mathsf{T}\theta}{\|\theta\|_{\ell_q}} + \frac{\left\|\Sigma^{1/2}\theta\right\|_{\ell_2}}{\|\theta\|_{\ell_q}}\nu\Big|.
$$

Letting $a_\theta := \frac{\mu^\mathsf{T}\theta}{\left\|\Sigma^{1/2}\theta\right\|_{\ell_2}}$, (41) can be written as

$$\begin{aligned}
\phi_\gamma(\theta; z_0) &= \mathbb{I}(\nu \le -a_\theta) \\
&\quad + \max\Big(1 - \frac{\gamma}{2}\frac{\left\|\Sigma^{1/2}\theta\right\|_{\ell_2}^2}{\|\theta\|_{\ell_q}^2}(\nu + a_\theta)^2, 0\Big) \cdot \mathbb{I}(\nu > -a_\theta) \\
&= \mathbb{I}(\nu \le -a_\theta) \\
&\quad + \left(1 - \frac{b_\theta\gamma}{2}(\nu + a_\theta)^2\right)\mathbb{I}\left(\sqrt{\frac{2}{b_\theta\gamma}} - a_\theta > \nu > -a_\theta\right),
\end{aligned} \tag{43}$$

where $b_\theta := \frac{\|\Sigma^{1/2}\theta\|_{\ell_2}^2}{\|\theta\|_{\ell_q}^2}$. By simple algebraic calculation, we get

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}_Z}[\phi_\gamma(\theta; z)] &= \Phi\Big(\sqrt{\frac{2}{b_\theta\gamma}} - a_\theta\Big) \\
&\quad + \frac{b_\theta\gamma}{2}\Big\{\Big(a_\theta + \sqrt{\frac{2}{b_\theta\gamma}}\Big)\varphi\Big(a_\theta - \sqrt{\frac{2}{b_\theta\gamma}}\Big) - a_\theta\varphi(a_\theta) \\
&\qquad + (a_\theta^2 + 1)\Big[\Phi\Big(a_\theta - \sqrt{\frac{2}{b_\theta\gamma}}\Big) - \Phi(a_\theta)\Big]\Big\}.
\end{aligned}$$

$\square$

The claim of Proposition 3.5 follows readily from Lemma A.1 and the fact that strong duality holds for the dual problem (8), where we use the change of variable $\gamma \mapsto \frac{\gamma}{b_\theta}$.

A.5. Proof of Remark 3.7

Recall the objective (25) and define

$$\begin{aligned}
\mathcal{R}(a) :=& \lambda\Phi(-a) + \gamma\varepsilon^2 + \Phi\left(\sqrt{\frac{2}{\gamma}} - a\right) \\
&+ \frac{\gamma}{2}\Big\{(a + \sqrt{\frac{2}{\gamma}})\varphi\left(a - \sqrt{\frac{2}{\gamma}}\right) - a\varphi(a) \\
&\quad + (a^2 + 1)\left(\Phi\left(a - \sqrt{\frac{2}{\gamma}}\right) - \Phi(a)\right)\Big\}.
\end{aligned}$$

Then, we get

$$\begin{aligned}
\frac{d\mathcal{R}(a)}{da} =& -\lambda\varphi(-a) \\
&+ \gamma\Big\{\varphi\left(\sqrt{\frac{2}{\gamma}} - a\right) - \varphi(a) \\
&\quad + a\left(\Phi\left(\sqrt{\frac{2}{\gamma}} - a\right) - \Phi(a)\right)\Big\}.
\end{aligned} \tag{44}$$

Note that

$$\begin{aligned}
&\frac{\partial}{\partial t}\Big(\varphi(t-a) - \varphi(a) + a\left(\Phi(t-a) - \Phi(a)\right)\Big) \\
&= \varphi(t-a)(2a - t),
\end{aligned} \tag{45}$$

and therefore the maximum of $\varphi(t-a) - \varphi(a) + a\left(\Phi(t-a) - \Phi(a)\right)$ is achieved at $t = 2a$. As a result $\frac{d\mathcal{R}(a)}{da} \le -\lambda\varphi(-a) < 0$, which implies that the objective (25) is decreasing in a. Since $|a| \le \|\mu\|_{\ell_2}$, 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 $\mathcal{Z} = \{z = (x, y),\ x \in \mathbb{R}^d, y \in \mathbb{R}\}$ 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 ℓ and derive the first-order approximation for the Wasserstein adversarial risk. Denoting by $u_j \in \mathbb{R}^d$ be the jth row of matrix U, for $j = 1, 2, ..., N$, we have

$$\begin{aligned}
\nabla_x \ell(\theta; Z) &= \nabla_x (y - \theta^\mathsf{T}\sigma(Ux))^2 \\
&= 2(\theta^\mathsf{T}\sigma(Ux) - y)\sum_{j=1}^N \theta_j \sigma'(u_j^\mathsf{T}x)u_j \\
&= 2(\theta^\mathsf{T}\sigma(Ux) - y)U^\mathsf{T}\mathrm{diag}(\sigma'(Ux))\theta.
\end{aligned} \tag{46}$$

As we work with Wasserstein of order $p = 2$, we have conjugate order $q = 2$. Therefore, Proposition 2.3 gives us $V_{P_Z,q}(\ell) = \left(\mathbb{E}[||\nabla_z\ell(\theta; Z)||_*^2]\right)^{1/2}$. By using (46) we get

$$V_{P_Z,q}(\ell) = 2\left(\mathbb{E}\left[(\theta^\mathsf{T}\sigma(Ux) - y)^2\left\|U^\mathsf{T}\mathrm{diag}(\sigma'(Ux))\theta\right\|_{\ell_2}^2\right]\right)^{1/2}.$$

Finally, relation $\mathsf{AR}(\theta) = \mathsf{SR}(\theta) + \varepsilon V_{P_Z,q}(\ell) + O(\varepsilon^2)$ from Proposition 2.3 completes the proof. We just need to verify that the necessary condition in Proposition 2.3 holds for the loss $\ell(\theta; z) = (y - \theta^\mathsf{T}\sigma(Wx))^2$. By the setting of the problem, we have $x \in \mathbb{S}^{d-1}(\sqrt{d})$ and $u_j \in \mathbb{S}^{d-1}(1)$. Therefore $\|x\|_{\ell_2} \le \sqrt{d}$ and $\|U\|_{\mathrm{op}} \le \sqrt{\max(N, d)}$.

In the following lemma we show that the solution θ_λ to (14) is bounded as λ varies in $[0, \infty)$.

Lemma A.2. *Under the setting of Corollary 3.8, and for θ_λ given by (14), there exist constants c_0 and c_1, independent of λ, such that with probability at least $1 - e^{-c_0 d}$ we have $\|\theta_\lambda\|_{\ell_2} \le c_1$.*

Using Lemma A.2 we can restrict ourselves to the ball of ℓ_2 radius c_1. More specifically, we can define a 'surrogate' loss for (14) where θ 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 $D = \mathrm{diag}(\sigma'(Ux))$, $\tilde{D} =$

$\mathrm{diag}(\sigma'(U\tilde{x}))$, $s=\sigma(Ux)$, and $\tilde{s}=\sigma(U\tilde{x})$, and write

$$
\begin{aligned}
&\frac{1}{2}\|\nabla_z\ell(\theta;z)-\nabla_z\ell(\theta;\tilde{z})\|_*\\
&=\frac{1}{2}\left\|\nabla_x\ell(\theta;z)-\nabla_x\ell(\theta;\tilde{z})\right\|_{\ell_2}\\
&\overset{(a)}{=}\left\|(\theta^\mathsf{T}s-y)U^\mathsf{T}D\theta-(\theta^\mathsf{T}\tilde{s}-\tilde{y})U^\mathsf{T}\tilde{D}\theta\right\|_{\ell_2}\\
&\overset{(b)}{\le}\left\|\theta^\mathsf{T}sU^\mathsf{T}(D-\tilde{D})\theta\right\|_{\ell_2}+\left\|\theta^\mathsf{T}(s-\tilde{s})U^\mathsf{T}\tilde{D}\theta\right\|_{\ell_2}\\
&+\left\|yU^\mathsf{T}(D-\tilde{D})\theta\right\|_{\ell_2}+\left\|(y-\tilde{y})U^\mathsf{T}\tilde{D}\theta\right\|_{\ell_2}\\
&\overset{(c)}{\le}Nc_1^2+\sqrt{N}c_1^2\left\|s-\tilde{s}\right\|_{\ell_2}+\sqrt{N}c_1^2+\sqrt{N}c_1|y-\tilde{y}|\\
&\overset{(d)}{\le}(N+\sqrt{N})c_1^2+Nc_1^2\left\|x-\tilde{x}\right\|_{\ell_2}+\sqrt{N}c_1|y-\tilde{y}|\\
&\le(N+\sqrt{N})c_1^2+Nc_1^2\left(\|x-\tilde{x}\|_{\ell_2}+\infty\,\mathbb{I}_{\{y\neq\tilde{y}\}}\right)\\
&\overset{(e)}{\le}M+L\|z-\tilde{z}\|\,,
\end{aligned}
$$

where (a) comes from (46), in (b) we used triangle inequality, (c) is a direct result of Cauchy inequality and the fact that $\sigma(u)\le u$, (d) comes from Lipschitz continuity of σ, 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 θ_λ and 0 and using the optimality of θ_λ we get

$$
\begin{aligned}
&(1+\lambda)\mathsf{SR}(\theta_\lambda)\\
&\le(1+\lambda)\mathsf{SR}(\theta_\lambda)\\
&\quad+2\varepsilon\,\mathbb{E}_x\Big[\left[(f_d(x)-\theta_\lambda^\mathsf{T}\sigma(Ux))^2+\sigma^2\right]\\
&\qquad\qquad\times\left\|U^\mathsf{T}\mathrm{diag}(\sigma'(Ux))\theta_\lambda\right\|_{\ell_2}^2\Big]^{1/2}\\
&\le(1+\lambda)\mathsf{SR}(0)\,.
\end{aligned}
$$

Therefore by invoking (31) we get

$$\mathbb{E}_x\left[(f_d(x)-\theta_\lambda^\mathsf{T}\sigma(Ux))^2\right]\le\mathbb{E}_x\left[f_d(x)^2\right] \tag{47}$$

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

$$\mathbb{E}[(\theta_\lambda^\mathsf{T}\sigma(Ux))^2]\le 4\mathbb{E}_x[f_d(x)^2]<c_2\,, \tag{48}$$

with probability at least $1-e^{-c_3 d}$ for some constants $c_2,c_3>0$. We next lower bound the eigenvalues of $\mathbb{E}[\sigma(Ux)\sigma(Ux)^\mathsf{T}]$ from which we can upper bound $\|\theta_\lambda\|_{\ell_2}$.

Define the dual activation of σ as

$$\tilde{\sigma}(\rho)=\mathbb{E}_{(v,w)\sim\mathsf{N}_\rho}[\sigma(v)\sigma(w)]$$

where N_ρ denotes the two dimensional Gaussian with mean zero and covariance $\begin{pmatrix}1&\rho\\\rho&1\end{pmatrix}$. With this definition, we have $\mathbb{E}[(\sigma(Ux)\sigma(Ux)^\mathsf{T})_{ij}]=\tilde{\sigma}(u_i^\mathsf{T}u_j)$ for $i,j=1,\dots,N$. Let $\{a_r\}_{r=0}^\infty$ denote the Hermite coefficients defined by

$$a_r:=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\sigma(g)h_r(g)e^{-\frac{g^2}{2}}\mathrm{d}g\,,$$

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

$$h_r(x):=\frac{1}{\sqrt{r!}}(-1)^r e^{\frac{x^2}{2}}\frac{\mathrm{d}^r}{\mathrm{d}x^r}e^{-\frac{x^2}{2}}\,.$$

Using the properties of normalized Hermite polynomials we have

$$\tilde{\sigma}(\rho)=\mathbb{E}_{(v,w)\sim\mathsf{N}_\rho}\Big[(\sum_{r=0}^\infty a_r h_r(v))(\sum_{\tilde{r}=0}^\infty a_{\tilde{r}}h_{\tilde{r}}(u))\Big]=\sum_{r=0}^\infty a_r^2\rho^r. \tag{49}$$

Writing in matrix form we obtain

$$\mathbb{E}[(\sigma(Ux)\sigma(Ux)^\mathsf{T})]=\tilde{\sigma}(UU^\mathsf{T})=\sum_{r=0}^\infty a_r^2(UU^\mathsf{T})^{\odot r}\,, \tag{50}$$

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

We next use the identity $(AA^\mathsf{T})\odot(BB^\mathsf{T})=(A*B)(A*B)^\mathsf{T}$, 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:

$$(AA^\mathsf{T})^{\odot r}=(A^{*r})(A^{*r})^\mathsf{T}\,, \tag{51}$$

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

$$
\begin{aligned}
\mathbb{E}[(\sigma(Ux)\sigma(Ux)^\mathsf{T})]&=\sum_{r=0}^\infty a_r^2(UU^\mathsf{T})^{\odot r}\\
&=\sum_{r=0}^\infty(a_rU^{*r})(a_rU^{*r})^\mathsf{T}\\
&\succeq a_r^2(U^{*r})(U^{*r})^\mathsf{T}\,,
\end{aligned}
$$

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

$$\mathbb{E}[(\sigma(Ux)\sigma(Ux)^\mathsf{T})]\succeq\frac{1}{4\pi}(U*U)\ge c_4\,, \tag{52}$$

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 (**?**)Corollary 7.5]soltanolkotabi2018theoretical.

Combining Equations (48) and (52) gives us $\|\theta_\lambda\|_{\ell_2}\le\sqrt{c_2/c_4}$, which completes the proof.