A. Definitions of quantities in the main text

A.1. Full definitions of $\overline{\mathcal{U}}$, $\overline{\mathcal{T}}$, $\mathcal{A}_U$, and $\mathcal{A}_T$ in Proposition 2

We first define functions $m_1(\cdot), m_2(\cdot)$, which could be understood as the limiting partial Stieltjes transforms of $\boldsymbol{A}(\boldsymbol{q})$ (c.f. Definition 1).

Definition 3 (Limiting partial Stieltjes transforms)**.** *For $\xi \in \mathbb{C}_+$ and $\boldsymbol{q} \in \mathcal{Q}$ where*

$$\mathcal{Q} = \{(s_1, s_2, t_1, t_2, p) : |s_2 t_2| \le \mu_1^2 (1+p)^2/2\}, \tag{25}$$

define functions $\mathsf{F}_1(\,\cdot\,,\,\cdot\,;\xi;\boldsymbol{q},\psi_1,\psi_2,\mu_1,\mu_\star), \mathsf{F}_2(\,\cdot\,,\,\cdot\,;\xi;\boldsymbol{q},\psi_1,\psi_2,\mu_1,\mu_\star) : \mathbb{C}\times\mathbb{C}\to\mathbb{C}$ via:

$$\begin{aligned}
\mathsf{F}_1(m_1, m_2; \xi; \boldsymbol{q}, \psi_1, \psi_2, \mu_1, \mu_\star) &\equiv \psi_1\Big(-\xi + s_1 - \mu_\star^2 m_2 + \frac{(1+t_2 m_2)s_2 - \mu_1^2(1+p)^2 m_2}{(1+s_2 m_1)(1+t_2 m_2) - \mu_1^2(1+p)^2 m_1 m_2}\Big)^{-1},\\
\mathsf{F}_2(m_1, m_2; \xi; \boldsymbol{q}, \psi_1, \psi_2, \mu_1, \mu_\star) &\equiv \psi_2\Big(-\xi + t_1 - \mu_\star^2 m_1 + \frac{(1+s_2 m_1)t_2 - \mu_1^2(1+p)^2 m_1}{(1+t_2 m_2)(1+s_2 m_1) - \mu_1^2(1+p)^2 m_1 m_2}\Big)^{-1}.
\end{aligned}$$

Let $m_1(\,\cdot\,;\boldsymbol{q};\boldsymbol{\psi})\; m_2(\,\cdot\,;\boldsymbol{q};\boldsymbol{\psi}) : \mathbb{C}_+ \to \mathbb{C}_+$ be defined, for $\Im(\xi) \ge C$ a sufficiently large constant, as the unique solution of the equations

$$\begin{aligned}
m_1 &= \mathsf{F}_1(m_1, m_2; \xi; \boldsymbol{q}, \psi_1, \psi_2, \mu_1, \mu_\star),\\
m_2 &= \mathsf{F}_2(m_1, m_2; \xi; \boldsymbol{q}, \psi_1, \psi_2, \mu_1, \mu_\star)
\end{aligned} \tag{26}$$

subject to the condition $|m_1| \le \psi_1/\Im(\xi)$, $|m_2| \le \psi_2/\Im(\xi)$. Extend this definition to $\Im(\xi) > 0$ by requiring m_1, m_2 to be analytic functions in $\mathbb{C}_+$.

We next define the function $g(\cdot)$ that will be shown to be the limiting log determinant of $\boldsymbol{A}(\boldsymbol{q})$.

Definition 4 (Limiting log determinants)**.** *For $\boldsymbol{q} = (s_1, s_2, t_1, t_2, p)$ and $\boldsymbol{\psi} = (\psi_1, \psi_2)$, define*

$$\begin{aligned}
\Xi(\xi, z_1, z_2; \boldsymbol{q}; \boldsymbol{\psi}) \equiv\ & \log[(s_2 z_1 + 1)(t_2 z_2 + 1) - \mu_1^2 (1+p)^2 z_1 z_2] - \mu_\star^2 z_1 z_2\\
&+ s_1 z_1 + t_1 z_2 - \psi_1 \log(z_1/\psi_1) - \psi_2 \log(z_2/\psi_2) - \xi(z_1 + z_2) - \psi_1 - \psi_2.
\end{aligned} \tag{27}$$

Let $m_1(\xi;\boldsymbol{q};\boldsymbol{\psi}), m_2(\xi;\boldsymbol{q};\boldsymbol{\psi})$ be defined as the analytic continuation of solution of Eq. (26) as defined in Definition 3. Define

$$g(\xi; \boldsymbol{q}; \boldsymbol{\psi}) = \Xi(\xi, m_1(\xi;\boldsymbol{q};\boldsymbol{\psi}), m_2(\xi;\boldsymbol{q};\boldsymbol{\psi}); \boldsymbol{q}; \boldsymbol{\psi}). \tag{28}$$

We next give the definitions of $\overline{\mathcal{U}}$, $\overline{\mathcal{T}}$, $\mathcal{A}_U$, and $\mathcal{A}_T$.

Definition 5 ($\overline{\mathcal{U}}$, $\overline{\mathcal{T}}$, $\mathcal{A}_U$, and $\mathcal{A}_T$ in Proposition 2)**.** *For any $\lambda \in \Lambda_U$, define*

$$\begin{aligned}
\mathcal{A}_U(\lambda, \psi_1, \psi_2) &= -\lim_{u\to 0_+}\Big[\psi_1\left(F_1^2\mu_1^2\partial_{s_1 s_2} + F_1^2\partial_{s_1 p} + F_1^2\partial_{s_1 t_2} + \tau^2\partial_{s_1 t_1}\right) g(\boldsymbol{i}u;\boldsymbol{q};\boldsymbol{\psi})\Big|_{\boldsymbol{q}=\boldsymbol{q}_U}\Big],\\
\overline{\mathcal{U}}(\lambda, \psi_1, \psi_2) &= F_1^2 + \tau^2 - \lim_{u\to 0_+}\Big[\big(F_1^2\mu_1^2\partial_{s_2} + F_1^2\partial_p + F_1^2\partial_{t_2} + \tau^2\partial_{t_1}\big) g(\boldsymbol{i}u;\boldsymbol{q};\boldsymbol{\psi})\Big|_{\boldsymbol{q}=\boldsymbol{q}_U}\Big],\\
\mathcal{A}_T(\lambda, \psi_1, \psi_2) &= -\lim_{u\to 0_+}\Big[\psi_1\left(F_1^2\mu_1^2\partial_{s_1 s_2} + F_1^2\partial_{s_1 p} + F_1^2\partial_{s_1 t_2} + \tau^2\partial_{s_1 t_1}\right) g(\boldsymbol{i}u;\boldsymbol{q};\boldsymbol{\psi})\Big|_{\boldsymbol{q}=\boldsymbol{q}_T}\Big],\\
\overline{\mathcal{T}}(\lambda, \psi_1, \psi_2) &= F_1^2 + \tau^2 - \lim_{u\to 0_+}\Big[\big(F_1^2\mu_1^2\partial_{s_2} + F_1^2\partial_p + F_1^2\partial_{t_2} + \tau^2\partial_{t_1}\big) g(\boldsymbol{i}u;\boldsymbol{q};\boldsymbol{\psi})\Big|_{\boldsymbol{q}=\boldsymbol{q}_T}\Big],
\end{aligned}$$

where $\boldsymbol{q}_U = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, \psi_2, 0, 0), \boldsymbol{q}_T = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, 0, 0, 0)$.

In the following, we give a simplified expression for $\overline{\mathcal{U}}$ and $\mathcal{A}_U$.

Remark 2 (Simplification of $\overline{\mathcal{U}}$ and $\mathcal{A}_U$)**.** *Define $\zeta, \overline{\lambda}$ as the rescaled version of μ_1^2 and λ*

$$\zeta = \frac{\mu_1^2}{\mu_\star^2},\ \ \overline{\lambda} = \frac{\lambda}{\mu_\star^2}.$$

Let $m_1(\,\cdot\,;\boldsymbol{\psi})\ m_2(\,\cdot\,;\boldsymbol{\psi}) : \mathbb{C}_+ \to \mathbb{C}_+$ be defined, for $\Im(\xi) \geq C$ a sufficiently large constant, as the unique solution of the equations

$$
\begin{aligned}
m_1 &= \psi_1 \left[-\xi + (1-\overline{\lambda}\psi_1) - m_2 + \frac{\zeta(1-m_2)}{1+\zeta m_1 - \zeta m_1 m_2} \right]^{-1}, \\
m_2 &= -\psi_2 \left[\xi + \psi_2 - m_1 - \frac{\zeta m_1}{1+\zeta m_1 - \zeta m_1 m_2} \right]^{-1},
\end{aligned}
\tag{29}
$$

subject to the condition $|m_1| \leq \psi_1/\Im(\xi)$, $|m_2| \leq \psi_2/\Im(\xi)$. Extend this definition to $\Im(\xi) > 0$ by requiring m_1, m_2 to be analytic functions in $\mathbb{C}_+$. Let

$$
\begin{aligned}
\overline{m}_1 &= \lim_{u\to\infty} m_1(\boldsymbol{i}u, \boldsymbol{\psi}), \\
\overline{m}_2 &= \lim_{u\to\infty} m_2(\boldsymbol{i}u, \boldsymbol{\psi}).
\end{aligned}
$$

Define

$$
\begin{aligned}
\chi_1 &= \overline{m}_1\zeta - \overline{m}_1\overline{m}_2\zeta + 1, \\
\chi_2 &= \overline{m}_1 - \psi_2 + \frac{\overline{m}_1\zeta}{\chi_1}, \\
\chi_3 &= \overline{\lambda}\psi_1 + \overline{m}_2 - 1 + \frac{\zeta\left(\overline{m}_2 - 1\right)}{\chi_1}.
\end{aligned}
$$

Define two polynomials $\mathcal{E}_1, \mathcal{E}_2$ as

$$
\begin{aligned}
\mathcal{E}_1(\psi_1,\psi_2,\overline{\lambda},\zeta) &= \psi_1^2(\psi_2\chi_1^4 + \psi_2\chi_1^2\zeta), \\
\mathcal{E}_2(\psi_1,\psi_2,\overline{\lambda},\zeta) &= \psi_1^2(\chi_1^2\chi_2^2\overline{m}_2^2\zeta - 2\chi_1^2\chi_2^2\overline{m}_2\zeta + \chi_1^2\chi_2^2\zeta + \psi_2\chi_1^2 - \psi_2\overline{m}_1^2\overline{m}_2^2\zeta^3 + 2\psi_2\overline{m}_1^2\overline{m}_2\zeta^3 - \psi_2\overline{m}_1^2\zeta^3 + \psi_2\zeta), \\
\mathcal{E}_3(\psi_1,\psi_2,\overline{\lambda},\zeta) &= -\chi_1^4\chi_2^2\chi_3^2 + \psi_1\psi_2\chi_1^4 + \psi_1\chi_1^2\chi_2^2\overline{m}_2^2\zeta^2 - 2\psi_1\chi_1^2\chi_2^2\overline{m}_2\zeta^2 + \psi_1\chi_1^2\chi_2^2\zeta^2 \\
&\quad + \psi_2\chi_1^2\chi_3^2\overline{m}_1^2\zeta^2 + 2\psi_1\psi_2\chi_1^2\zeta - \psi_1\psi_2\overline{m}_1^2\overline{m}_2^2\zeta^4 + 2\psi_1\psi_2\overline{m}_1^2\overline{m}_2\zeta^4 - \psi_1\psi_2\overline{m}_1^2\zeta^4 + \psi_1\psi_2\zeta^2.
\end{aligned}
$$

Then

$$
\begin{aligned}
\overline{\mathcal{U}}(\overline{\lambda},\psi_1,\psi_2) &= -\frac{\left(\overline{m}_2 - 1\right)\left(\tau^2\chi_1(\psi_1,\psi_2,\overline{\lambda},\zeta) + F_1^2\right)}{\chi_1(\psi_1,\psi_2,\overline{\lambda},\zeta)}, \\
\mathcal{A}_U(\overline{\lambda},\psi_1,\psi_2) &= \frac{\tau^2\mathcal{E}_1(\psi_1,\psi_2,\overline{\lambda},\zeta) + F_1^2\mathcal{E}_1(\psi_1,\psi_2,\overline{\lambda},\zeta)}{\mathcal{E}_2(\psi_1,\psi_2,\overline{\lambda},\zeta)}.
\end{aligned}
$$

Remark 3 (Simplification of $\overline{\mathcal{T}}$ and $\mathcal{A}_T$)**.** *Define $\zeta, \overline{\lambda}$ as the rescaled version of μ_1^2 and λ*

$$
\zeta = \frac{\mu_1^2}{\mu_\star^2}, \quad \overline{\lambda} = \frac{\lambda}{\mu_\star^2}.
$$

Let $m_1(\,\cdot\,;\boldsymbol{\psi})\ m_2(\,\cdot\,;\boldsymbol{\psi}) : \mathbb{C}_+ \to \mathbb{C}_+$ be defined, for $\Im(\xi) \geq C$ a sufficiently large constant, as the unique solution of the equations

$$
\begin{aligned}
m_1 &= \psi_1 \left[-\xi + (1-\overline{\lambda}\psi_1) - m_2 + \frac{\zeta(1-m_2)}{1+\zeta m_1 - \zeta m_1 m_2} \right]^{-1}, \\
m_2 &= -\psi_2 \left[\xi + m_1 + \frac{\zeta m_1}{1+\zeta m_1 - \zeta m_1 m_2} \right]^{-1},
\end{aligned}
\tag{30}
$$

subject to the condition $|m_1| \leq \psi_1/\Im(\xi)$, $|m_2| \leq \psi_2/\Im(\xi)$. Extend this definition to $\Im(\xi) > 0$ by requiring m_1, m_2 to be analytic functions in $\mathbb{C}_+$. Let

$$
\begin{aligned}
\overline{m}_1 &= \lim_{u\to\infty} m_1(\boldsymbol{i}u, \boldsymbol{\psi}), \\
\overline{m}_2 &= \lim_{u\to\infty} m_2(\boldsymbol{i}u, \boldsymbol{\psi}).
\end{aligned}
$$

Define

$$
\chi_4 = \overline{m}_1 + \frac{\overline{m}_1\zeta}{\chi_1(\overline{m}_1,\overline{m}_2,\zeta)},
$$

and

$$\chi_1 = \overline{m}_1\zeta - \overline{m}_1\overline{m}_2\zeta + 1,$$
$$\chi_3 = \overline{\lambda}\psi_1 + \overline{m}_2 - 1 + \frac{\zeta\left(\overline{m}_2 - 1\right)}{\chi_1},$$

where the definitions of χ_1, χ_3 are the same as in Remark 2. Define three polynomials $\mathcal{E}_3, \mathcal{E}_4, \mathcal{E}_5$ as

$$\begin{aligned}
\mathcal{E}_4(\psi_1,\psi_2,\overline{\lambda},\zeta) =&\psi_1\Big(\psi_2\chi_1^4\chi_4^3 + \chi_1^4\chi_4^2\overline{m}_1^3\overline{m}_2^2\zeta^3 - 2\chi_1^4\chi_4^2\overline{m}_1^3\overline{m}_2\zeta^3 + \chi_1^4\chi_4^2\overline{m}_1^3\zeta^3 + 2\chi_1^3\chi_4^2\overline{m}_1^3\overline{m}_2^2\zeta^2\\
&- 4\chi_1^3\chi_4^2\overline{m}_1^3\overline{m}_2\zeta^2 + 2\chi_1^3\chi_4^2\overline{m}_1^3\zeta^2 - \psi_2\chi_1^3\chi_4^2\overline{m}_1\zeta + \chi_1^2\chi_4^2\overline{m}_1^3\overline{m}_2^2\zeta - 2\chi_1^2\chi_4^2\overline{m}_1^3\overline{m}_2\zeta\\
&+ \chi_1^2\chi_4^2\overline{m}_1^3\zeta + \psi_2\chi_1^2\chi_4^2\overline{m}_1\zeta - \psi_2\chi_1^2\overline{m}_1^5\overline{m}_2^2\zeta^5 + 2\psi_2\chi_1^2\overline{m}_1^5\overline{m}_2\zeta^5 - \psi_2\chi_1^2\overline{m}_1^5\zeta^5\\
&- 2\psi_2\chi_1\overline{m}_1^5\overline{m}_2^2\zeta^4 + 4\psi_2\chi_1\overline{m}_1^5\overline{m}_2\zeta^4 - 2\psi_2\chi_1\overline{m}_1^5\zeta^4 - \psi_2\overline{m}_1^5\overline{m}_2^2\zeta^3\\
&+ 2\psi_2\overline{m}_1^5\overline{m}_2\zeta^3 - \psi_2\overline{m}_1^5\zeta^3\Big),\\
\mathcal{E}_5(\psi_1,\psi_2,\overline{\lambda},\zeta) =&\overline{m}_1\Big(\zeta + 1 + \overline{m}_1\zeta - \overline{m}_1\overline{m}_2\zeta\Big)^2\Big(-\chi_1^4\chi_3^2\chi_4^2\overline{m}_1^2\\
&+ \psi_1\psi_2\chi_1^4\chi_4^2 - 2\psi_1\psi_2\chi_1^3\chi_4\overline{m}_1\zeta + \psi_2\chi_1^2\chi_3^2\overline{m}_1^4\zeta^2 + \psi_1\chi_1^2\chi_4^2\overline{m}_1^2\overline{m}_2^2\zeta^2\\
&- 2\psi_1\chi_1^2\chi_4^2\overline{m}_1^2\overline{m}_2\zeta^2 + \psi_1\chi_1^2\chi_4^2\overline{m}_1^2\zeta^2 + 2\psi_1\psi_2\chi_1^2\chi_4\overline{m}_1\zeta + \psi_1\psi_2\chi_1^2\overline{m}_1^2\zeta^2\\
&- 2\psi_1\psi_2\chi_1\overline{m}_1^2\zeta^2 - \psi_1\psi_2\overline{m}_1^4\overline{m}_2^2\zeta^4 + 2\psi_1\psi_2\overline{m}_1^4\overline{m}_2\zeta^4 - \psi_1\psi_2\overline{m}_1^4\zeta^4 + \psi_1\psi_2\overline{m}_1^2\zeta^2\Big),\\
\mathcal{E}_6(\psi_1,\psi_2,\overline{\lambda},\zeta) =&\chi_1^2\chi_4^2\psi_1\psi_2\Big(\chi_4\chi_1^2 - \overline{m}_1\chi_1\zeta + \overline{m}_1\zeta\Big)\Big(\overline{m}_1\zeta - \overline{m}_1\overline{m}_2\zeta + 1\Big)^2.
\end{aligned}$$

Then

$$\overline{\mathcal{T}}(\overline{\lambda},\psi_1,\psi_2) = -\frac{\left(\overline{m}_2 - 1\right)\left(\tau^2\chi_1(\psi_1,\psi_2,\overline{\lambda},\zeta) + F_1^2\right)}{\chi_1(\psi_1,\psi_2,\overline{\lambda},\zeta)},$$
$$\mathcal{A}_T(\overline{\lambda},\psi_1,\psi_2) = -\psi_1\frac{F_1^2\mathcal{E}_4(\psi_1,\psi_2,\overline{\lambda},\zeta) + \tau^2\mathcal{E}_6(\psi_1,\psi_2,\overline{\lambda},\zeta)}{\mathcal{E}_5(\psi_1,\psi_2,\overline{\lambda},\zeta)}.$$

A.2. Definitions of $\mathcal{R}$ and $\mathcal{A}$

In this section, we present the expression of $\mathcal{R}$ and $\mathcal{A}$ from Mei & Montanari (2019) which are used in our results and plots.

Definition 6 (Formula for the prediction error of minimum norm interpolator)**.** *Define*

$$\zeta = \mu_1^2/\mu_\star^2, \;\; \rho = F_1^2/\tau^2$$

Let the functions $\nu_1, \nu_2 : \mathbb{C}_+ \to \mathbb{C}_+$ be be uniquely defined by the following conditions: (i) ν_1, ν_2 are analytic on $\mathbb{C}_+$; (ii) For $\Im(\xi) > 0$, $\nu_1(\xi)$, $\nu_2(\xi)$ satisfy the following equations

$$\begin{aligned}
\nu_1 &= \psi_1\Big(-\xi - \nu_2 - \frac{\zeta^2\nu_2}{1 - \zeta^2\nu_1\nu_2}\Big)^{-1},\\
\nu_2 &= \psi_2\Big(-\xi - \nu_1 - \frac{\zeta^2\nu_1}{1 - \zeta^2\nu_1\nu_2}\Big)^{-1};
\end{aligned} \tag{31}$$

(iii) $(\nu_1(\xi), \nu_2(\xi))$ is the unique solution of these equations with $|\nu_1(\xi)| \le \psi_1/\Im(\xi)$, $|\nu_2(\xi)| \le \psi_2/\Im(\xi)$ for $\Im(\xi) > C$, with C a sufficiently large constant.

Let

$$\chi \equiv \lim_{u\to 0} \nu_1(\boldsymbol{i}u)\cdot\nu_2(\boldsymbol{i}u), \tag{32}$$

and

$$\begin{aligned}
E_0(\zeta,\psi_1,\psi_2) \equiv&\; -\chi^5\zeta^6 + 3\chi^4\zeta^4 + (\psi_1\psi_2 - \psi_2 - \psi_1 + 1)\chi^3\zeta^6 - 2\chi^3\zeta^4 - 3\chi^3\zeta^2\\
&+ (\psi_1 + \psi_2 - 3\psi_1\psi_2 + 1)\chi^2\zeta^4 + 2\chi^2\zeta^2 + \chi^2 + 3\psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2\,,\\
E_1(\zeta,\psi_1,\psi_2) \equiv&\; \psi_2\chi^3\zeta^4 - \psi_2\chi^2\zeta^2 + \psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2\,,\\
E_2(\zeta,\psi_1,\psi_2) \equiv&\; \chi^5\zeta^6 - 3\chi^4\zeta^4 + (\psi_1 - 1)\chi^3\zeta^6 + 2\chi^3\zeta^4 + 3\chi^3\zeta^2 + (-\psi_1 - 1)\chi^2\zeta^4 - 2\chi^2\zeta^2 - \chi^2\,.
\end{aligned} \tag{33}$$

Then the expression for the asymptotic risk of minimum norm interpolator gives

$$\mathcal{R}(\psi_1,\psi_2) = F_1^2\frac{E_1(\zeta,\psi_1,\psi_2)}{E_0(\zeta,\psi_1,\psi_2)} + \tau^2\frac{E_2(\zeta,\psi_1,\psi_2)}{E_0(\zeta,\psi_1,\psi_2)} + \tau^2.$$

The expression for the norm of the minimum norm interpolator gives

$$\begin{aligned}
A_1 =& \frac{\rho}{1+\rho}\Big[-\chi^2(\chi\zeta^4 - \chi\zeta^2 + \psi_2\zeta^2 + \zeta^2 - \chi\psi_2\zeta^4 + 1)\Big] + \frac{1}{1+\rho}\Big[\chi^2(\chi\zeta^2-1)(\chi^2\zeta^4 - 2\chi\zeta^2 + \zeta^2 + 1)\Big],\\
A_0 =& -\chi^5\zeta^6 + 3\chi^4\zeta^4 + (\psi_1\psi_2 - \psi_2 - \psi_1 + 1)\chi^3\zeta^6 - 2\chi^3\zeta^4 - 3\chi^3\zeta^2\\
&+ (\psi_1 + \psi_2 - 3\psi_1\psi_2 + 1)\chi^2\zeta^4 + 2\chi^2\zeta^2 + \chi^2 + 3\psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2,\\
\mathcal{A}(\psi_1,\psi_2) =& \psi_1(F_1^2+\tau^2)A_1/(\mu_\star^2 A_0).
\end{aligned}$$

B. Experimental setup for simulations in Figure 2

In this section, we present additional details for Figure 2. We choose $y_i = \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle$ for some $\|\boldsymbol{\beta}\|_2^2 = 1$, the ReLU activation function $\sigma(x) = \max\{x, 0\}$, and $\psi_1 = N/d = 2.5$ and $\psi_2 = n/d = 1.5$.

For the theoretical curves (in solid lines), we choose $\lambda \in [0.426, 2]$, so that $\mathcal{A}_U(\lambda) \in [0, 15]$, and plot the parametric curve $(\mathcal{A}_U(\lambda), \overline{\mathcal{U}}(\lambda) + \lambda\mathcal{A}_U(\lambda))$ for the uniform convergence. For the uniform convergence over interpolators, we choose $\lambda \in [0.21, 2]$ so that $\mathcal{A}_T(\lambda) \in [6.4, 15]$, and plot $(\mathcal{A}_T(\lambda), \overline{\mathcal{T}}(\lambda) + \lambda\mathcal{A}_T(\lambda))$. The definitions of these theoretical predictions are given in Definition 5, Remark 2 and Remark 3

For the empirical simulations (in dots), first recall that in Proposition 2, we defined

$$\begin{aligned}
\boldsymbol{a}_U(\lambda) =& \arg\max_{\boldsymbol{a}} \Big[R(\boldsymbol{a}) - \widehat{R}_n(\boldsymbol{a}) - \psi_1\lambda\|\boldsymbol{a}\|_2^2\Big],\\
\boldsymbol{a}_T(\lambda) =& \arg\max \boldsymbol{a} \inf_{\boldsymbol{\mu}} \Big[R(\boldsymbol{a}) - \lambda\psi_1\|\boldsymbol{a}\|_2^2 + 2\langle \boldsymbol{\mu}, \boldsymbol{Z}\boldsymbol{a} - \boldsymbol{y}/\sqrt{d}\,\rangle\Big].
\end{aligned}$$

After picking a value of λ, we sample 20 independent problem instances, with the number of features $N = 500$, number of samples $n = 300$, covariate dimension $d = 200$. We compute the corresponding $(\psi_1\|\boldsymbol{a}_U\|_2^2, R(\boldsymbol{a}_U) - \hat{R}_n(\boldsymbol{a}_U))$ and $(\psi_1\|\boldsymbol{a}_T\|_2^2, R(\boldsymbol{a}_T))$ for each instance. Then, we plot the empirical mean and $1/\sqrt{20}$ times the empirical standard deviation (around the mean) of each coordinate.

C. Proof of Proposition 1

The proof of Proposition 1 contains two parts: standard uniform convergence U and uniform convergence over interpolators T. The proof for the two cases are essentially the same, both based on the fact that strong duality holds for quadratic program with single quadratic constraint (c.f. Boyd & Vandenberghe (2004), Appendix A.1).

C.1. Standard uniform convergence U

Recall that the uniform convergence bound U is defined as in Eq. (4)

$$U(A, N, n, d) = \sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A} \Big(R(\boldsymbol{a}) - \widehat{R}_n(\boldsymbol{a})\Big).$$

Since the maximization problem in (4) is a quadratic program with a single quadratic constraint, the strong duality holds. So we have

$$\sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A^2} R(\boldsymbol{a}) - \widehat{R}_n(\boldsymbol{a}) = \inf_{\lambda\ge 0}\sup_{\boldsymbol{a}} \Big[R(\boldsymbol{a}) - \widehat{R}_n(\boldsymbol{a}) - \psi_1\lambda(\|\boldsymbol{a}\|_2^2 - \psi_1^{-1}A)\Big].$$

Finally, by the definition of $\overline{U}$ as in Eq. (20), we get

$$U(A, N, n, d) = \inf_{\lambda\ge 0}\Big[\overline{U}(\lambda, N, n, d) + \lambda A\Big].$$

C.2. Uniform convergence over interpolators T

Without loss of generality, we consider the regime when $N > n$.

Recall that the uniform convergence over interpolators T is defined as in Eq. (5)

$$T(A, N, n, d) = \sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A, \widehat{R}_n(\boldsymbol{a})=0} R(\boldsymbol{a}).$$

When the set $\{\boldsymbol{a} \in \mathbb{R}^N : (N/d)\|\boldsymbol{a}\|_2^2 \le A, \widehat{R}_n(\boldsymbol{a}) = 0\}$ is empty, we have

$$T(A, N, n, d) = \inf_{\lambda \ge 0} \Big[\overline{T}(\lambda, N, n, d) + \lambda A\Big] = -\infty.$$

In the following, we assume that the set $\{\boldsymbol{a} \in \mathbb{R}^N : (N/d)\|\boldsymbol{a}\|_2^2 \le A, \widehat{R}_n(\boldsymbol{a}) = 0\}$ is non-empty, i.e., there exists $\boldsymbol{a} \in \mathbb{R}^N$ such that $\widehat{R}_n(\boldsymbol{a}) = 0$ and $(N/d)\|\boldsymbol{a}\|_2^2 \le A$.

Let m be the dimension of the null space of $\boldsymbol{Z} \in \mathbb{R}^{n\times N}$, i.e. $m = \dim(\{\boldsymbol{u} : \boldsymbol{Z}\boldsymbol{u} = \boldsymbol{0}\})$. Note that $\boldsymbol{Z} \in \mathbb{R}^{N\times n}$ and $N > n$, we must have $N - n \le m \le N$. We let $\boldsymbol{R} \in \mathbb{R}^{N\times m}$ be a matrix whose column space gives the null space of matrix $\boldsymbol{Z}$. Let $\boldsymbol{a}_0$ be the minimum norm interpolating solution (whose existence is given by the assumption that $\{\boldsymbol{a} \in \mathbb{R}^N : \widehat{R}_n(\boldsymbol{a}) = 0\}$ is non-empty)

$$\boldsymbol{a}_0 = \lim_{\lambda\to 0_+} \arg\min_{\boldsymbol{a}\in\mathbb{R}^N} \Big[\widehat{R}_n(\boldsymbol{a}) + \lambda\|\boldsymbol{a}\|_2^2\Big] = \arg\min_{\boldsymbol{a}\in\mathbb{R}^N:\widehat{R}_n(\boldsymbol{a})=0} \|\boldsymbol{a}\|_2^2.$$

Then we have

$$\{\boldsymbol{a} \in \mathbb{R}^N : \widehat{R}_n(\boldsymbol{a}) = 0\} = \{\boldsymbol{a} \in \mathbb{R}^N : \boldsymbol{y} = \sqrt{d}\boldsymbol{Z}\boldsymbol{a}\} = \{\boldsymbol{R}\boldsymbol{u} + \boldsymbol{a}_0 : \boldsymbol{u} \in \mathbb{R}^m\}.$$

Then T can be rewritten as a maximization problem in terms of $\boldsymbol{u}$:

$$\begin{aligned}\sup_{(N/d)\|\boldsymbol{a}\|_2^2\le A,\widehat{R}_n(\boldsymbol{a})=0} R(\boldsymbol{a}) &= \sup_{\boldsymbol{u}\in\mathbb{R}^m:\|\boldsymbol{R}\boldsymbol{u}+\boldsymbol{a}_0\|_2^2\le\psi_1^{-1}A} \Big[\langle \boldsymbol{R}\boldsymbol{u}+\boldsymbol{a}_0, \boldsymbol{U}(\boldsymbol{R}\boldsymbol{u}+\boldsymbol{a}_0)\rangle - 2\langle \boldsymbol{R}\boldsymbol{u}+\boldsymbol{a}_0, \boldsymbol{v}\rangle + \mathbb{E}(y^2)\Big]\\ &= R(\boldsymbol{a}_0) + \sup_{\boldsymbol{u}\in\mathbb{R}^m:\|\boldsymbol{R}\boldsymbol{u}+\boldsymbol{a}_0\|_2^2\le\psi_1^{-1}A} \Big[\langle \boldsymbol{u}, \boldsymbol{R}^\mathsf{T}\boldsymbol{U}\boldsymbol{R}\boldsymbol{u}\rangle + 2\langle \boldsymbol{R}\boldsymbol{u}, \boldsymbol{U}\boldsymbol{a}_0 - \boldsymbol{v}\rangle\Big].\end{aligned}$$

Note that the optimization problem only has non-feasible region when $A > (N/d)\|\boldsymbol{a}_0\|_2^2$. By strong duality of quadratic programs with a single quadratic constraint, we have

$$\begin{aligned}&\sup_{\boldsymbol{u}\in\mathbb{R}^m:\|\boldsymbol{R}\boldsymbol{u}+\boldsymbol{a}_0\|_2^2\le\psi_1^{-1}A} \Big[\langle \boldsymbol{u}, \boldsymbol{R}^\mathsf{T}\boldsymbol{U}\boldsymbol{R}\boldsymbol{u}\rangle + 2\langle \boldsymbol{R}\boldsymbol{u}, \boldsymbol{U}\boldsymbol{a}_0 - \boldsymbol{v}\rangle\Big]\\ =& \inf_{\lambda\ge 0}\sup_{\boldsymbol{u}\in\mathbb{R}^m} \Big[\langle \boldsymbol{u}, \boldsymbol{R}^\mathsf{T}\boldsymbol{U}\boldsymbol{R}\boldsymbol{u}\rangle + 2\langle \boldsymbol{R}\boldsymbol{u}, \boldsymbol{U}\boldsymbol{a}_0 - \boldsymbol{v}\rangle - \lambda(\psi_1\|\boldsymbol{R}\boldsymbol{u}+\boldsymbol{a}_0\|_2^2 - A)\Big].\end{aligned}$$

The maximization over $\boldsymbol{u}$ can be restated as the maximization over $\boldsymbol{a}$:

$$R(\boldsymbol{a}_0) + \sup_{\boldsymbol{u}\in\mathbb{R}^m} \Big[\langle \boldsymbol{u}, \boldsymbol{R}^\mathsf{T}\boldsymbol{U}\boldsymbol{R}\boldsymbol{u}\rangle + 2\langle \boldsymbol{R}\boldsymbol{u}, \boldsymbol{U}\boldsymbol{a}_0 - \boldsymbol{v}\rangle - \lambda\psi_1\|\boldsymbol{R}\boldsymbol{u}+\boldsymbol{a}_0\|_2^2\Big] = \sup_{\boldsymbol{a}:\widehat{R}_n(\boldsymbol{a})=0} \Big[R(\boldsymbol{a}) - \lambda\psi_1\|\boldsymbol{a}\|_2^2\Big].$$

Moreover, since $\sup_{\boldsymbol{a}:\widehat{R}_n(\boldsymbol{a})=0}[R(\boldsymbol{a}) - \lambda\psi_1\|\boldsymbol{a}\|_2^2]$ is a quadratic programming with linear constraints, we have

$$\sup_{\boldsymbol{a}:\widehat{R}_n(\boldsymbol{a})=0} \Big[R(\boldsymbol{a}) - \lambda\psi_1\|\boldsymbol{a}\|_2^2\Big] = \sup_{\boldsymbol{a}}\inf_{\boldsymbol{\mu}} \Big[R(\boldsymbol{a}) - \lambda\psi_1\|\boldsymbol{a}\|_2^2 + 2\langle \boldsymbol{\mu}, \boldsymbol{Z}\boldsymbol{a} - \boldsymbol{y}/\sqrt{d}\,\rangle\Big].$$

Combining all the equality above and the definition of $\overline{T}$ as in Eq. (21), we have

$$\begin{aligned}
T(A,N,n,d) &= \sup_{(N/d)\|\boldsymbol{a}\|_2^2\le A,\widehat{R}_n(\boldsymbol{a})=0} R(\boldsymbol{a})\\
&= R(\boldsymbol{a}_0) + \sup_{\boldsymbol{u}\in\mathbb{R}^m:\|\boldsymbol{R}\boldsymbol{u}+\boldsymbol{a}_0\|_2^2\le\psi_1^{-1}A}\Big[\langle \boldsymbol{u},\boldsymbol{R}^\mathsf{T}\boldsymbol{U}\boldsymbol{R}\boldsymbol{u}\rangle + 2\langle \boldsymbol{R}\boldsymbol{u},\boldsymbol{U}\boldsymbol{a}_0-\boldsymbol{v}\rangle\Big]\\
&= R(\boldsymbol{a}_0) + \inf_{\lambda\ge 0}\sup_{\boldsymbol{u}}\Big[\langle \boldsymbol{u},\boldsymbol{R}^\mathsf{T}\boldsymbol{U}\boldsymbol{R}\boldsymbol{u}\rangle + 2\langle \boldsymbol{R}\boldsymbol{u},\boldsymbol{U}\boldsymbol{a}_0-\boldsymbol{v}\rangle - \lambda(\psi_1\|\boldsymbol{R}\boldsymbol{u}+\boldsymbol{a}_0\|_2^2-A)\Big]\\
&= \inf_{\lambda\ge 0}\Big\{\lambda A + R(\boldsymbol{a}_0) + \sup_{\boldsymbol{u}}\big[\langle \boldsymbol{u},\boldsymbol{R}^\mathsf{T}\boldsymbol{U}\boldsymbol{R}\boldsymbol{u}\rangle + 2\langle \boldsymbol{R}\boldsymbol{u},\boldsymbol{U}\boldsymbol{a}_0-\boldsymbol{v}\rangle - \lambda\psi_1\|\boldsymbol{R}\boldsymbol{u}+\boldsymbol{a}_0\|_2^2\big]\Big\}\\
&= \inf_{\lambda\ge 0}\Big\{\lambda A + \sup_{\boldsymbol{a}:\widehat{R}_n(\boldsymbol{a})=0}\Big[R(\boldsymbol{a}) - \lambda\psi_1\|\boldsymbol{a}\|_2^2\Big]\Big\}\\
&= \inf_{\lambda\ge 0}\Big\{\lambda A + \sup_{\boldsymbol{a}}\inf_{\boldsymbol{\mu}}\Big[R(\boldsymbol{a}) - \lambda\psi_1\|\boldsymbol{a}\|_2^2 + 2\langle\boldsymbol{\mu},\boldsymbol{Z}\boldsymbol{a}-\boldsymbol{y}/\sqrt{d}\,\rangle\Big]\Big\}\\
&= \inf_{\lambda\ge 0}\Big[\overline{T}(\lambda,N,n,d)+\lambda A\Big].
\end{aligned}$$

This concludes the proof.

D. Proof of Proposition 2

Note that the definitions of $\overline{U}$ and $\overline{T}$ as in Eq. (20) and (21) depend on $\boldsymbol{\beta}=\boldsymbol{\beta}^{(d)}$, where $\boldsymbol{\beta}^{(d)}$ gives the coefficients of the target function $f_d(\boldsymbol{x})=\langle\boldsymbol{x},\boldsymbol{\beta}^{(d)}\rangle$. Suppose we explicitly write their dependence on $\boldsymbol{\beta}=\boldsymbol{\beta}^{(d)}$, i.e., $\overline{U}(\lambda,N,n,d)=\overline{U}(\boldsymbol{\beta},\lambda,N,n,d)$ and $\overline{T}(\lambda,N,n,d)=\overline{T}(\boldsymbol{\beta},\lambda,N,n,d)$, then we can see that for any fixed $\boldsymbol{\beta}_\star$ and $\tilde{\boldsymbol{\beta}}$ with $\|\tilde{\boldsymbol{\beta}}\|_2=\|\boldsymbol{\beta}_\star\|_2$, we have $\overline{U}(\boldsymbol{\beta}_\star,\lambda,N,n,d)\overset{d}{=}\overline{U}(\tilde{\boldsymbol{\beta}},\lambda,N,n,d)$ and $\overline{T}(\boldsymbol{\beta}_\star,\lambda,N,n,d)\overset{d}{=}\overline{T}(\tilde{\boldsymbol{\beta}},\lambda,N,n,d)$ where the randomness comes from $\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon}$. This is by the fact that the distribution of $\boldsymbol{x}_i$'s and $\boldsymbol{\theta}_a$'s are rotationally invariant. As a consequence, for any fixed deterministic $\boldsymbol{\beta}_\star$, if we take $\boldsymbol{\beta}\sim\mathrm{Unif}(\mathbb{S}^{d-1}(\|\boldsymbol{\beta}_\star\|_2))$, we have

$$\begin{aligned}
\overline{U}(\boldsymbol{\beta}_\star,\lambda,N,n,d)&\overset{d}{=}\overline{U}(\boldsymbol{\beta},\lambda,N,n,d),\\
\overline{T}(\boldsymbol{\beta}_\star,\lambda,N,n,d)&\overset{d}{=}\overline{T}(\boldsymbol{\beta},\lambda,N,n,d).
\end{aligned}$$

where the randomness comes from $\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon},\boldsymbol{\beta}$.

Consequently, as long as we are able to show the equation

$$\overline{U}(\boldsymbol{\beta},\lambda,N,n,d)=\overline{\mathcal{U}}(\lambda,\psi_1,\psi_2)+o_{d,\mathbb{P}}(1)$$

for random $\boldsymbol{\beta}\sim\mathrm{Unif}(\mathbb{S}^{n-1}(F_1))$, this equation will also hold for any deterministic $\boldsymbol{\beta}_\star$ with $\|\boldsymbol{\beta}_\star\|_2^2=F_1^2$. Vice versa for $\overline{T}$, $\|\overline{\boldsymbol{a}}_U\|_2^2$ and $\|\overline{\boldsymbol{a}}_T\|_2^2$.

As a result, in the following, we work with the assumption that $\boldsymbol{\beta}=\boldsymbol{\beta}^{(d)}\sim\mathrm{Unif}(\mathbb{S}^{d-1}(F_1))$. That is, in proving Proposition 2, we replace Assumption 1 by Assumption 6 below. By the argument above, as long as Proposition 2 holds under Assumption 6, it also holds under the original assumption, i.e., Assumption 1.

Assumption 6 (Linear Target Function)**.** *We assume that $f_d\in L^2(\mathbb{S}^{d-1}(\sqrt{d}))$ with $f_d(\boldsymbol{x})=\langle\boldsymbol{\beta}^{(d)},\boldsymbol{x}\rangle$, where $\boldsymbol{\beta}^{(d)}\sim$* $\mathrm{Unif}(\mathbb{S}^{d-1}(F_1))$.

D.1. Expansions

Denote $\boldsymbol{v}=(v_i)_{i\in[N]}\in\mathbb{R}^N$ and $\boldsymbol{U}=(U_{ij})_{i,j\in[N]}\in\mathbb{R}^{N\times N}$ where their elements are defined via

$$\begin{aligned}
v_i&\equiv\mathbb{E}_{\varepsilon,\boldsymbol{x}}[y\sigma(\langle\boldsymbol{x},\boldsymbol{\theta}_i\rangle/\sqrt{d})],\\
U_{ij}&\equiv\mathbb{E}_{\boldsymbol{x}}[\sigma(\langle\boldsymbol{x},\boldsymbol{\theta}_i\rangle/\sqrt{d})\sigma(\langle\boldsymbol{x},\boldsymbol{\theta}_j\rangle/\sqrt{d})].
\end{aligned}$$

Here, $y=\langle\boldsymbol{x},\boldsymbol{\beta}\rangle+\varepsilon$, where $\boldsymbol{\beta}\sim\mathrm{Unif}(\mathbb{S}^{d-1}(F_1))$, $\boldsymbol{x}\sim\mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, $\varepsilon\sim\mathcal{N}(0,\tau^2)$, and $(\boldsymbol{\theta}_j)_{j\in[N]}\sim_{iid}\mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ are mutually independent. The expectations are taken with respect to the test sample $\boldsymbol{x}\sim\mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and $\varepsilon\sim\mathcal{N}(0,\tau^2)$ (especially, the expectations are conditional on $\boldsymbol{\beta}$ and $(\boldsymbol{\theta}_i)_{i\in[N]}$).

Moreover, we denote $\boldsymbol{y} = (y_1, \ldots, y_n)^\mathsf{T} \in \mathbb{R}^n$ where $y_i = \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle + \boldsymbol{\varepsilon}_i$. Recall that $(\boldsymbol{x}_i)_{i\in[n]} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and $(\varepsilon_i)_{i\in[n]} \sim_{iid} \mathcal{N}(0, \tau^2)$ are mutually independent and independent from $\boldsymbol{\beta} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$. We further denote $\boldsymbol{Z} = (Z_{ij})_{i\in[n], j\in[N]}$ where its elements are defined via

$$Z_{ij} = \sigma(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_j \rangle / \sqrt{d}) / \sqrt{d}.$$

The population risk (1) can be reformulated as

$$R(\boldsymbol{a}) = \langle \boldsymbol{a}, \boldsymbol{U}\boldsymbol{a} \rangle - 2\langle \boldsymbol{a}, \boldsymbol{v} \rangle + \mathbb{E}[y^2],$$

where $\boldsymbol{a} = (a_1, \ldots, a_N) \in \mathbb{R}^N$. The empirical risk (2) can be reformulated as

$$\widehat{R}_n(\boldsymbol{a}) = \psi_2^{-1} \langle \boldsymbol{a}, \boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\boldsymbol{a} \rangle - 2\psi_2^{-1} \frac{\langle \boldsymbol{Z}^\mathsf{T}\boldsymbol{y}, \boldsymbol{a} \rangle}{\sqrt{d}} + \frac{1}{n}\|\boldsymbol{y}\|_2^2.$$

By the Appendix A in Mei & Montanari (2019) (we include in the Appendix F for completeness), we can expand $\sigma(x)$ in terms of Gegenbauer polynommials

$$\sigma(x) = \sum_{k=0}^{\infty} \lambda_{d,k}(\sigma) B(d,k) Q_k^{(d)}(\sqrt{d} \cdot x),$$

where $Q_k^{(d)}$ is the k'th Gegenbauer polynomial in d dimensions, $B(d, k)$ is the dimension of the space of polynomials on $\mathbb{S}^{d-1}(\sqrt{d})$ with degree exactly k. Finally, $\lambda_{d,k}(\sigma)$ is the k'th Gegenbauer coefficient. More details of this expansion can be found in Appendix F.

By the properties of Gegenbauer polynomials (c.f. Appendix F.2), we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}\sim\mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))}[\boldsymbol{x} Q_k(\langle \boldsymbol{x}, \boldsymbol{\theta}_i \rangle)] = \boldsymbol{0}, \qquad &\forall k \neq 1,\\
&\mathbb{E}_{\boldsymbol{x}\sim\mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))}[\boldsymbol{x} Q_1(\langle \boldsymbol{x}, \boldsymbol{\theta}_i \rangle)] = \boldsymbol{\theta}_i / d, \qquad &k = 1.
\end{aligned}$$

As a result, we have

$$v_i = \mathbb{E}_{\varepsilon,\boldsymbol{x}}[y\sigma(\langle \boldsymbol{x}, \boldsymbol{\theta}_i \rangle / \sqrt{d})] = \sum_{k=0}^{\infty} \lambda_{d,k}(\sigma) B(d,k) \mathbb{E}_{\boldsymbol{x}}[\langle \boldsymbol{x}, \boldsymbol{\beta} \rangle Q_k^{(d)}(\sqrt{d} \cdot x)] = \lambda_{d,1}(\sigma) \langle \boldsymbol{\theta}_i, \boldsymbol{\beta} \rangle. \tag{34}$$

D.2. Removing the perturbations

By Lemma 6 and 7 as in Appendix D.6, we have the following decomposition

$$\boldsymbol{U} = \mu_1^2 \boldsymbol{Q} + \mu_\star^2 \mathbf{I}_N + \boldsymbol{\Delta}, \tag{35}$$

with $\boldsymbol{Q} = \boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T}/d$, $\mathbb{E}[\|\boldsymbol{\Delta}\|_{\mathrm{op}}^2] = o_d(1)$, and μ_1^2 and $\mu_\star^2$ are given in Assumption 2.

In the following, we would like to show that $\boldsymbol{\Delta}$ has vanishing effects in the asymptotics of $\overline{U}$, $\overline{T}$, $\|\overline{\boldsymbol{a}}_U\|_2^2$ and $\|\overline{\boldsymbol{a}}_T\|_2^2$.

For this purpose, we denote

$$\begin{aligned}
\boldsymbol{U}_c &= \mu_1^2 \boldsymbol{Q} + \mu_\star^2 \mathbf{I}_N,\\
R_c(\boldsymbol{a}) &= \langle \boldsymbol{a}, \boldsymbol{U}_c \boldsymbol{a} \rangle - 2\langle \boldsymbol{a}, \boldsymbol{v} \rangle + \mathbb{E}[y^2],\\
\widehat{R}_{c,n}(\boldsymbol{a}) &= \langle \boldsymbol{a}, \psi_2^{-1} \boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\boldsymbol{a} \rangle - 2\langle \boldsymbol{a}, \psi_2^{-1} \boldsymbol{Z}^\mathsf{T}\boldsymbol{y}/\sqrt{d} \rangle + \mathbb{E}[y^2],\\
\overline{U}_c(\lambda, N, n, d) &= \sup_{\boldsymbol{a}} \Big(R_c(\boldsymbol{a}) - \widehat{R}_{c,n}(\boldsymbol{a}) - \psi_1 \lambda \|\boldsymbol{a}\|_2^2 \Big),\\
\overline{T}_c(\lambda, N, n, d) &= \sup_{\boldsymbol{a}} \inf_{\boldsymbol{\mu}} \Big[R_c(\boldsymbol{a}) - \lambda\psi_1 \|\boldsymbol{a}\|_2^2 + 2\langle \boldsymbol{\mu}, \boldsymbol{Z}\boldsymbol{a} - \boldsymbol{y}/\sqrt{d}\, \rangle \Big].
\end{aligned} \tag{36}$$

For a fixed $\lambda \in \Lambda_U$, note we have

$$\begin{aligned}\overline{U}_c(\lambda, N, n, d) &= \sup_{\boldsymbol{a}} \Big(\langle \boldsymbol{a}, (\boldsymbol{U}_c - \psi_2^{-1} \boldsymbol{Z}^\mathsf{T} \boldsymbol{Z} - \psi_1 \lambda \mathbf{I}_N) \boldsymbol{a} \rangle - 2 \langle \boldsymbol{a}, \boldsymbol{v} - \psi_2^{-1} \frac{\boldsymbol{Z}^\mathsf{T} \boldsymbol{y}}{\sqrt{d}} \rangle \Big) \\ &= \sup_{\boldsymbol{a}} \Big(\langle \boldsymbol{a}, \overline{\boldsymbol{M}} \boldsymbol{a} \rangle - 2 \langle \boldsymbol{a}, \overline{\boldsymbol{v}} \rangle \Big)\end{aligned} \tag{37}$$

where $\overline{\boldsymbol{M}} = \boldsymbol{U}_c - \psi_2^{-1} \boldsymbol{Z}^\mathsf{T} \boldsymbol{Z} - \psi_1 \lambda \mathbf{I}_N$ and $\overline{\boldsymbol{v}} = \boldsymbol{v} - \psi_2^{-1} \boldsymbol{Z}^\mathsf{T} \boldsymbol{y} / \sqrt{d}$. When $\boldsymbol{X}, \boldsymbol{\Theta}$ are such that the good event in Assumption 4 happens (which says that $\overline{\boldsymbol{M}} \preceq -\varepsilon \mathbf{I}_N$ for some $\varepsilon > 0$), the inner maximization can be uniquely achieved at

$$\overline{\boldsymbol{a}}_{U,c}(\lambda) = \arg\max_{\boldsymbol{a}} \Big(\langle \boldsymbol{a}, \overline{\boldsymbol{M}} \boldsymbol{a} \rangle - 2 \langle \boldsymbol{a}, \overline{\boldsymbol{v}} \rangle \Big) = \overline{\boldsymbol{M}}^{-1} \overline{\boldsymbol{v}}. \tag{38}$$

and when the good event $\{\|\boldsymbol{\Delta}\|_{\rm op} \le \varepsilon/2\}$ also happens, the maximizer in the definition of $\overline{U}(\lambda, N, n, d)$ (c.f. Eq. (20)) can be uniquely achieved at

$$\overline{\boldsymbol{a}}_U(\lambda) = \arg\max_{\boldsymbol{a}} \Big(\langle \boldsymbol{a}, (\overline{\boldsymbol{M}} + \boldsymbol{\Delta}) \boldsymbol{a} \rangle - 2 \langle \boldsymbol{a}, \overline{\boldsymbol{v}} \rangle \Big) = (\overline{\boldsymbol{M}} + \boldsymbol{\Delta})^{-1} \overline{\boldsymbol{v}}.$$

Note we have

$$\overline{\boldsymbol{a}}_U(\lambda) - \overline{\boldsymbol{a}}_{U,c}(\lambda) = (\overline{\boldsymbol{M}} + \boldsymbol{\Delta})^{-1} \overline{\boldsymbol{v}} - \overline{\boldsymbol{M}}^{-1} \overline{\boldsymbol{v}} = (\overline{\boldsymbol{M}} + \boldsymbol{\Delta})^{-1} \boldsymbol{\Delta} \overline{\boldsymbol{M}}^{-1} \overline{\boldsymbol{v}},$$

so by the fact that $\|\boldsymbol{\Delta}\|_{\rm op} = o_{d,\mathbb{P}}(1)$, we have

$$\|\overline{\boldsymbol{a}}_U(\lambda) - \overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2 \le \|(\overline{\boldsymbol{M}} + \boldsymbol{\Delta})^{-1} \boldsymbol{\Delta}\|_{\rm op} \|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2 = o_{d,\mathbb{P}}(1) \|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2.$$

This gives $\|\overline{\boldsymbol{a}}_U(\lambda)\|_2^2 = (1 + o_{d,\mathbb{P}}(1)) \|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2^2$.

Moreover, by the fact that $\|\boldsymbol{\Delta}\|_{\rm op} = o_{d,\mathbb{P}}(1)$, we have

$$\begin{aligned}\overline{U}_c(\lambda, N, n, d) &= \sup_{\boldsymbol{a}} \Big(R(\boldsymbol{a}) - \widehat{R}_n(\boldsymbol{a}) - \psi_1 \lambda \|\boldsymbol{a}\|_2^2 - \langle \boldsymbol{a}, \boldsymbol{\Delta} \boldsymbol{a} \rangle \Big) + \mathbb{E}[y^2] - \|\boldsymbol{y}\|_2^2 / n \\ &= \overline{U}(\lambda, N, n, d) + o_{d,\mathbb{P}}(1) (\|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2^2 + 1).\end{aligned}$$

As a consequence, as long as we can prove the asymptotics of $\overline{U}_c$ and $\|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2^2$, it also gives the asymptotics of $\overline{U}$ and $\|\overline{\boldsymbol{a}}_U(\lambda)\|_2^2$. Vice versa for $\overline{T}$ and $\|\overline{\boldsymbol{a}}_T(\lambda)\|_2^2$.

D.3. The asymptotics of $\overline{U}_c$ and $\psi_1 \|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2^2$

In the following, we derive the asymptotics of $\overline{U}_c(\lambda, N, n, d)$ and $\psi_1 \|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2^2$. When we refer to $\overline{\boldsymbol{a}}_{U,c}(\lambda)$, it is always well defined with high probability, since it can be well defined under the condition that the good event in Assumption 4 happens. Note that this good event only depend on $\boldsymbol{X}, \boldsymbol{\Theta}$ and is independent of $\boldsymbol{\beta}, \boldsymbol{\varepsilon}$.

By Eq. (37) and (38), simple calculation shows that

$$\begin{aligned}\overline{U}_c(\lambda, N, n, d) &\equiv -\langle \overline{\boldsymbol{v}}, \overline{\boldsymbol{M}}^{-1} \overline{\boldsymbol{v}} \rangle = -\Psi_1 - \Psi_2 - \Psi_3, \\ \|\overline{\boldsymbol{a}}_{U,c}\|_2^2 &\equiv \langle \overline{\boldsymbol{v}}, \overline{\boldsymbol{M}}^{-2} \overline{\boldsymbol{v}} \rangle = \Phi_1 + \Phi_2 + \Phi_3,\end{aligned}$$

where

$$\begin{aligned}\Psi_1 &= \langle \boldsymbol{v}, \overline{\boldsymbol{M}}^{-1} \boldsymbol{v} \rangle, & \Phi_1 &= \langle \boldsymbol{v}, \overline{\boldsymbol{M}}^{-2} \boldsymbol{v} \rangle, \\ \Psi_2 &= -2\psi_2^{-1} \langle \frac{\boldsymbol{Z}^\mathsf{T} \boldsymbol{y}}{\sqrt{d}}, \overline{\boldsymbol{M}}^{-1} \boldsymbol{v} \rangle, & \Phi_2 &= -2\psi_2^{-1} \langle \frac{\boldsymbol{Z}^\mathsf{T} \boldsymbol{y}}{\sqrt{d}}, \overline{\boldsymbol{M}}^{-2} \boldsymbol{v} \rangle, \\ \Psi_3 &= \psi_2^{-2} \langle \frac{\boldsymbol{Z}^\mathsf{T} \boldsymbol{y}}{\sqrt{d}}, \overline{\boldsymbol{M}}^{-1} \frac{\boldsymbol{Z}^\mathsf{T} \boldsymbol{y}}{\sqrt{d}} \rangle, & \Phi_3 &= \psi_2^{-2} \langle \frac{\boldsymbol{Z}^\mathsf{T} \boldsymbol{y}}{\sqrt{d}}, \overline{\boldsymbol{M}}^{-2} \frac{\boldsymbol{Z}^\mathsf{T} \boldsymbol{y}}{\sqrt{d}} \rangle.\end{aligned}$$

The following lemma gives the expectation of Ψ_i's and Φ_i's with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$.

Lemma 1 (Expectation of Ψ_i's and Φ_i's)**.** *Denote* $\boldsymbol{q}_U(\lambda, \boldsymbol{\psi}) = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, \psi_2, 0, 0)$. *We have*

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_1] &= \mu_1^2 F_1^2 \cdot \frac{1}{d}\mathrm{Tr}\Big(\overline{\boldsymbol{M}}^{-1}\boldsymbol{Q}\Big) \times (1 + o_d(1)),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_2] &= -\frac{2F_1^2}{\psi_2} \cdot \frac{1}{d}\mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{Z}_1^\mathsf{T}\Big) \times (1 + o_d(1)),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_3] &= \frac{F_1^2}{\psi_2^2} \cdot \frac{1}{d}\mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{H}\Big) + \frac{\tau^2}{\psi_2^2} \cdot \frac{1}{d}\mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{Z}^\mathsf{T}\Big),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_1] &= \mu_1^2 F_1^2 \cdot \frac{1}{d}\mathrm{Tr}\Big(\overline{\boldsymbol{M}}^{-2}\boldsymbol{Q}\Big) \times (1 + o_d(1)),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_2] &= -\frac{2F_1^2}{\psi_2} \cdot \frac{1}{d}\mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-2}\boldsymbol{Z}_1^\mathsf{T}\Big) \times (1 + o_d(1)),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_3] &= \frac{F_1^2}{\psi_2^2} \cdot \frac{1}{d}\mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-2}\boldsymbol{Z}^\mathsf{T}\boldsymbol{H}\Big) + \frac{\tau^2}{\psi_2^2} \cdot \frac{1}{d}\mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-2}\boldsymbol{Z}^\mathsf{T}\Big).
\end{aligned}
$$

Here the definitions of $\boldsymbol{Q}$, $\boldsymbol{H}$, *and* $\boldsymbol{Z}_1$ *are given by Eq. (19).*

Furthermore, we have

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_1] &= \mu_1^2 F_1^2 \cdot \partial_{s_2} G_d(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi})) \times (1 + o_d(1)),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_2] &= F_1^2 \cdot \partial_p G_d(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi})) \times (1 + o_d(1)),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_3] &= F_1^2 \cdot (\partial_{t_2} G_d(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi})) - 1) + \tau^2 \cdot (\partial_{t_1} G_d(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi})) - 1),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_1] &= -\mu_1^2 F_1^2 \cdot \partial_{s_1}\partial_{s_2} G_d(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi})) \times (1 + o_d(1)),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_2] &= -F_1^2 \cdot \partial_{s_1}\partial_p G_d(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi})) \times (1 + o_d(1)),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_3] &= -F_1^2 \cdot \partial_{s_1}\partial_{t_2} G_d(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi})) - \tau^2 \cdot \partial_{s_1}\partial_{t_1} G_d(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi})).
\end{aligned}
$$

The definition of G_d *is as in Definition 1, and* $\nabla_{\boldsymbol{q}}^k G_d(0_+; \boldsymbol{q})$ *for* $k \in \{1, 2\}$ *stands for the* k*'th derivatives (as a vector or a matrix) of* $G_d(\boldsymbol{i}u; \boldsymbol{q})$ *with respect to* $\boldsymbol{q}$ *in the* $u \to 0+$ *limit (with its elements given by partial derivatives)*

$$
\nabla_{\boldsymbol{q}}^k G_d(0_+; \boldsymbol{q}) = \lim_{u \to 0+} \nabla_{\boldsymbol{q}}^k G_d(\boldsymbol{i}u; \boldsymbol{q}).
$$

We next state the asymptotic characterization of the log-determinant which was proven in (Mei & Montanari, 2019).

Proposition 3 (Proposition 8.4 in (Mei & Montanari, 2019))**.** *Define*

$$
\begin{aligned}
\Xi(\xi, z_1, z_2; \boldsymbol{q}; \boldsymbol{\psi}) \equiv\ & \log[(s_2 z_1 + 1)(t_2 z_2 + 1) - \mu_1^2 (1 + p)^2 z_1 z_2] - \mu_\star^2 z_1 z_2\\
&+ s_1 z_1 + t_1 z_2 - \psi_1 \log(z_1/\psi_1) - \psi_2 \log(z_2/\psi_2) - \xi(z_1 + z_2) - \psi_1 - \psi_2.
\end{aligned}
\tag{39}
$$

For $\xi \in \mathbb{C}_+$ *and* $\boldsymbol{q} \in \mathcal{Q}$ *(c.f. Eq. (25)), let* $m_1(\xi; \boldsymbol{q}; \boldsymbol{\psi}), m_2(\xi; \boldsymbol{q}; \boldsymbol{\psi})$ *be defined as the analytic continuation of solution of Eq. (26) as defined in Definition 3. Define*

$$
g(\xi; \boldsymbol{q}; \boldsymbol{\psi}) = \Xi(\xi, m_1(\xi; \boldsymbol{q}; \boldsymbol{\psi}), m_2(\xi; \boldsymbol{q}; \boldsymbol{\psi}); \boldsymbol{q}; \boldsymbol{\psi}). \tag{40}
$$

Consider proportional asymptotics $N/d \to \psi_1$, $N/d \to \psi_2$, *as per Assumption 3. Then for any fixed* $\xi \in \mathbb{C}_+$ *and* $\boldsymbol{q} \in \mathcal{Q}$, *we have*

$$
\lim_{d \to \infty} \mathbb{E}[|G_d(\xi; \boldsymbol{q}) - g(\xi; \boldsymbol{q}; \boldsymbol{\psi})|] = 0. \tag{41}
$$

Moreover, for any fixed $u \in \mathbb{R}_+$ *and* $\boldsymbol{q} \in \mathcal{Q}$, *we have*

$$
\lim_{d \to \infty} \mathbb{E}[\|\partial_{\boldsymbol{q}} G_d(\boldsymbol{i}u; \boldsymbol{q}) - \partial_{\boldsymbol{q}} g(\boldsymbol{i}u; \boldsymbol{q}; \boldsymbol{\psi})\|_2] = 0, \tag{42}
$$

$$
\lim_{d \to \infty} \mathbb{E}[\|\nabla_{\boldsymbol{q}}^2 G_d(\boldsymbol{i}u; \boldsymbol{q}) - \nabla_{\boldsymbol{q}}^2 g(\boldsymbol{i}u; \boldsymbol{q}; \boldsymbol{\psi})\|_{\mathrm{op}}] = 0. \tag{43}
$$

Remark 4. *Note that Proposition 8.4 in (Mei & Montanari, 2019) stated that the Eq. (42) and (43) holds at* $\boldsymbol{q} = \boldsymbol{0}$. *However, by a simple modification of their proof, one can show that these equations also holds at any* $\boldsymbol{q} \in \mathcal{Q}$.

Combining Assumption 5 with Proposition 3, we have

Proposition 4. *Let Assumption 5 holds. For any $\lambda \in \Lambda_U$, denote $\boldsymbol{q}_U = \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}) = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, \psi_2, 0, 0)$, then we have, for $k = 1, 2$,*

$$\|\nabla_{\boldsymbol{q}}^k G_d(0_+; \boldsymbol{q}_U) - \lim_{u \to 0_+} \nabla_{\boldsymbol{q}}^k g(\boldsymbol{i}u; \boldsymbol{q}_U; \boldsymbol{\psi})\| = o_{d,\mathbb{P}}(1).$$

As a consequence of Proposition 4, we can calculate the asymptotics of Ψ_i's and Φ_i's. Combined with the concentration result in Lemma 2 latter in the section, the proposition below completes the proof of the part of Proposition 2 regarding the standard uniform convergence U. Its correctness follows directly from Lemma 1 and Proposition 4.

Proposition 5. *Follow the assumptions of Proposition 2. For any $\lambda \in \Lambda_U$, denote $\boldsymbol{q}_U(\lambda, \boldsymbol{\psi}) = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, \psi_2, 0, 0)$, then we have*

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_1] &\xrightarrow{\mathbb{P}} \mu_1^2 F_1^2 \cdot \partial_{s_2} g(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_2] &\xrightarrow{\mathbb{P}} F_1^2 \cdot \partial_p g(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_3] &\xrightarrow{\mathbb{P}} F_1^2 \cdot \Big(\partial_{t_2} g(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}) - 1\Big) + \tau^2 \Big(\partial_{t_1} g(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}) - 1\Big),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_1] &\xrightarrow{\mathbb{P}} -\mu_1^2 F_1^2 \cdot \partial_{s1}\partial_{s2} g(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_2] &\xrightarrow{\mathbb{P}} -F_1^2 \cdot \partial_{s1}\partial_p g(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_3] &\xrightarrow{\mathbb{P}} -F_1^2 \cdot \partial_{s_1}\partial_{t_2} g(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}) - \tau^2 \cdot \partial_{s_1}\partial_{t_1} g(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}),
\end{aligned}$$

where $\nabla_{\boldsymbol{q}}^k g(0_+; \boldsymbol{q}; \boldsymbol{\psi})$ for $k \in \{1, 2\}$ stands for the k'th derivatives (as a vector or a matrix) of $g(\boldsymbol{i}u; \boldsymbol{q}; \boldsymbol{\psi})$ with respect to $\boldsymbol{q}$ in the $u \to 0+$ limit (with its elements given by partial derivatives)

$$\nabla_{\boldsymbol{q}}^k g(0_+; \boldsymbol{q}; \boldsymbol{\psi}) = \lim_{u \to 0_+} \nabla_{\boldsymbol{q}}^k g(\boldsymbol{i}u; \boldsymbol{q}; \boldsymbol{\psi}).$$

As a consequence, we have

$$\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\overline{U}_c(\lambda, N, n, d)] \xrightarrow{\mathbb{P}} \overline{\mathcal{U}}(\lambda, \psi_1, \psi_2), \qquad \mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\psi_1 \|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2^2] \xrightarrow{\mathbb{P}} \mathcal{A}_U(\lambda, \psi_1, \psi_2),$$

where the definitions of $\overline{\mathcal{U}}$ and $\mathcal{A}_U$ are given in Definition 5. Here $\xrightarrow{\mathbb{P}}$ stands for convergence in probability as $N/d \to \psi_1$ and $n/d \to \psi_2$ (with respect to the randomness of $\boldsymbol{X}$ and $\boldsymbol{\Theta}$).

Lemma 2. *Follow the assumptions of Proposition 2. For any $\lambda \in \Lambda_U$, we have*

$$\begin{aligned}
\mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_1], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_2], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_3] &= o_{d,\mathbb{P}}(1),\\
\mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_1], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_2], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_3] &= o_{d,\mathbb{P}}(1),
\end{aligned}$$

so that

$$\mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\overline{U}_c(\lambda, N, n, d)], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2^2] = o_{d,\mathbb{P}}(1).$$

Here, $o_{d,\mathbb{P}}(1)$ stands for converges to 0 in probability (with respect to the randomness of $\boldsymbol{X}$ and $\boldsymbol{\Theta}$) as $N/d \to \psi_1$ and $n/d \to \psi_2$ and $d \to \infty$.

Now, combining Lemma 2 and Proposition 5, we have

$$\overline{U}_c(\lambda, N, n, d) \xrightarrow{\mathbb{P}} \overline{\mathcal{U}}(\lambda, \psi_1, \psi_2), \qquad \psi_1 \|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2^2 \xrightarrow{\mathbb{P}} \mathcal{A}_U(\lambda, \psi_1, \psi_2),$$

Finally, combining with the arguments in Appendix D.2 proves the asymptotics of $\overline{U}$ and $\psi_1 \|\overline{\boldsymbol{a}}_U(\lambda)\|_2^2$.

D.4. The asymptotics of $\overline{T}_c$ and $\psi_1 \|\overline{\boldsymbol{a}}_{T,c}(\lambda)\|_2^2$

In the following, we derive the asymptotics of $\overline{T}_c(\lambda, N, n, d)$ and $\psi_1 \|\overline{\boldsymbol{a}}_{T,c}(\lambda)\|_2^2$. This follows the same steps as the proof of the asymptotics of $\overline{U}_c$ and $\psi_1 \|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2^2$. We will give an overview of its proof. The detailed proof is the same as that of $\overline{U}_c$, and we will not include them for brevity.

For a fixed $\lambda \in \Lambda_T$, recalling that the definition of $\overline{T}_c$ as in Eq. (36), we have

$$\begin{aligned}\overline{T}_c(\lambda, N, n, d) &= \sup_{\boldsymbol{a}} \inf_{\boldsymbol{\mu}} \left[R_c(\boldsymbol{a}) - \lambda\psi_1\|\boldsymbol{a}\|_2^2 + 2\langle \boldsymbol{\mu}, \boldsymbol{Z}\boldsymbol{a} - \boldsymbol{y}/\sqrt{d}\,\rangle \right] \\ &= \sup_{\boldsymbol{a}} \inf_{\boldsymbol{\mu}} \left(\langle \boldsymbol{a}, (\boldsymbol{U}_c - \lambda\psi_1 \boldsymbol{I}_N)\boldsymbol{a}\rangle - 2\langle \boldsymbol{a}, \boldsymbol{v}\rangle + 2\langle \boldsymbol{\mu}, \boldsymbol{Z}\boldsymbol{a}\rangle - 2\langle \boldsymbol{\mu}, \boldsymbol{y}/\sqrt{d}\rangle \right) + \mathbb{E}[y^2] \\ &= \sup_{\sqrt{d}\boldsymbol{Z}\boldsymbol{a}=y} \langle \boldsymbol{a}, (\boldsymbol{U}_c - \lambda\psi_1 \boldsymbol{I}_N)\boldsymbol{a}\rangle - 2\langle \boldsymbol{a}, \boldsymbol{v}\rangle + \mathbb{E}[y^2]\end{aligned} \tag{44}$$

Whenever the good event in Assumption 4 happens, $(\boldsymbol{U}_c - \lambda\psi_1 \boldsymbol{I}_N)$ is negative definite in null$(\boldsymbol{Z})$. The optimum of the above variational equation exists. By KKT condition, the optimal $\boldsymbol{a}$ and dual variable $\boldsymbol{\mu}$ satisfies

- Stationary condition: $(\boldsymbol{U}_c - \lambda\psi_1 \boldsymbol{I}_N)\boldsymbol{a} + \boldsymbol{Z}^{\mathsf{T}}\mu = \boldsymbol{v}$.
- Primal Feasible: $\boldsymbol{Z}\boldsymbol{a} = \boldsymbol{y}/\sqrt{d}$.

The two conditions can be written compactly as

$$\begin{bmatrix} \boldsymbol{U}_c - \psi_1\lambda\mathbf{I}_N & \boldsymbol{Z}^{\mathsf{T}} \\ \boldsymbol{Z} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{y}/\sqrt{d} \end{bmatrix}. \tag{45}$$

We define

$$\overline{\boldsymbol{M}} \equiv \begin{bmatrix} \boldsymbol{U}_c - \psi_1\lambda\mathbf{I}_N & \boldsymbol{Z}^{\mathsf{T}} \\ \boldsymbol{Z} & \mathbf{0} \end{bmatrix}, \qquad \overline{\boldsymbol{v}} \equiv \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{y}/\sqrt{d} \end{bmatrix}. \tag{46}$$

Under Assumption 4, $\overline{\boldsymbol{M}}$ is invertible. To see this, suppose there exists vector $[\boldsymbol{a}_1^{\mathsf{T}}, \boldsymbol{\mu}_1^{\mathsf{T}}]^{\mathsf{T}} \neq \mathbf{0} \in \mathbb{R}^{N+n}$ such that $\overline{\boldsymbol{M}}[\boldsymbol{a}_1^{\mathsf{T}}, \boldsymbol{\mu}_1^{\mathsf{T}}]^{\mathsf{T}} = \mathbf{0}$, then

$$\begin{aligned}(\boldsymbol{U}_c - \lambda\psi_1 \boldsymbol{I}_N)\boldsymbol{a}_1 + \boldsymbol{Z}^{\mathsf{T}}\boldsymbol{\mu}_1 &= 0, \\ \boldsymbol{Z}\boldsymbol{a}_1 &= 0.\end{aligned}$$

As in Assumption 4, let $\mathsf{P}_{\text{null}} = \mathbf{I}_N - \boldsymbol{Z}^{\dagger}\boldsymbol{Z}$. We write $\boldsymbol{a}_1 = \mathsf{P}_{\text{null}}\boldsymbol{v}_1$ for some $\boldsymbol{v}_1 \neq \mathbf{0} \in \mathbb{R}^N$. Then,

$$\begin{aligned}&(\boldsymbol{U}_c - \lambda\psi_1 \boldsymbol{I}_N)\mathsf{P}_{\text{null}}\boldsymbol{v}_1 + \boldsymbol{Z}^{\mathsf{T}}\boldsymbol{\mu}_1 = 0, \\ \Rightarrow\ & \mathsf{P}_{\text{null}}(\boldsymbol{U}_c - \lambda\psi_1 \boldsymbol{I}_N)\mathsf{P}_{\text{null}}\boldsymbol{v}_1 + \mathsf{P}_{\text{null}}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{\mu}_1 = 0, \\ \Rightarrow\ & \mathsf{P}_{\text{null}}(\boldsymbol{U}_c - \lambda\psi_1 \boldsymbol{I}_N)\mathsf{P}_{\text{null}}\boldsymbol{v}_1 = 0,\end{aligned}$$

where the last relation come from the fact that $\boldsymbol{Z}\mathsf{P}_{\text{null}} = \mathbf{0}$. However by Assumption 4, $\mathsf{P}_{\text{null}}(\boldsymbol{U}_c - \lambda\psi_1 \boldsymbol{I}_N)\mathsf{P}_{\text{null}}$ is negative definite, which leads to a contradiction.

In the following, we assume the event in Assumption 4 happens so that $\overline{\boldsymbol{M}}$ is invertible. In this case, the maximizer in Eq. (44) can be well defined as

$$\overline{\boldsymbol{a}}_{T,c}(\lambda) = [\mathbf{I}_N, \mathbf{0}_{N\times n}]\overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{v}}.$$

Moreover, we can write $\overline{T}_c$ as

$$\overline{T}_c(\lambda, N, n, d) = \mathbb{E}[y^2] - \overline{\boldsymbol{v}}^{\mathsf{T}}\overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{v}}.$$

We further define

$$\overline{\boldsymbol{v}}_1 = [\boldsymbol{v}^{\mathsf{T}}, \mathbf{0}_{n\times 1}^{\mathsf{T}}]^{\mathsf{T}}, \quad \overline{\boldsymbol{v}}_2 = [\mathbf{0}_{N\times 1}^{\mathsf{T}}, \boldsymbol{y}^{\mathsf{T}}/\sqrt{d}]^{\mathsf{T}}, \quad \boldsymbol{E} \equiv \begin{bmatrix} \mathbf{I}_N & \mathbf{0}_{N\times n} \\ \mathbf{0}_{n\times N} & \mathbf{0}_{n\times n} \end{bmatrix}.$$

Simple calculation shows that

$$\begin{aligned}\overline{T}_c(\lambda, N, n, d) &\equiv \mathbb{E}[y^2] - \langle \overline{\boldsymbol{v}}, \overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{v}}\rangle = F_1^2 + \tau^2 - \Psi_1 - \Psi_2 - \Psi_3, \\ \|\overline{\boldsymbol{a}}_{U,c}\|_2^2 &\equiv \langle \overline{\boldsymbol{v}}, \overline{\boldsymbol{M}}^{-1}\boldsymbol{E}\overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{v}}\rangle = \Phi_1 + \Phi_2 + \Phi_3,\end{aligned}$$

where

$$\begin{aligned}\Psi_1 &= \langle \overline{\boldsymbol{v}}_1, \overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{v}}_1\rangle, & \Phi_1 &= \langle \overline{\boldsymbol{v}}_1, \overline{\boldsymbol{M}}^{-1}\boldsymbol{E}\overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{v}}_1\rangle, \\ \Psi_2 &= 2\langle \overline{\boldsymbol{v}}_2, \overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{v}}_1\rangle, & \Phi_2 &= 2\langle \overline{\boldsymbol{v}}_2, \overline{\boldsymbol{M}}^{-1}\boldsymbol{E}\overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{v}}_1\rangle, \\ \Psi_3 &= \langle \overline{\boldsymbol{v}}_2, \overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{v}}_2\rangle, & \Phi_3 &= \langle \overline{\boldsymbol{v}}_2, \overline{\boldsymbol{M}}^{-1}\boldsymbol{E}\overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{v}}_2\rangle.\end{aligned}$$

The following lemma gives the expectation of Ψ_i's and Φ_i's with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$.

Lemma 3 (Expectation of Ψ_i's and Φ_i's)**.** *Denote* $\boldsymbol{q}_T(\lambda,\boldsymbol{\psi}) = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, 0, 0, 0)$. *We have*

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_1] &= \mu_1^2 F_1^2 \cdot \partial_{s_2} G_d(0_+; \boldsymbol{q}_T(\lambda,\boldsymbol{\psi})) \times (1 + o_d(1)),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_2] &= F_1^2 \cdot \partial_p G_d(0_+; \boldsymbol{q}_T(\lambda,\boldsymbol{\psi})) \times (1 + o_d(1)),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_3] &= F_1^2 \cdot \partial_{t_2} G_d(0_+; \boldsymbol{q}_T(\lambda,\boldsymbol{\psi})) + \tau^2 \cdot \partial_{t_1} G_d(0_+; \boldsymbol{q}_T(\lambda,\boldsymbol{\psi})),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_1] &= -\mu_1^2 F_1^2 \cdot \partial_{s_1}\partial_{s_2} G_d(0_+; \boldsymbol{q}_T(\lambda,\boldsymbol{\psi})) \times (1 + o_d(1)),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_2] &= -F_1^2 \cdot \partial_{s_1}\partial_p G_d(0_+; \boldsymbol{q}_T(\lambda,\boldsymbol{\psi})) \times (1 + o_d(1)),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_3] &= -F_1^2 \cdot \partial_{s_1}\partial_{t_2} G_d(0_+; \boldsymbol{q}_T(\lambda,\boldsymbol{\psi})) - \tau^2 \cdot \partial_{s_1}\partial_{t_1} G_d(0_+; \boldsymbol{q}_T(\lambda,\boldsymbol{\psi})).
\end{aligned}
$$

The definition of G_d is as in Definition 1, and $\nabla_{\boldsymbol{q}}^k G_d(0_+;\boldsymbol{q})$ for $k \in \{1,2\}$ stands for the k'th derivatives (as a vector or a matrix) of $G_d(\boldsymbol{i}u;\boldsymbol{q})$ with respect to $\boldsymbol{q}$ in the $u \to 0+$ limit (with its elements given by partial derivatives)

$$\nabla_{\boldsymbol{q}}^k G_d(0_+;\boldsymbol{q}) = \lim_{u\to 0+} \nabla_{\boldsymbol{q}}^k G_d(\boldsymbol{i}u;\boldsymbol{q}).$$

The proof of Lemma 3 follows from direct calculation and is identical to the proof of Lemma 1. Combining Assumption 5 with Proposition 3, we have

Proposition 6. *Let Assumption 5 holds. For any $\lambda \in \Lambda_T$, denote $\boldsymbol{q}_T = \boldsymbol{q}_T(\lambda,\boldsymbol{\psi}) = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, 0, 0, 0)$, then we have, for $k = 1, 2$,*

$$\|\nabla_{\boldsymbol{q}}^k G_d(0_+;\boldsymbol{q}_T) - \lim_{u\to 0+} \nabla_{\boldsymbol{q}}^k g(\boldsymbol{i}u;\boldsymbol{q}_T;\boldsymbol{\psi})\| = o_{d,\mathbb{P}}(1).$$

As a consequence of Proposition 6, we can calculate the asymptotics of Ψ_i's and Φ_i's.

Proposition 7. *Follow the assumptions of Proposition 2. For any $\lambda \in \Lambda_T$, denote $\boldsymbol{q}_T(\lambda,\boldsymbol{\psi}) = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, 0, 0, 0)$, then we have*

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_1] &\xrightarrow{\mathbb{P}} \mu_1^2 F_1^2 \cdot \partial_{s_2} g(0_+; \boldsymbol{q}_T(\lambda,\boldsymbol{\psi});\boldsymbol{\psi}),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_2] &\xrightarrow{\mathbb{P}} F_1^2 \cdot \partial_p g(0_+; \boldsymbol{q}_T(\lambda,\boldsymbol{\psi});\boldsymbol{\psi}),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_3] &\xrightarrow{\mathbb{P}} F_1^2 \cdot \partial_{t_2} g(0_+; \boldsymbol{q}_T(\lambda,\boldsymbol{\psi});\boldsymbol{\psi}) + \tau^2 \cdot \partial_{t_1} g(0_+; \boldsymbol{q}_T(\lambda,\boldsymbol{\psi});\boldsymbol{\psi}),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_1] &\xrightarrow{\mathbb{P}} -\mu_1^2 F_1^2 \cdot \partial_{s1}\partial_{s2} g(0_+; \boldsymbol{q}_T(\lambda,\boldsymbol{\psi});\boldsymbol{\psi}),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_2] &\xrightarrow{\mathbb{P}} -F_1^2 \cdot \partial_{s1}\partial_p g(0_+; \boldsymbol{q}_T(\lambda,\boldsymbol{\psi});\boldsymbol{\psi}),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_3] &\xrightarrow{\mathbb{P}} -F_1^2 \cdot \partial_{s_1}\partial_{t_2} g(0_+; \boldsymbol{q}_T(\lambda,\boldsymbol{\psi});\boldsymbol{\psi}) - \tau^2 \cdot \partial_{s_1}\partial_{t_1} g(0_+; \boldsymbol{q}_T(\lambda,\boldsymbol{\psi});\boldsymbol{\psi}),
\end{aligned}
$$

where $\nabla_{\boldsymbol{q}}^k g(0_+;\boldsymbol{q};\boldsymbol{\psi})$ for $k \in \{1,2\}$ stands for the k'th derivatives (as a vector or a matrix) of $g(\boldsymbol{i}u;\boldsymbol{q};\boldsymbol{\psi})$ with respect to $\boldsymbol{q}$ in the $u \to 0+$ limit (with its elements given by partial derivatives)

$$\nabla_{\boldsymbol{q}}^k g(0_+;\boldsymbol{q};\boldsymbol{\psi}) = \lim_{u\to 0_+} \nabla_{\boldsymbol{q}}^k g(\boldsymbol{i}u;\boldsymbol{q};\boldsymbol{\psi}).$$

As a consequence, we have

$$\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\overline{T}_c(\lambda,N,n,d)] \xrightarrow{\mathbb{P}} \overline{\mathcal{T}}(\lambda,\psi_1,\psi_2), \qquad \mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\psi_1\|\overline{\boldsymbol{a}}_{T,c}(\lambda)\|_2^2] \xrightarrow{\mathbb{P}} \mathcal{A}_T(\lambda,\psi_1,\psi_2),$$

where the definitions of $\overline{\mathcal{T}}$ and $\mathcal{A}_T$ are given in Definition 5. Here $\xrightarrow{\mathbb{P}}$ stands for convergence in probability as $N/d \to \psi_1$ and $n/d \to \psi_2$ (with respect to the randomness of $\boldsymbol{X}$ and $\boldsymbol{\Theta}$).

The Proposition above suggests that Ψ_i and Φ_i concentrates with respect to the randomness in $\boldsymbol{X}$ and $\boldsymbol{\Theta}$. To complete the concentration proof, we need to show that Ψ_i and Φ_i concentrates with respect to the randomness in $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$.

Lemma 4. *Follow the assumptions of Proposition 2. For any $\lambda \in \Lambda_T$, we have*

$$
\begin{aligned}
\mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_1], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_2], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_3] &= o_{d,\mathbb{P}}(1),\\
\mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_1], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_2], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_3] &= o_{d,\mathbb{P}}(1),
\end{aligned}
$$

so that

$$\mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\overline{T}_c(\lambda, N, n, d)], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\|\overline{\boldsymbol{a}}_{T,c}(\lambda)\|_2^2] = o_{d,\mathbb{P}}(1).$$

Here, $o_{d,\mathbb{P}}(1)$ *stands for converges to* 0 *in probability (with respect to the randomness of* $\boldsymbol{X}$ *and* $\boldsymbol{\Theta}$*) as* $N/d \to \psi_1$ *and* $n/d \to \psi_2$ *and* $d \to \infty$*.*

Now, combining Proposition 7 and 4, we have

$$\overline{T}_c(\lambda, N, n, d) \xrightarrow{\mathbb{P}} \overline{\mathcal{T}}(\lambda, \psi_1, \psi_2), \qquad \psi_1 \|\overline{\boldsymbol{a}}_{T,c}(\lambda)\|_2^2 \xrightarrow{\mathbb{P}} \mathcal{A}_T(\lambda, \psi_1, \psi_2).$$

The results above combined with the arguments in Appendix D.2 completes the proof for the asymptotics of $\overline{T}$ and $\psi_1\|\overline{\boldsymbol{a}}_T(\lambda)\|_2^2$.

D.5. Proof of Lemma 1 and Lemma 2

Proof of Lemma 1. Note that by Assumption 4, the matrix $\overline{\boldsymbol{M}} = \boldsymbol{U}_c - \psi_2^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} - \psi_1\lambda\mathbf{I}_N$ is negative definite (so that it is invertible) with high probability. Moreover, whenever $\overline{\boldsymbol{M}}$ is negative definite, the matrix $\boldsymbol{A}(\boldsymbol{q}_U)$ for $\boldsymbol{q}_U = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, \psi_2, 0, 0)$ is also invertible. In the following, we condition on this good event happens.

From the expansion for $\boldsymbol{v}_i$ in (34), we have

$$\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}\Psi_1 = \mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}\Big[\mathrm{Tr}\Big(\overline{\boldsymbol{M}}^{-1}\boldsymbol{v}\boldsymbol{v}^\mathsf{T}\Big)\Big] = \frac{1}{d}\lambda_{d,1}(\sigma)^2 F_1^2 \cdot \Big[\mathrm{Tr}\Big(\overline{\boldsymbol{M}}^{-1}\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T}\Big)\Big] = \frac{1}{d}\mu_1^2 F_1^2 \mathrm{Tr}\left(\overline{\boldsymbol{M}}^{-1}\frac{\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T}}{d}\right) \times (1 + o_d(1)),$$

where we used the relation $\lambda_{d,1} = \mu_1/\sqrt{d} \times (1 + o_d(1))$ as in Eq. (66). Similarly, the second term is

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}\Psi_2 =& -\frac{2}{\psi_2\sqrt{d}}\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}\Big[\mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{v}\boldsymbol{y}^\mathsf{T}\Big)\Big] \\
=& -\frac{2}{\psi_2 d\sqrt{d}}\lambda_{d,1}(\sigma)F_1^2 \cdot \mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{\Theta}\boldsymbol{X}^\mathsf{T}\Big) \\
=& -\frac{2}{\psi_2 d^2}\mu_1 F_1^2 \cdot \mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{\Theta}\boldsymbol{X}^\mathsf{T}\Big) \times (1 + o_d(1)).
\end{aligned}$$

To compute Ψ_3, note we have

$$\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}[\boldsymbol{y}\boldsymbol{y}^\mathsf{T}] = F_1^2 \cdot (\boldsymbol{X}\boldsymbol{X}^\mathsf{T})/d + \tau^2\mathbf{I}_n.$$

This gives the expansion for Ψ_3

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}\Psi_3 =& \psi_2^{-2}d^{-1}\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}\mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{y}\boldsymbol{y}^\mathsf{T}\Big) \\
=& \psi_2^{-2}d^{-2}F_1^2\mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\Big) + \psi_2^{-2}d^{-1}\mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{Z}\Big)\tau^2.
\end{aligned}$$

Through the same algebraic manipulation above, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}\Phi_1 =& \frac{1}{d}\mu_1^2 F_1^2 \mathrm{Tr}\left(\overline{\boldsymbol{M}}^{-2}\frac{\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T}}{d}\right) \times (1 + o_d(1)), \\
\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}\Phi_2 =& -\frac{2}{\psi_2 d^2}\mu_1 F_1^2 \cdot \mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-2}\boldsymbol{\Theta}\boldsymbol{X}^\mathsf{T}\Big) \times (1 + o_d(1)), \\
\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}\Phi_3 =& \psi_2^{-2}d^{-2}F_1^2 \cdot \mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-2}\boldsymbol{Z}^\mathsf{T}\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\Big) + \psi_2^{-2}d^{-1}\tau^2\mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-2}\boldsymbol{Z}^\mathsf{T}\Big).
\end{aligned}$$

Next, we express the trace of matrices products as the derivative of the function $G_d(\xi, \boldsymbol{q})$ (c.f. Definition 1). The derivatives of G_d are (which can we well-defined at $\boldsymbol{q} = \boldsymbol{q}_U = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, \psi_2, 0, 0)$ with high probability by Assumption 4)

$$\partial_{q_i}G_d(0, \boldsymbol{q}) = \frac{1}{d}\mathrm{Tr}(\boldsymbol{A}(\boldsymbol{q})^{-1}\partial_i\boldsymbol{A}(\boldsymbol{q})), \quad \partial_{q_i}\partial_{q_j}G_d(0, \boldsymbol{q}) = -\frac{1}{d}\mathrm{Tr}(\boldsymbol{A}(\boldsymbol{q})^{-1}\partial_{q_i}\boldsymbol{A}(\boldsymbol{q})\boldsymbol{A}(\boldsymbol{q})^{-1}\partial_{q_j}\boldsymbol{A}(\boldsymbol{q})). \tag{47}$$

As an example, we consider evaluating $\partial_{s_2} G_d(0, \boldsymbol{q})$ at $\boldsymbol{q} = \boldsymbol{q}_U \equiv (\mu_\star^2 - \lambda\psi_1, \mu_1^2, \psi_2, 0, 0)$. Using the formula for block matrix inversion, we have

$$\boldsymbol{A}(\mu_\star^2 - \lambda\psi_1, \mu_1^2, \psi_2, 0, 0)^{-1} = \begin{bmatrix} (\mu_\star^2 - \lambda\psi_1)\mathbf{I}_N + \mu_1^2 \boldsymbol{Q} & \boldsymbol{Z}^\mathsf{T} \\ \boldsymbol{Z} & \psi_2 \mathbf{I}_n \end{bmatrix}^{-1} = \begin{bmatrix} (\boldsymbol{U}_c - \psi_2^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} - \psi_1\lambda\mathbf{I}_N)^{-1} & \cdots \\ \cdots & \cdots \end{bmatrix}.$$

Then we have

$$\partial_{s_2} G_d(0, \boldsymbol{q}_U) = \frac{1}{d}\mathrm{Tr}\left(\begin{bmatrix} \overline{\boldsymbol{M}}^{-1} & \cdots \\ \cdots & \cdots \end{bmatrix} \begin{bmatrix} \boldsymbol{Q} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right) = \mathrm{Tr}(\overline{\boldsymbol{M}}^{-1}\boldsymbol{Q})/d.$$

Applying similar argument to compute other derivatives, we get

1. $\mathrm{Tr}(\overline{\boldsymbol{M}}^{-1}\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T})/d^2 = \mathrm{Tr}(\overline{\boldsymbol{M}}^{-1}\boldsymbol{Q})/d = \partial_{s_2} G_d(0, \boldsymbol{q}_U)$.
2. $\mu_1 \cdot \mathrm{Tr}(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{\Theta}\boldsymbol{X}^\mathsf{T})/d^2 = \mathrm{Tr}(\overline{\boldsymbol{M}}^{-1}\boldsymbol{Z}_1^\mathsf{T}\boldsymbol{Z})/d = -\psi_2\partial_p G_d(0, \boldsymbol{q}_U)/2$.
3. $\mathrm{Tr}(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{X}\boldsymbol{X}^\mathsf{T})/d^2 = \mathrm{Tr}(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{H})/d = \psi_2^2\partial_{t_2} G_d(0, \boldsymbol{q}_U) - \psi_2^2$.
4. $\mathrm{Tr}(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{Z}^\mathsf{T})/d = \psi_2^2\partial_{t_1} G_d(0, \boldsymbol{q}_U) - \psi_2^2$.
5. $\mathrm{Tr}(\overline{\boldsymbol{M}}^{-2}\boldsymbol{Q})/d = -\partial_{s_1}\partial_{s_2} G_d(0, \boldsymbol{q}_U)$.
6. $(2/d\psi_2) \cdot \mathrm{Tr}(\boldsymbol{Z}_1^\mathsf{T}\boldsymbol{Z}\overline{\boldsymbol{M}}^{-2}) = \partial_{s_1}\partial_p G_d(0, \boldsymbol{q}_U)$.
7. $\mathrm{Tr}(\overline{\boldsymbol{M}}^{-2}\boldsymbol{Z}^\mathsf{T}\boldsymbol{H}\boldsymbol{Z})/(d\psi_2^2) = -\partial_{s_1}\partial_{t_2} G_d(0, \boldsymbol{q}_U)$.
8. $\mathrm{Tr}(\overline{\boldsymbol{M}}^{-2}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z})/(d\psi_2^2) = -\partial_{s_1}\partial_{t_1} G_d(0, \boldsymbol{q}_U)$.

Combining these equations concludes the proof. □

Proof of Lemma 2. We prove this lemma by assuming that $\boldsymbol{\beta}$ follows a different distribution: $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, (\|F_1\|_2^2/d)\mathbf{I}_d)$. The case when $\boldsymbol{\beta} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(F_1))$ can be treated similarly.

By directly calculating the variance, we can show that, there exists scalers $(c_{ik}^{(d)})_{k\in[K_i]}$ with $c_{ik}^{(d)} = \Theta_d(1)$, and matrices $(\boldsymbol{A}_{ik}, \boldsymbol{B}_{ik})_{k\in[K_i]} \subseteq \{\mathbf{I}_N, \boldsymbol{Q}, \boldsymbol{Z}^\mathsf{T}\boldsymbol{H}\boldsymbol{Z}, \boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\}$, such that the variance of Ψ_i's can be expressed in form

$$\mathrm{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}(\Psi_i) = \frac{1}{d}\sum_{k=1}^{K_i} c_{ik}^{(d)}\mathrm{Tr}(\overline{\boldsymbol{M}}^{-1}\boldsymbol{A}_{ik}\overline{\boldsymbol{M}}^{-1}\boldsymbol{B}_{ik})/d.$$

For example, by Lemma 8, we have

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{\beta}\sim\mathcal{N}(\mathbf{0},(F_1^2/d)\mathbf{I}_d)}(\Psi_1) &= \lambda_{d,1}(\sigma)^4\mathrm{Var}_{\boldsymbol{\beta}\sim\mathcal{N}(\mathbf{0},(F_1^2/d)\mathbf{I}_d))}(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Theta}^\mathsf{T}\overline{\boldsymbol{M}}^{-1}\boldsymbol{\Theta}\boldsymbol{\beta}) = 2\lambda_{d,1}(\sigma)^4F_1^4\|\boldsymbol{\Theta}^\mathsf{T}\overline{\boldsymbol{M}}^{-1}\boldsymbol{\Theta}\|_F^2/d^2 \\ &= c_1^{(d)}\mathrm{Tr}(\overline{\boldsymbol{M}}^{-1}\boldsymbol{Q}\overline{\boldsymbol{M}}^{-1}\boldsymbol{Q})/d^2,\end{aligned}$$

where $c_1^{(d)} = 2d^2\lambda_{d,1}(\sigma)^4F_1^4 = O_d(1)$. The variance of Ψ_2 and Ψ_3 can be calculated similarly.

Note that each $\mathrm{Tr}(\overline{\boldsymbol{M}}^{-1}\boldsymbol{A}_{ik}\overline{\boldsymbol{M}}^{-1}\boldsymbol{B}_{ik})/d$ can be expressed as an entry of $\nabla_{\boldsymbol{q}}^2 G_d(0; \boldsymbol{q})$ (c.f. Eq. (47)), and by Proposition 4, they are of order $O_{d,\mathbb{P}}(1)$. This gives

$$\mathrm{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}(\Psi_i) = o_{d,\mathbb{P}}(1).$$

Similarly, for the same set of scalers $(c_{ik}^{(d)})_{k\in[K_i]}$ and matrices $(\boldsymbol{A}_{ik}, \boldsymbol{B}_{ik})_{k\in[K_i]}$, we have

$$\mathrm{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}(\Phi_i) = \frac{1}{d}\sum_{k=1}^{K_i} c_{ik}\mathrm{Tr}(\overline{\boldsymbol{M}}^{-2}\boldsymbol{A}_{ik}\overline{\boldsymbol{M}}^{-2}\boldsymbol{B}_{ik})/d.$$

Note that for two semidefinite matrices $\boldsymbol{A}, \boldsymbol{B}$, we have $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{B}) \le \|\boldsymbol{A}\|_{\mathrm{op}}\mathrm{Tr}(\boldsymbol{B})$. Moreover, note we have $\|\overline{\boldsymbol{M}}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$ (by Assumption 4). This gives

$$\mathrm{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}(\Phi_i) = o_{d,\mathbb{P}}(1).$$

This concludes the proof. □

D.6. Auxiliary Lemmas

The following lemma (Lemma 5) is a reformulation of Proposition 3 in (Ghorbani et al., 2019). We present it in a stronger form, but it can be easily derived from the proof of Proposition 3 in (Ghorbani et al., 2019). This lemma was first proved in (El Karoui, 2010) in the Gaussian case. (Notice that the second estimate —on $Q_k(\boldsymbol{\Theta}\boldsymbol{X}^{\mathsf{T}})$— follows by applying the first one whereby $\boldsymbol{\Theta}$ is replaced by $\boldsymbol{W} = [\boldsymbol{\Theta}^{\mathsf{T}}|\boldsymbol{X}^{\mathsf{T}}]^{\mathsf{T}}$

Lemma 5. *Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)^{\mathsf{T}} \in \mathbb{R}^{N\times d}$ with $(\boldsymbol{\theta}_a)_{a\in[N]} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{\mathsf{T}} \in \mathbb{R}^{n\times d}$ with $(\boldsymbol{x}_i)_{i\in[n]} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$. Assume $1/c \le n/d, N/d \le c$ for some constant $c \in (0,\infty)$. Then*

$$\mathbb{E}\Big[\sup_{k\ge 2} \|Q_k(\boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathsf{T}}) - \mathbf{I}_N\|_{\mathrm{op}}^2\Big] = o_d(1)\,, \tag{48}$$

$$\mathbb{E}\Big[\sup_{k\ge 2} \|Q_k(\boldsymbol{\Theta}\boldsymbol{X}^{\mathsf{T}})\|_{\mathrm{op}}^2\Big] = o_d(1). \tag{49}$$

Notice that the second estimate —on $Q_k(\boldsymbol{\Theta}\boldsymbol{X}^{\mathsf{T}})$— follows by applying the first one —Eq. (48)— whereby $\boldsymbol{\Theta}$ is replaced by $\boldsymbol{W} = [\boldsymbol{\Theta}^{\mathsf{T}}|\boldsymbol{X}^{\mathsf{T}}]^{\mathsf{T}}$, and we use $\|Q_k(\boldsymbol{\Theta}\boldsymbol{X}^{\mathsf{T}})\|_{\mathrm{op}} \le \|Q_k(\boldsymbol{W}\boldsymbol{W}^{\mathsf{T}}) - \mathbf{I}_{N+n}\|_{\mathrm{op}}$.

The following lemma (Lemma 6) can be easily derived from Lemma 5. Again, this lemma was first proved in (El Karoui, 2010) in the Gaussian case.

Lemma 6. *Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)^{\mathsf{T}} \in \mathbb{R}^{N\times d}$ with $(\boldsymbol{\theta}_a)_{a\in[N]} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$. Let activation function σ satisfies Assumption 2. Assume $1/c \le N/d \le c$ for some constant $c \in (0,\infty)$. Denote*

$$\boldsymbol{U} = \Big(\mathbb{E}_{\boldsymbol{x}\sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))}[\sigma(\langle \boldsymbol{\theta}_a, \boldsymbol{x}\rangle/\sqrt{d})\sigma(\langle \boldsymbol{\theta}_b, \boldsymbol{x}\rangle/\sqrt{d})]\Big)_{a,b\in[N]} \in \mathbb{R}^{N\times N}.$$

Then we can rewrite the matrix $\boldsymbol{U}$ to be

$$\boldsymbol{U} = \lambda_{d,0}(\sigma)^2 \mathbf{1}_N \mathbf{1}_N^{\mathsf{T}} + \mu_1^2 \boldsymbol{Q} + \mu_\star^2(\mathbf{I}_N + \boldsymbol{\Delta}),$$

with $\boldsymbol{Q} = \boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathsf{T}}/d$ and $\mathbb{E}[\|\boldsymbol{\Delta}\|_{\mathrm{op}}^2] = o_d(1)$.

In the following, we show that, under sufficient regularity condition of σ, we have $\lambda_{d,0}(\sigma) = O(1/d)$.

Lemma 7. *Let $\sigma \in C^2(\mathbb{R})$ with $|\sigma'(x)|, |\sigma''(x)| < c_0 e^{c_1|x|}$ for some $c_0, c_1 \in \mathbb{R}$. Assume that $\mathbb{E}_{G\sim\mathcal{N}(0,1)}[\sigma(G)] = 0$. Then we have*

$$\lambda_{d,0}(\sigma) \equiv \mathbb{E}_{\boldsymbol{x}\sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))}[\sigma(x_1)] = O(1/d).$$

Proof of Lemma 7. Let $\boldsymbol{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and $\gamma \sim \chi(d)/\sqrt{d}$ independently. Then we have $\gamma\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, so that by the assumption, we have $\mathbb{E}[\sigma(\gamma x_1)] = 0$.

As a consequence, by the second order Taylor expansion, and by the independence of γ and $\boldsymbol{x}$, we have (for $\xi(x_1) \in [\gamma, 1]$)

$$\begin{aligned}
|\lambda_{d,0}(\sigma)| = |\mathbb{E}[\sigma(x_1)]| \le |\mathbb{E}[\sigma(x_1)] - \mathbb{E}[\sigma(\gamma x_1)]| &\le \Big|\mathbb{E}[\sigma'(x_1)x_1]\mathbb{E}[\gamma - 1]\Big| + \Big|(1/2)\mathbb{E}[\sigma''(\xi(x_1)x_1)(\gamma-1)^2]\Big| \\
&\le \Big|\mathbb{E}[\sigma'(x_1)x_1]\Big| \cdot \Big|\mathbb{E}[\gamma - 1]\Big| + (1/2)\mathbb{E}\Big[\sup_{u\in[\gamma,1]} \sigma''(ux_1)^2\Big]^{1/2} \mathbb{E}[(\gamma-1)^4]^{1/2}.
\end{aligned}$$

By the assumption that $|\sigma'(x)|, |\sigma''(x)| < c_0 e^{c_1|x|}$ for some $c_0, c_1 \in \mathbb{R}$, there exists constant K that only depends on c_0 and c_1 such that

$$\sup_d \Big|\mathbb{E}[\sigma'(x_1)x_1]\Big| \le K, \qquad \sup_d \Big|(1/2)\mathbb{E}\Big[\sup_{u\in[\gamma,1]} \sigma''(ux_1)^2\Big]^{1/2}\Big| \le K.$$

Moreover, by property of the χ distribution, we have

$$|\mathbb{E}[\gamma - 1]| = O(d^{-1}), \qquad \mathbb{E}[(\gamma-1)^4]^{1/2} = O(d^{-1}).$$

This concludes the proof. $\square$

The following lemma is a simple variance calculation and can be found as Lemma C.5 in (Mei & Montanari, 2019). We restate here for completeness.

Lemma 8. *Let $\boldsymbol{A} \in \mathbb{R}^{n\times N}$ and $\boldsymbol{B} \in \mathbb{R}^{n\times n}$. Let $\boldsymbol{g} = (g_1, \ldots, g_n)^\mathsf{T}$ with $g_i \sim_{iid} \mathbb{P}_g$, $\mathbb{E}_g[g] = 0$, and $\mathbb{E}_g[g^2] = 1$. Let $\boldsymbol{h} = (h_1, \ldots, h_N)^\mathsf{T}$ with $h_i \sim_{iid} \mathbb{P}_h$, $\mathbb{E}_h[h] = 0$, and $\mathbb{E}_h[h^2] = 1$. Further we assume that $\boldsymbol{h}$ is independent of $\boldsymbol{g}$. Then we have*

$$
\begin{aligned}
\mathit{Var}(\boldsymbol{g}^\mathsf{T}\boldsymbol{A}\boldsymbol{h}) &= \|\boldsymbol{A}\|_F^2,\\
\mathit{Var}(\boldsymbol{g}^\mathsf{T}\boldsymbol{B}\boldsymbol{g}) &= \sum_{i=1}^n B_{ii}^2(\mathbb{E}[g^4]-3) + \|\boldsymbol{B}\|_F^2 + \mathrm{Tr}(\boldsymbol{B}^2).
\end{aligned}
$$

E. Proof of Theorem 1

Here we give the whole proof for U. The proof for T is the same.

For fixed $A^2 \in \Gamma_U \equiv \{\mathcal{A}_U(\lambda, \psi_1, \psi_2) : \lambda \in \Lambda_U\}$, we denote

$$
\lambda_\star(A^2) = \inf_\lambda \Big\{\lambda : \mathcal{A}_U(\lambda, \psi_1, \psi_2) = A^2\Big\}.
$$

By the definition of Γ_U, the set $\{\lambda : \mathcal{A}_U(\lambda, \psi_1, \psi_2) = A^2\}$ is non-empty and lower bounded, so that $\lambda_\star(A^2)$ can be well-defined. Moreover, we have $\lambda_\star(A^2) \in \Lambda_U$. It is also easy to see that we have

$$
\lambda_\star(A^2) \in \arg\min_{\lambda\ge 0}\Big[\overline{\mathcal{U}}(\lambda, \psi_1, \psi_2) + \lambda A^2\Big]. \tag{50}
$$

E.1. Upper bound

Note we have

$$
\begin{aligned}
U(A, N, n, d) = &\sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A^2}\Big(R(\boldsymbol{a}) - \widehat{R}_n(\boldsymbol{a})\Big)\\
&\le \inf_\lambda \sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A^2}\Big(R(\boldsymbol{a}) - \widehat{R}_n(\boldsymbol{a}) - \psi_1\lambda(\|\boldsymbol{a}\|_2^2 - \psi_1^{-1}A^2)\Big)\\
&\le \inf_\lambda \Big[\overline{U}(\lambda, N, n, d) + \lambda A^2\Big]\\
&\le \overline{U}(\lambda_\star(A^2), N, n, d) + \lambda_\star(A^2)A^2.
\end{aligned}
$$

Note that $\lambda_\star(A^2) \in \Lambda_U$, so by Lemma 5, in the limit of Assumption 3, we have

$$
U(A, N, n, d) \le \overline{\mathcal{U}}(\lambda_\star(A^2), \psi_1, \psi_2) + \lambda_\star(A^2)A^2 + o_{d,\mathbb{P}}(1) = \mathcal{U}(A, \psi_1, \psi_2) + o_{d,\mathbb{P}}(1),
$$

where the last equality is by Eq. (50). This proves the upper bound.

E.2. Lower bound

For any $A^2 > 0$, we define a random variable $\hat\lambda(A^2)$ (which depend on $\boldsymbol{X}$, $\boldsymbol{\Theta}$, $\boldsymbol{\beta}$, $\boldsymbol{\varepsilon}$) by

$$
\hat\lambda(A^2) = \inf\Big\{\lambda : \lambda \in \arg\min_{\lambda\ge 0}\Big[\overline{U}(\lambda, N, n, d) + \lambda A^2\Big]\Big\}.
$$

By Proposition 1, the set is should always be non-empty, so that $\hat\lambda(A^2)$ can always be well-defined.

Moreover, since $\lambda_\star(A^2) \in \Lambda_U$, by Assumption 4, as we have shown in the proof in Proposition 2, we can uniquely define $\overline{\boldsymbol{a}}_U(\lambda_\star(A^2))$ with high probability, where

$$
\overline{\boldsymbol{a}}_U(\lambda_\star(A^2)) = \arg\max_{\boldsymbol{a}}\Big[R(\boldsymbol{a}) - \widehat{R}_n(\boldsymbol{a}) - \psi_1\lambda_\star(A^2)\|\boldsymbol{a}\|_2^2\Big].
$$

As a consequence, for a small $\varepsilon > 0$, the following event $\mathcal{E}_{\varepsilon,d}$ can be well-defined with high probability

$$
\begin{aligned}
\mathcal{E}_{\varepsilon,d} &= \Big\{\psi_1\|\overline{\boldsymbol{a}}_U(\lambda_\star(A^2))\|_2^2 \ge A^2 - \varepsilon\Big\} \cap \Big\{\hat\lambda(A^2+\varepsilon) \le \lambda_\star(A^2)\Big\}\\
&= \Big\{A^2 - \varepsilon \le \psi_1\|\overline{\boldsymbol{a}}_U(\lambda_\star(A^2))\|_2^2 \le A^2 + \varepsilon\Big\}.
\end{aligned}
$$

Now, by Proposition 2, in the limit of Assumption 3, we have

$$\lim_{d\to\infty} \mathbb{P}_{\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\beta},\boldsymbol{\varepsilon}}(\mathcal{E}_{\varepsilon,d}) = 1, \tag{51}$$

and we have

$$\mathcal{U}(\lambda_\star(A^2),\psi_1,\psi_2) = \overline{\mathcal{U}}(\lambda_\star(A^2),\psi_1,\psi_2) + o_{d,\mathbb{P}}(1). \tag{52}$$

By the strong duality as in Proposition 1, for any $A^2 \in \Gamma_U$, we have

$$U(A,N,n,d) = \overline{U}(\hat\lambda(A^2),N,n,d) + \hat\lambda(A^2)A^2.$$

Consequently, for small $\varepsilon > 0$, when the event $\mathcal{E}_{\varepsilon,d}$ happens, we have

$$\begin{aligned}
&U((A^2+\varepsilon)^{1/2},N,n,d)\\
=&\sup_{\boldsymbol{a}} \Big(R(\boldsymbol{a}) - \widehat{R}_n(\boldsymbol{a}) - \psi_1\hat\lambda(A^2+\varepsilon)\cdot\big(\|\boldsymbol{a}\|_2^2 - \psi_1^{-1}(A^2+\varepsilon)\big)\Big)\\
\ge& R(\overline{\boldsymbol{a}}_U(\lambda_\star(A^2))) - \widehat{R}_n(\overline{\boldsymbol{a}}_U(\lambda_\star(A^2))) - \psi_1\hat\lambda(A^2+\varepsilon)\cdot\big(\|\overline{\boldsymbol{a}}_U(\lambda_\star(A^2))\|_2^2 - \psi_1^{-1}(A^2+\varepsilon)\big)\\
\ge& R(\overline{\boldsymbol{a}}_U(\lambda_\star(A^2))) - \widehat{R}_n(\overline{\boldsymbol{a}}_U(\lambda_\star(A^2))) - \psi_1\hat\lambda(A^2+\varepsilon)\cdot\big(\|\overline{\boldsymbol{a}}_U(\lambda_\star(A^2))\|_2^2 - \psi_1^{-1}(A^2-\varepsilon)\big)\\
\ge& R(\overline{\boldsymbol{a}}_U(\lambda_\star(A^2))) - \widehat{R}_n(\overline{\boldsymbol{a}}_U(\lambda_\star(A^2))) - \psi_1\lambda_\star(A^2)\cdot\big(\|\overline{\boldsymbol{a}}_U(\lambda_\star(A^2))\|_2^2 - \psi_1^{-1}(A^2-\varepsilon)\big)\\
=&\overline{U}(\lambda_\star(A^2),N,n,d) + \lambda_\star(A^2)\cdot(A^2-\varepsilon).
\end{aligned}$$

As a consequence, by Eq. (51) and (52), we have

$$U((A^2+\varepsilon)^{1/2},N,n,d) \ge \overline{\mathcal{U}}(\lambda_\star(A^2),\psi_1,\psi_2) + \lambda_\star(A^2)\cdot(A^2-\varepsilon) - o_{d,\mathbb{P}}(1) = \mathcal{U}(A,\psi_1,\psi_2) - \varepsilon\lambda_\star(A^2) - o_{d,\mathbb{P}}(1).$$

where the last equality is by the definition of $\mathcal{U}$ as in Definition 2, and by the fact that $\lambda_\star(A^2) \in \arg\min_{\lambda\ge 0}[\overline{\mathcal{U}}(\lambda,\psi_1,\psi_2) + \lambda A^2]$. Taking ε sufficiently small proves the lower bound. This concludes the proof of Theorem 1.

F. Technical background

In this section we introduce additional technical background useful for the proofs. In particular, we will use decompositions in (hyper-)spherical harmonics on the $\mathbb{S}^{d-1}(\sqrt{d})$ and in Hermite polynomials on the real line. We refer the readers to (Efthimiou & Frye, 2014; Szego, Gabor, 1939; Chihara, 2011; Ghorbani et al., 2019; Mei & Montanari, 2019) for further information on these topics.

F.1. Functional spaces over the sphere

For $d \ge 1$, we let $\mathbb{S}^{d-1}(r) = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = r\}$ denote the sphere with radius r in $\mathbb{R}^d$. We will mostly work with the sphere of radius $\sqrt{d}$, $\mathbb{S}^{d-1}(\sqrt{d})$ and will denote by γ_d the uniform probability measure on $\mathbb{S}^{d-1}(\sqrt{d})$. All functions in the following are assumed to be elements of $L^2(\mathbb{S}^{d-1}(\sqrt{d}),\gamma_d)$, with scalar product and norm denoted as $\langle\,\cdot\,,\,\cdot\,\rangle_{L^2}$ and $\|\,\cdot\,\|_{L^2}$:

$$\langle f,g\rangle_{L^2} \equiv \int_{\mathbb{S}^{d-1}(\sqrt{d})} f(\boldsymbol{x})\,g(\boldsymbol{x})\,\gamma_d(\mathrm{d}\boldsymbol{x})\,. \tag{53}$$

For $\ell \in \mathbb{Z}_{\ge 0}$, let $\tilde V_{d,\ell}$ be the space of homogeneous harmonic polynomials of degree ℓ on $\mathbb{R}^d$ (i.e. homogeneous polynomials $q(\boldsymbol{x})$ satisfying $\Delta q(\boldsymbol{x}) = 0$), and denote by $V_{d,\ell}$ the linear space of functions obtained by restricting the polynomials in $\tilde V_{d,\ell}$ to $\mathbb{S}^{d-1}(\sqrt{d})$. With these definitions, we have the following orthogonal decomposition

$$L^2(\mathbb{S}^{d-1}(\sqrt{d}),\gamma_d) = \bigoplus_{\ell=0}^{\infty} V_{d,\ell}\,. \tag{54}$$

The dimension of each subspace is given by

$$\dim(V_{d,\ell}) = B(d,\ell) = \frac{2\ell+d-2}{\ell}\binom{\ell+d-3}{\ell-1}\,. \tag{55}$$

For each $\ell \in \mathbb{Z}_{\geq 0}$, the spherical harmonics $\{Y_{\ell j}^{(d)}\}_{1 \leq j \leq B(d,\ell)}$ form an orthonormal basis of $V_{d,\ell}$:

$$\langle Y_{ki}^{(d)}, Y_{sj}^{(d)} \rangle_{L^2} = \delta_{ij}\delta_{ks}.$$

Note that our convention is different from the more standard one, that defines the spherical harmonics as functions on $\mathbb{S}^{d-1}(1)$. It is immediate to pass from one convention to the other by a simple scaling. We will drop the superscript d and write $Y_{\ell,j} = Y_{\ell,j}^{(d)}$ whenever clear from the context.

We denote by P_k the orthogonal projections to $V_{d,k}$ in $L^2(\mathbb{S}^{d-1}(\sqrt{d}), \gamma_d)$. This can be written in terms of spherical harmonics as

$$\mathsf{P}_k f(\boldsymbol{x}) \equiv \sum_{l=1}^{B(d,k)} \langle f, Y_{kl} \rangle_{L^2} Y_{kl}(\boldsymbol{x}). \tag{56}$$

Then for a function $f \in L^2(\mathbb{S}^{d-1}(\sqrt{d}))$, we have

$$f(\boldsymbol{x}) = \sum_{k=0}^{\infty} \mathsf{P}_k f(\boldsymbol{x}) = \sum_{k=0}^{\infty} \sum_{l=1}^{B(d,k)} \langle f, Y_{kl} \rangle_{L^2} Y_{kl}(\boldsymbol{x}).$$

F.2. Gegenbauer polynomials

The ℓ-th Gegenbauer polynomial $Q_\ell^{(d)}$ is a polynomial of degree ℓ. Consistently with our convention for spherical harmonics, we view $Q_\ell^{(d)}$ as a function $Q_\ell^{(d)} : [-d, d] \to \mathbb{R}$. The set $\{Q_\ell^{(d)}\}_{\ell \geq 0}$ forms an orthogonal basis on $L^2([-d, d], \tilde{\tau}_d)$ (where $\tilde{\tau}_d$ is the distribution of $\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle$ when $\boldsymbol{x}_1, \boldsymbol{x}_2 \sim_{i.i.d.} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$), satisfying the normalization condition:

$$\langle Q_k^{(d)}, Q_j^{(d)} \rangle_{L^2(\tilde{\tau}_d)} = \frac{1}{B(d,k)} \, \delta_{jk} \, . \tag{57}$$

In particular, these polynomials are normalized so that $Q_\ell^{(d)}(d) = 1$. As above, we will omit the superscript d when clear from the context (write it as Q_ℓ for notation simplicity).

Gegenbauer polynomials are directly related to spherical harmonics as follows. Fix $\boldsymbol{v} \in \mathbb{S}^{d-1}(\sqrt{d})$ and consider the subspace of V_ℓ formed by all functions that are invariant under rotations in $\mathbb{R}^d$ that keep $\boldsymbol{v}$ unchanged. It is not hard to see that this subspace has dimension one, and coincides with the span of the function $Q_\ell^{(d)}(\langle \boldsymbol{v}, \cdot \rangle)$.

We will use the following properties of Gegenbauer polynomials

1. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{d-1}(\sqrt{d})$

$$\langle Q_j^{(d)}(\langle \boldsymbol{x}, \cdot \rangle), Q_k^{(d)}(\langle \boldsymbol{y}, \cdot \rangle) \rangle_{L^2(\mathbb{S}^{d-1}(\sqrt{d}), \gamma_d)} = \frac{1}{B(d,k)} \delta_{jk} Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y} \rangle). \tag{58}$$

2. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{d-1}(\sqrt{d})$

$$Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y} \rangle) = \frac{1}{B(d,k)} \sum_{i=1}^{B(d,k)} Y_{ki}^{(d)}(\boldsymbol{x}) Y_{ki}^{(d)}(\boldsymbol{y}). \tag{59}$$

Note in particular that property 2 implies that –up to a constant– $Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y} \rangle)$ is a representation of the projector onto the subspace of degree-k spherical harmonics

$$(\mathsf{P}_k f)(\boldsymbol{x}) = B(d,k) \int_{\mathbb{S}^{d-1}(\sqrt{d})} Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y} \rangle) \, f(\boldsymbol{y}) \, \gamma_d(\mathrm{d}\boldsymbol{y}) \, . \tag{60}$$

For a function $\sigma \in L^2([-\sqrt{d}, \sqrt{d}], \tau_d)$ (where τ_d is the distribution of $\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle/\sqrt{d}$ when $\boldsymbol{x}_1, \boldsymbol{x}_2 \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$), denoting its spherical harmonics coefficients $\lambda_{d,k}(\sigma)$ to be

$$\lambda_{d,k}(\sigma) = \int_{[-\sqrt{d},\sqrt{d}]} \sigma(x) Q_k^{(d)}(\sqrt{d}x) \tau_d(x), \tag{61}$$

then we have the following equation holds in $L^2([-\sqrt{d}, \sqrt{d}], \tau_d)$ sense

$$\sigma(x) = \sum_{k=0}^{\infty} \lambda_{d,k}(\sigma) B(d,k) Q_k^{(d)}(\sqrt{d}x). \tag{62}$$

F.3. Hermite polynomials

The Hermite polynomials $\{\mathrm{He}_k\}_{k\geq 0}$ form an orthogonal basis of $L^2(\mathbb{R}, \mu_G)$, where $\mu_G(\mathrm{d}x) = e^{-x^2/2}\mathrm{d}x/\sqrt{2\pi}$ is the standard Gaussian measure, and He_k has degree k. We will follow the classical normalization (here and below, expectation is with respect to $G \sim \mathsf{N}(0,1)$):

$$\mathbb{E}\big\{\mathrm{He}_j(G)\,\mathrm{He}_k(G)\big\} = k!\,\delta_{jk}\,. \tag{63}$$

As a consequence, for any function $\sigma \in L^2(\mathbb{R}, \mu_G)$, we have the decomposition

$$\sigma(x) = \sum_{k=1}^{\infty} \frac{\mu_k(\sigma)}{k!}\,\mathrm{He}_k(x)\,, \qquad \mu_k(\sigma) \equiv \mathbb{E}\big\{\sigma(G)\,\mathrm{He}_k(G)\big\}\,. \tag{64}$$

The Hermite polynomials can be obtained as high-dimensional limits of the Gegenbauer polynomials introduced in the previous section. Indeed, the Gegenbauer polynomials (up to a $\sqrt{d}$ scaling in domain) are constructed by Gram-Schmidt orthogonalization of the monomials $\{x^k\}_{k\geq 0}$ with respect to the measure τ_d, while Hermite polynomial are obtained by Gram-Schmidt orthogonalization with respect to μ_G. Since $\tau_d \Rightarrow \mu_G$ (here $\Rightarrow$ denotes weak convergence), it is immediate to show that, for any fixed integer k,

$$\lim_{d\to\infty} \mathrm{Coeff}\{Q_k^{(d)}(\sqrt{d}x)\,B(d,k)^{1/2}\} = \mathrm{Coeff}\left\{\frac{1}{(k!)^{1/2}}\,\mathrm{He}_k(x)\right\}\,. \tag{65}$$

Here and below, for P a polynomial, $\mathrm{Coeff}\{P(x)\}$ is the vector of the coefficients of P. As a consequence, for any fixed integer k, we have

$$\mu_k(\sigma) = \lim_{d\to\infty} \lambda_{d,k}(\sigma)(B(d,k)k!)^{1/2}, \tag{66}$$

where $\mu_k(\sigma)$ and $\lambda_{d,k}(\sigma)$ are given in Eq. (64) and (61).