A. Details of σ_l and other Φ^{adv}

Assumption of σ_l in [25]. To prove Eq. (2) according to Theorem 1.5 in [76], [25] considers $f_{\widetilde{\mathbf{W}}}$ such that $\left| ||\mathbf{W}_l||_2 - ||\widetilde{\mathbf{W}}_l||_2 \right| \leq \frac{1}{n} ||\widetilde{\mathbf{W}}_l||_2$, then they assume $\sigma_l = \frac{||\widetilde{\mathbf{W}}_l||_2}{\beta_{\widetilde{\mathbf{W}}_l}} \sigma$, where $\beta_{\widetilde{\mathbf{W}}_l} := \left(\prod_{l=1}^n ||\widetilde{\mathbf{W}}_l||_2 \right)^{\frac{1}{n}}$.

Assumption of σ_l in [54]. To prove the PAC-Bayesian bound in [54] according to Theorem 1.5 in [76], [54] assumes all variances are same across layers, that is, $\sigma_l = \sigma$.

Our assumption of σ_l. We can prove Lem. 3.2 under both of above assumptions. To make the main paper more clear, we assume that $\sigma_l = \sigma$ in the main paper. And we provide the proofs of Lem. 3.2 for $\sigma_l = \sigma$ and $\sigma_l = \frac{||\widetilde{\mathbf{W}}_l||_2}{\beta_{\widetilde{\mathbf{W}}_l}} \sigma$ in Appendix B (the assumption of $\sigma_l = \frac{||\widetilde{\mathbf{W}}_l||_2}{\beta_{\widetilde{\mathbf{W}}_l}} \sigma$ includes the assumption of $\sigma_l = \sigma$).

PGM attack for Φ^{adv}. For a PGM attack with noise power ϵ given Euclidean norm $||\cdot||$, r iterations for attack and step size $\mathcal{Z}$, let $\kappa \leq ||\nabla_{\mathbf{s}''} \mathcal{L}(f_{\mathbf{W}}(\mathbf{s}''))||$ hold for every $\mathbf{s}'' \in \{\mathcal{D} \cup \mathcal{D}'\}$ with constant $\kappa > 0$, then we get [25]

$$\Phi^{\mathrm{adv}} = \Big\{ \prod_{l=1}^n ||\mathbf{W}_l||_2 \Big(1 + \frac{\mathcal{Z}}{\kappa} \frac{1 - (2\mathcal{Z}/\kappa)^r \overline{\mathrm{lip}}(\nabla \mathcal{L} \circ f_{\mathbf{W}})^r}{1 - (2\mathcal{Z}/\kappa) \overline{\mathrm{lip}}(\nabla \mathcal{L} \circ f_{\mathbf{W}})} (\prod_{l=1}^n ||\mathbf{W}_l||_2) \sum_{l=1}^n \prod_{j=1}^l ||\mathbf{W}_j||_2 \Big) \Big\}^2 \sum_{l=1}^n \frac{||\mathbf{W}_l||_F^2}{||\mathbf{W}_l||_2^2}, \tag{22}$$

where

$$\overline{\mathrm{lip}}(\nabla \mathcal{L} \circ f_{\mathbf{W}}) := (\prod_{l=1}^n ||\mathbf{W}_l||_2) \sum_{l=1}^n \prod_{j=1}^l ||\mathbf{W}_j||_2$$

gives an upper bound on the Lipschitz constant of $\nabla_{\mathbf{s}} \mathcal{L}(f_{\mathbf{W}}(\mathbf{s}))$.

B. Proof of Lem. 3.2

We provide our proofs based on the proofs in [25], to be clearer about the proofs, we suggest readers go through Appendix C.2 in [25] firstly. To prove Eq. (2), [25] considers $f_{\widetilde{\mathbf{W}}}$ such that $\left| ||\mathbf{W}_l||_2 - ||\widetilde{\mathbf{W}}_l||_2 \right| \leq \frac{1}{n} ||\widetilde{\mathbf{W}}_l||_2$, since $(1 + \frac{1}{n})^n \leq e$ and $\frac{1}{e} \leq (1 - \frac{1}{n})^{n-1}$, we get

$$\left(\frac{1}{e}\right)^{\frac{n}{n-1}} \prod_{l=1}^n ||\widetilde{\mathbf{W}}_l||_2 \leq \prod_{l=1}^n ||\mathbf{W}_l||_2 \leq e \prod_{l=1}^n ||\widetilde{\mathbf{W}}_l||_2, \tag{23}$$

and for each j, we get

$$\frac{1}{||\mathbf{W}_j||_2} \prod_{l=1}^n ||\mathbf{W}_l||_2 \leq \frac{e}{||\widetilde{\mathbf{W}}_j||_2} \prod_{l=1}^n ||\widetilde{\mathbf{W}}_l||_2 \tag{24}$$

and

$$\begin{aligned} \frac{1}{||\widetilde{\mathbf{W}}_j||_2} \prod_{l=1}^n ||\widetilde{\mathbf{W}}_l||_2 &\leq (1 - \frac{1}{n})^{-(n-1)} \frac{1}{||\mathbf{W}_j||_2} \prod_{l=1}^n ||\mathbf{W}_l||_2 \\ &\leq \frac{e}{||\mathbf{W}_j||_2} \prod_{l=1}^n ||\mathbf{W}_l||_2. \end{aligned} \tag{25}$$

Then let $\sigma_l = \sigma$ ($\frac{||\widetilde{\mathbf{W}}_l||_2}{\beta_{\widetilde{\mathbf{W}}_l}} = 1$) or $\sigma_l = \frac{||\widetilde{\mathbf{W}}_l||_2}{\beta_{\widetilde{\mathbf{W}}_l}} \sigma$ and let FGM perturbs vector be

$$\delta_{\mathbf{W}}^{\mathrm{fgm}}(\mathbf{s}) := \arg\max_{||\boldsymbol{\delta}|| \leq \epsilon} \boldsymbol{\delta}^\intercal \nabla_{\mathbf{s}} \mathcal{L}(f_{\mathbf{W}}(\mathbf{s})). \tag{26}$$

According to Appendix C.2 Eq. (22) in [25], we get the following inequation

$$\begin{aligned} &||f_{\mathbf{W}+\mathbf{U}}\big(\mathbf{s} + \delta_{\mathbf{W}+\mathbf{U}}^{\mathrm{fgm}}(\mathbf{s})\big) - f_{\mathbf{W}}\big(\mathbf{s} + \delta_{\mathbf{W}}^{\mathrm{fgm}}(\mathbf{s})\big)|| \\ &\leq e(B + \epsilon) \prod_{l=1}^n ||\mathbf{W}_l||_2 \sum_{l=1}^n \frac{||\mathbf{U}_l||_2}{||\mathbf{W}_l||_2} \\ &\quad + 2e^2 \frac{\epsilon}{\kappa} \prod_{l=1}^n ||\mathbf{W}_l||_2^2 \sum_{l=1}^n \Big[\frac{||\mathbf{U}_l||_2}{||\mathbf{W}_l||_2} \\ &\quad + B(\prod_{j=1}^l ||\mathbf{W}_j||_2) \sum_{j=1}^l \frac{||\mathbf{U}_j||_2}{||\mathbf{W}_j||_2} \Big]. \end{aligned} \tag{27}$$

According to Section 1.1 in [5], we have

$$\begin{aligned} \mathbb{E}||\mathbf{U}_l||_2 &\lesssim (1 + \sqrt{\ln h}) ||\mathbb{E}(\mathbf{U}_l^\intercal \mathbf{U}_l)||_2^{\frac{1}{2}} + ||\mathbb{E}(\mathbf{U}_l \mathbf{U}_l^\intercal)||_2^{\frac{1}{2}} \\ &\leq c\Big((1 + \sqrt{\ln h}) ||\mathbb{E}(\mathbf{U}_l^\intercal \mathbf{U}_l)||_2^{\frac{1}{2}} + ||\mathbb{E}(\mathbf{U}_l \mathbf{U}_l^\intercal)||_2^{\frac{1}{2}} \Big), \end{aligned}$$

$$\mathbb{P}\Big(\Big| ||\mathbf{U}_l||_2 - \mathbb{E}||\mathbf{U}_l||_2 \Big| \geq t \Big) \leq 2e^{-t^2/2\sigma_*(\mathbf{U}_l)^2},$$

$$\sigma_*(\mathbf{U}_l) \leq ||\mathbb{E}(\mathbf{U}_l^\intercal \mathbf{U}_l)||_2^{\frac{1}{2}},$$

where $c > 0$ is a universal constant. Taking a union bond over the layers, we get that, with probability $> \frac{1}{2}$, the spectral norm of $\mathbf{U}_l$ is bounded by $(\sqrt{2\ln(4n)} + c + c\sqrt{\ln h}) ||\mathbb{E}(\mathbf{U}_l^\intercal \mathbf{U}_l)||_2^{\frac{1}{2}} + c||\mathbb{E}(\mathbf{U}_l \mathbf{U}_l^\intercal)||_2^{\frac{1}{2}}$, let $c_1 = \sqrt{2\ln(4n)} + c + c\sqrt{\ln h}$ and $c_2 = c$, we have

$$||\mathbf{U}_l||_2 \leq \Big(c_1 ||\mathbf{R}_l'||_2^{\frac{1}{2}} + c_2 ||\mathbf{R}_l''||_2^{\frac{1}{2}} \Big) \sigma_l. \tag{28}$$

Thus, $\frac{\beta_{\widetilde{\mathbf{W}}_l}}{||\widetilde{\mathbf{W}}_l||_2} ||\mathbf{U}_l||_2$ is bounded by $\Big(c_1 ||\mathbf{R}_l'||_2^{\frac{1}{2}} + c_2 ||\mathbf{R}_l''||_2^{\frac{1}{2}} \Big) \sigma$. Then, according to Appendix C.2 Eq.

(22) in [25], Eqs. (24) and (27), we can get

$$
\begin{aligned}
&||f_{\mathbf{W}+\mathbf{U}}(\mathbf{s}+\delta^{\mathrm{fgm}}_{\mathbf{W}+\mathbf{U}}(\mathbf{s})) - f_{\mathbf{W}}(\mathbf{s}+\delta^{\mathrm{fgm}}_{\mathbf{W}}(\mathbf{s}))|| \\
&\leq e^2(B+\epsilon)\prod_{l=1}^{n}||\widetilde{\mathbf{W}}_l||_2\sum_{l=1}^{n}\frac{||\mathbf{U}_l||_2}{||\widetilde{\mathbf{W}}_l||_2} \\
&\quad + 2e^5\frac{\epsilon}{\kappa}\prod_{l=1}^{n}||\widetilde{\mathbf{W}}_l||_2^2\sum_{l=1}^{n}\Big[\frac{||\mathbf{U}_l||_2}{||\widetilde{\mathbf{W}}_l||_2} \\
&\qquad + B(\prod_{j=1}^{l}||\widetilde{\mathbf{W}}_j||_2)\sum_{j=1}^{l}\frac{||\mathbf{U}_j||_2}{||\widetilde{\mathbf{W}}_j||_2}\Big] \\
&\leq 2e^5(B+\epsilon)\sigma\Big(\sum_{l=1}^{n}(c_1||\mathbf{R}'_l||_2^{\frac{1}{2}} + c_2||\mathbf{R}''_l||_2^{\frac{1}{2}})\Big) \\
&\Big\{\prod_{l=1}^{n}||\widetilde{\mathbf{W}}_l||_2^{\frac{n-1}{n}} + \frac{\epsilon}{\kappa}\Big(\prod_{l=1}^{n}||\widetilde{\mathbf{W}}_l||_2^{\frac{2n-1}{n}}\Big)\Big(\frac{1}{B}+\sum_{l=1}^{n}\prod_{j=1}^{l}||\widetilde{\mathbf{W}}_j||_2\Big)\Big\} \\
&\leq \frac{\gamma}{4},
\end{aligned}
$$

hence we choose

$$
\begin{aligned}
\sigma = &\frac{\gamma}{8e^5(B+\epsilon)(\sum_{l=1}^{n}(c_1||\mathbf{R}'_l||_2^{\frac{1}{2}} + c_2||\mathbf{R}''_l||_2^{\frac{1}{2}}))\prod_{l=1}^{n}||\widetilde{\mathbf{W}}_l||_2^{\frac{n-1}{n}}} \\
&\cdot\frac{1}{(1+\frac{\epsilon}{\kappa}\prod_{l=1}^{n}||\widetilde{\mathbf{W}}_l||_2(\frac{1}{B}+\sum_{l=1}^{n}\prod_{j=1}^{l}||\widetilde{\mathbf{W}}_j||_2))}
\end{aligned}
\tag{29}
$$

Then we can get

$$
\begin{aligned}
&\mathrm{KL}(Q_{\mathrm{vec}(\mathbf{W})+\mathbf{u}}||P) = \sum_{l=1}^{n}\Big(\frac{||\mathbf{W}_l||_F^2}{2\sigma_l^2} - \ln\det\mathbf{R}_l\Big) \\
&\leq \mathcal{O}\Bigg((B+\epsilon)^2\Big(\sum_{l=1}^{n}(c_1||\mathbf{R}'_l||_2^{\frac{1}{2}} + c_2||\mathbf{R}''_l||_2^{\frac{1}{2}})\Big)^2\prod_{l=1}^{n}||\widetilde{\mathbf{W}}_l||_2^2 \\
&\frac{(1+\frac{\epsilon}{\kappa}\prod_{l=1}^{n}||\widetilde{\mathbf{W}}_l||_2\sum_{l=1}^{n}\prod_{j=1}^{l}||\widetilde{\mathbf{W}}_j||_2)^2}{\gamma^2}\sum_{l=1}^{n}\frac{||\mathbf{W}_l||_F^2}{||\widetilde{\mathbf{W}}_l||_2^2} \\
&-\sum_{l=1}^{n}\ln\det\mathbf{R}_l\Bigg) \\
&\leq \mathcal{O}\Bigg((B+\epsilon)^2\Big(\sum_{l=1}^{n}(c_1||\mathbf{R}'_l||_2^{\frac{1}{2}} + c_2||\mathbf{R}''_l||_2^{\frac{1}{2}})\Big)^2\prod_{l=1}^{n}||\mathbf{W}_l||_2^2 \\
&\frac{(1+\frac{\epsilon}{\kappa}\prod_{l=1}^{n}||\mathbf{W}_l||_2\sum_{l=1}^{n}\prod_{j=1}^{l}||\mathbf{W}_j||_2)^2}{\gamma^2}\sum_{l=1}^{n}\frac{||\mathbf{W}_l||_F^2}{||\mathbf{W}_l||_2^2} \\
&-\sum_{l=1}^{n}\ln\det\mathbf{R}_l\Bigg).
\end{aligned}
\tag{30}
$$

Thus, we have

$$
\begin{aligned}
&\mathcal{L}_{\mathcal{D}'}(f_{\mathbf{W}}) \leq \mathcal{L}_{\gamma,\mathcal{S}'}(f_{\mathbf{W}}) + \mathcal{O}\Bigg(\Bigg(\frac{-\sum_l\ln\det\mathbf{R}_l + \ln\frac{m}{\delta}}{\gamma^2 m} \\
&+\frac{\Psi^{\mathrm{adv}}\Big(\sum_l(c_1||\mathbf{R}'_l||_2^{\frac{1}{2}} + c_2||\mathbf{R}''_l||_2^{\frac{1}{2}})\Big)^2}{\gamma^2 m}\Bigg)^{\frac{1}{2}}\Bigg),
\end{aligned}
$$

where $\Psi^{\mathrm{adv}} = (B+\epsilon)^2\Phi^{\mathrm{adv}}$. And

$$
\begin{aligned}
\Phi^{\mathrm{adv}} =& \prod_{l=1}^{n}||\mathbf{W}_l||_2^2\Big\{1+\frac{\epsilon}{\kappa}(\prod_{l=1}^{n}||\mathbf{W}_l||_2) \\
&\cdot\sum_{l=1}^{n}\prod_{j=1}^{l}||\mathbf{W}_j||_2\Big\}^2\sum_{l=1}^{n}\frac{||\mathbf{W}_l||_F^2}{||\mathbf{W}_l||_2^2}
\end{aligned}
\tag{31}
$$

for FGM attack.

Proofs for PGM attack are similar (combine Eqs. (28) and (30) and Appendix C.3 in [25]).

C. Sampling Method

We use sharpness-like method [34] to get a set of weight samples $(\mathbf{W}+\eta)$ such that $|\mathcal{L}(f_{\mathbf{W}+\eta}) - \mathcal{L}(f_{\mathbf{W}})| \leq \epsilon'$ (*e.g.*, $\epsilon' = 0.05$ for CIFAR-10/SVHN and $\epsilon' = 0.1$ for CIFAR-100), where $\mathrm{vec}(\eta)$ is a $\mathbf{0}$ mean Gaussian noise. To get the samples from the posteriori distribution steadily and fastly, we train the convergent network with learning rate 0.0001, noise η and 50 epochs, then collect corresponding 50 samples. As the samples are stabilized at (clean/adversarial) training loss and validation loss but with different weights, we can treat them as the samples from same (clean/adversarial) posteriori distribution and estimate the correlation matrix through these samples.

D. Proofs of Lems. 4.1, 4.2

As we assume $r_{\mathbf{s}}r_{\mathbf{s}'} \geq 0$ (above Lem. 4.1), we give the proofs with two cases ($r_{\mathbf{s}} \geq 0$ and $r_{\mathbf{s}} \leq 0$).

Proof for Lem. 4.1.
Let $r_{\mathbf{s}} \geq 0$ and $r_{\mathbf{s}'} \geq 0$, we get

$$
\begin{aligned}
\Lambda'_{l,\max} &= \max\big(||\mathbf{R}'_{l,\mathcal{S}}||_2^{\frac{1}{2}}, ||\mathbf{R}'_{l,\mathcal{S}'}||_2^{\frac{1}{2}}\big) \\
&= \sqrt{h\big(1+(h-1)\max(r_{\mathbf{s}}, r_{\mathbf{s}'})\big)}
\end{aligned}
\tag{32}
$$

and

$$
\begin{aligned}
\Lambda''_{l,\max} &= \max\big(||\mathbf{R}''_{l,\mathcal{S}}||_2^{\frac{1}{2}}, ||\mathbf{R}''_{l,\mathcal{S}'}||_2^{\frac{1}{2}}\big) \\
&= \sqrt{h\big(1+(h-1)\max(r_{\mathbf{s}}, r_{\mathbf{s}'})\big)}.
\end{aligned}
\tag{33}
$$

Thus, decreasing $||\mathbf{R}_{l,\mathcal{S}}||_F^2$ and $||\mathbf{R}_{l,\mathcal{S}'}||_F^2$ leads to a decline in $\Lambda'_{l,\max}$ and $\Lambda''_{l,\max}$.

Let $r_{\mathbf{s}} \leq 0$ and $r_{\mathbf{s}'} \leq 0$, we get

$$
\begin{aligned}
\Lambda'_{l,\max} &= \max\big(||\mathbf{R}'_{l,\mathcal{S}}||_2^{\frac{1}{2}}, ||\mathbf{R}'_{l,\mathcal{S}'}||_2^{\frac{1}{2}}\big) \\
&= \sqrt{h\big(1-\min(r_{\mathbf{s}}, r_{\mathbf{s}'})\big)}
\end{aligned}
\tag{34}
$$

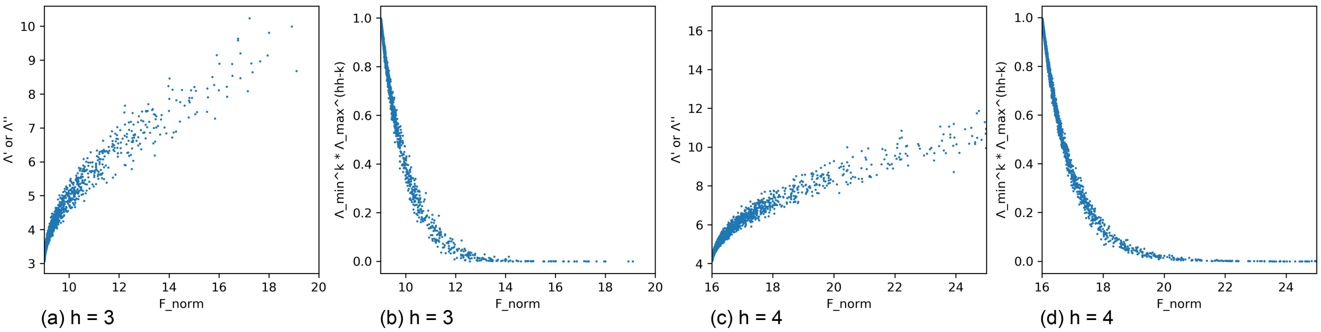

Figure 3. **(a)** We sample 10000 9-dimensional correlation matrices and demonstrate $||\mathbf{R}_l||_F^2$ w.r.t $\Lambda'_{l,\max}$ or $\Lambda''_{l,\max}$. **(b)** We sample 10000 9-dimensional correlation matrices and demonstrate $||\mathbf{R}_l||_F^2$ w.r.t $\Lambda_{l,\min}^{k_l}\Lambda_{l,\max}^{h^2-k_l}$. **(c)** We sample 10000 16-dimensional correlation matrices and demonstrate $||\mathbf{R}_l||_F^2$ w.r.t $\Lambda'_{l,\max}$ or $\Lambda''_{l,\max}$. **(d)** We sample 10000 16-dimensional correlation matrices and demonstrate $||\mathbf{R}_l||_F^2$ w.r.t $\Lambda_{l,\min}^{k_l}\Lambda_{l,\max}^{h^2-k_l}$.

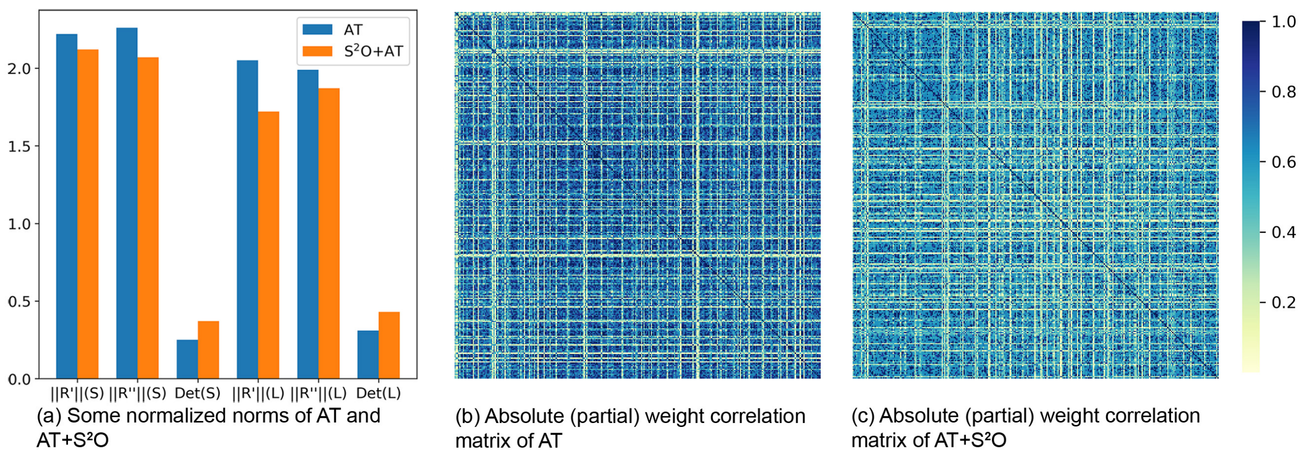

Figure 4. **(a)** shows the normalized spectral norm of $\mathbf{R}'_{\mathcal{S}}$, $\mathbf{R}''{}_{\mathcal{S}}$, and the determinant of $\mathbf{R}_{\mathcal{S}}$, with sampling estimation (S) and Laplace approximation (L) respectively. **(b)** and **(c)** demonstrate the absolute correlation matrix of partial weights (estimate under clean data), for AT and AT+S^{2}O respectively.

and

$$\begin{aligned}\Lambda''_{l,\max} &= \max\left(||\mathbf{R}''_{l,\mathcal{S}}||_2^{\frac{1}{2}}, ||\mathbf{R}''_{l,\mathcal{S}'}||_2^{\frac{1}{2}}\right) \\ &= \sqrt{h(1-\min(r_{\mathbf{s}}, r_{\mathbf{s}'}))}.\end{aligned} \tag{35}$$

Thus, decreasing $||\mathbf{R}_{l,\mathcal{S}}||_F^2$ and $||\mathbf{R}_{l,\mathcal{S}'}||_F^2$ leads to a decline in $\Lambda'_{l,\max}$ and $\Lambda''_{l,\max}$.

Proof for Lem. 4.2.
Let $r_{\mathbf{s}} \geq r_{\mathbf{s}'} \geq 0$, we get

$$\begin{aligned}c(r) &= \Lambda_{l,\min}^{k_l}\Lambda_{l,\max}^{h^2-k_l} \\ &= (1-r_{\mathbf{s}})^{h^2-1}(1+(h^2-1)r_{\mathbf{s}})\end{aligned} \tag{36}$$

and

$$\frac{\partial c(r)}{\partial r_{\mathbf{s}}} = -h^2(h^2-1)r_{\mathbf{s}}(1-r_{\mathbf{s}})^{h^2-2} \leq 0, \tag{37}$$

it is easy to get $c(r)$ is negative correlated with $r_{\mathbf{s}}$. Similarly, if $r_{\mathbf{s}'} \geq r_{\mathbf{s}} \geq 0$, we can get $c(r)$ is negative correlated with $r_{\mathbf{s}'}$. Thus, decreasing $||\mathbf{R}_{l,\mathcal{S}}||_F^2$ and $||\mathbf{R}_{l,\mathcal{S}'}||_F^2$ leads to an increase in $\Lambda_{l,\min}^{k_l}\Lambda_{l,\max}^{h^2-k_l}$.

Let $r_{\mathbf{s}} \leq r_{\mathbf{s}'} \leq 0$, we get

$$\begin{aligned}c(r) &= \Lambda_{l,\min}^{k_l}\Lambda_{l,\max}^{h^2-k_l} \\ &= (1+(h^2-1)r_{\mathbf{s}})(1-r_{\mathbf{s}})^{h^2-1}\end{aligned} \tag{38}$$

and

$$\frac{\partial c(r)}{\partial r_{\mathbf{s}}} = -h^2(h^2-1)r_{\mathbf{s}}(1-r_{\mathbf{s}})^{h^2-2} \geq 0, \tag{39}$$

it is also easy to get $c(r)$ is positive correlated with $r_{\mathbf{s}}$. Similarly, if $r_{\mathbf{s}'} \leq r_{\mathbf{s}} \leq 0$, we can get $c(r)$ is positive correlated with $r_{\mathbf{s}'}$. Thus, decreasing $||\mathbf{R}_{l,\mathcal{S}}||_F^2$ and $||\mathbf{R}_{l,\mathcal{S}'}||_F^2$ leads to an increase in $\Lambda_{l,\min}^{k_l}\Lambda_{l,\max}^{h^2-k_l}$.

E. Simulations of Lems. 4.1, 4.2 and Second-Order Statistics of Weights under Clean Data

As Fig. 3 shows, for 10000 random general 9-dimensional correlation matrices and 16-dimensional correlation matrices respectively, Lems. 4.1 and 4.2 also hold approximately.

The results in Fig. 4 also suggest that $\mathrm{S^2O}$ can decrease the spectral norm of $\mathbf{R}'_{\mathcal{S}}$, $\mathbf{R}''{}_{\mathcal{S}}$ and increases the determinant of $\mathbf{R}_{\mathcal{S}}$.

F. Approximate Optimization

We use a fast approximate method to update $\mathbf{g}(\mathbf{A})$, i.e., add a penalty term to the high correlated $\mathbf{a}_{l,i}$ and $\mathbf{a}_{l,j}$ to reduce their correlation. Details are given in the code.