

Supplementary material 1 for The Double Descent Behavior in Two Layer Neural Network for Binary Classification

S.1 Legendre transformation

Given a function $l : \mathbb{R} \rightarrow \mathbb{R}$, its *Legendre transform* \tilde{l} (the “conjugate” function) is defined by

$$\tilde{l}(y) = \max_{x \in \mathbb{R}} \{xy - l(x)\}. \quad (\text{S1})$$

Then the Legendre transformation transforms the pair $(x, l(x))$ into a new pair $(y, \tilde{l}(y))$ by the definition. The domain of \tilde{l} is the set of $y \in \mathbb{R}$ such that the supremum is finite. y is known as the conjugate variable. If $l(x)$ is a convex function, the inverse transformation gives $l(x)$ back and

$$l(x) = \max_{y \in \mathbb{R}} \{xy - \tilde{l}(y)\}. \quad (\text{S2})$$

In our study, we replace the convex loss function l by the inverse transformation of the Legendre transformation \tilde{l} (S2). For example, the square loss function $l(x) = \frac{1}{2}(1 - x)^2$ has the conjugate given by, $\tilde{l}(y) = \max_{x \in \mathbb{R}} \{xy - \frac{1}{2}(1 - x)^2\} = \frac{y^2}{2} + y$. This transformation allows us to convert the original minimization problem to a min-max problem.

S.2 Properties of Legendre transformation

The necessary condition for the existence of $\tilde{l}(y)$ is that the derivative of the function inside the maximum in (S1) with respect to x is zero, i.e.,

$$y - l'(x) = 0 \Rightarrow l'(x) = y. \quad (\text{S3})$$

This is to be viewed as an equation of x for a given y . Moreover, when the chosen $l(x)$ is a strictly convex function, the second derivative of $xy - l(x)$ with respect to x is $-l''(x)$, which is negative by assumption. Therefore $l'(x) = y$ is necessary and sufficient for the local maximum. It is possible that the equation (S3) has multiple solutions. However, the solution is unique if l satisfies the two conditions that $l'(x)$ is continuous and monotonically increasing and $l'(x) \rightarrow \infty$ for $x \rightarrow \infty$ and $l'(x) \rightarrow -\infty$ for $x \rightarrow -\infty$. Thus, under these conditions, we have an equivalent way to write \tilde{l} via the two equations

$$\tilde{l}(y) = xy - l(x) \quad \text{and} \quad y = l'(x). \quad (\text{S4})$$

This can also be reduced to, $\tilde{l}(y) = xl'(x) - l(x)$, provided that $y = l'(x)$ should be solved for x in terms of y . The differential of $\tilde{l}(y) = xy - l(x)$ can be written as,

$$d\tilde{l}(y) = ydx + xdy - l'(x)dx = ydx + xdy - ydx = xdy \Rightarrow \tilde{l}'(y) = x.$$

In conclusion, when the function l is strictly convex and satisfies the two conditions, for $x = \tilde{l}'(y)$, we have the relationship used in (17) as $l'(x) = y$ and $l(x) = y\tilde{l}'(y) - \tilde{l}(y)$.

S.3 Strategy of using Convex Gaussian Min-Max Theorem

The expression in (21) is a lower bound for the auxiliary problem (10). Since we are interested in high dimensional behaviors when $n, d \rightarrow \infty$, we do not need to compute these lower bounds for the auxiliary problem or local and global losses for the primary problem exactly. Instead, we use the relationships introduced in CGMT to show that, $\omega_\lambda^{(d)}(r, s)$ is a candidate to observe the long-run behavior of the global training loss introduced in (7).

Theorem S.1. *In higher dimension when $n, d \rightarrow \infty$, the global training loss L_λ^* can be approximated by the infimum of the lower bound of the local training loss $\omega_\lambda^{(d)}$ in auxiliary optimization problem, i.e.,*

$$\mathbb{P}\left(\lim_{n, d \rightarrow \infty} L_\lambda^*(r, s) = \lim_{n, d \rightarrow \infty} \omega_\lambda^*(r, s)\right) = 1. \quad (\text{S5})$$

Proof. For fixed r and s , we previously defined the local training loss $L_\lambda(r, s)$ in (6). We set the global training loss in (7). Using AO problem, in the local training loss minimization procedure, we have found a lower bound $\omega_\lambda^{(d)}(r, s)$ as in (21) such that $\tilde{L}_\lambda(r, s) \geq \omega_\lambda^{(d)}(r, s)$. Next we define,

$$\omega_\lambda^*(r, s) := \inf_{s^2 \leq r} \omega_\lambda^{(d)}(r, s). \quad (\text{S6})$$

The first statement of CGMT resulted in the following inequality in (14).

$$\mathbb{P}(L_\lambda(r, s) < c) \leq 2\mathbb{P}(\omega_\lambda^{(d)}(r, s) < c).$$

By (14), for any $\delta > 0$, $\mathbb{P}(L_\lambda(r, s) < \omega_\lambda^{(d)}(r, s) - \delta) \leq 2\mathbb{P}(\omega_\lambda^{(d)}(r, s) < \omega_\lambda^{(d)}(r, s) - \delta)$. The right side of the inequality becomes zero since $\delta > 0$ and it implies,

$$\mathbb{P}(L_\lambda(r, s) \geq \omega_\lambda^{(d)}(r, s) - \delta) = 1.$$

Then using (7) and (S6), we can rewrite the above as,

$$\mathbb{P}(L_\lambda^*(r, s) \geq \omega_\lambda^*(r, s) - \delta) = 1, \quad \text{and} \quad \mathbb{P}(L_\lambda^*(r, s) - \omega_\lambda^*(r, s) \geq -\delta) = 1.$$

Since our interest is on the high dimensional behavior of the loss, next we consider the limits when $n, d \rightarrow \infty$.

$$\mathbb{P}\left(\lim_{n, d \rightarrow \infty} L_\lambda^*(r, s) - \lim_{n, d \rightarrow \infty} \omega_\lambda^*(r, s) \geq -\delta\right) = 1. \quad (\text{S7})$$

Recall the $\psi(\boldsymbol{\beta}, \mathbf{u})$ in (11). When $u_i y_i > 0$, then $\psi(\boldsymbol{\beta}, \mathbf{u})$ is convex with respect to $\boldsymbol{\beta}$ due to its absolute-valued term with positive $u_i y_i$ and $\psi(\boldsymbol{\beta}, \mathbf{u})$ is concave with respect to u_i since $-\tilde{l}(u_i)$ is concave by its definition (Supplementary material 1 S1). Hence $\psi(\boldsymbol{\beta}, \mathbf{u})$ is convex-concave on $\mathbb{R}^d \times \mathbb{R}^n$ where $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\mathbf{u} \in \mathbb{R}^n$. Using the convex-concave property of $\psi(\boldsymbol{\beta}, \mathbf{u})$ [1], for any $c \in \mathbb{R}$, we have

$$\mathbb{P}(L_\lambda(r, s) \geq c) \leq 2\mathbb{P}(\tilde{L}_\lambda(r, s) \geq c).$$

Let $c = \omega_\lambda^{(d)}(r, s) + \delta$ for any $\delta > 0$ and it yields,

$$\mathbb{P}(L_\lambda(r, s) \geq \omega_\lambda^{(d)}(r, s) + \delta) \leq 2\mathbb{P}(\tilde{L}_\lambda(r, s) \geq \omega_\lambda^{(d)}(r, s) + \delta).$$

Then using (7) and (S6), we claim that the infimum of $L_\lambda(r, s)$ is greater than the infimum of $\omega_\lambda^{(d)}(r, s) + \delta$, since $L_\lambda(r, s) \geq \omega_\lambda^{(d)}(r, s) + \delta$. Same argument follows for the right side of

the inequality above and we get, $\mathbb{P}(L_\lambda^*(r, s) \geq \omega_\lambda^*(r, s) + \delta) \leq 2\mathbb{P}(\tilde{L}_\lambda^*(r, s) \geq \omega_\lambda^*(r, s) + \delta)$. We rewrite this as,

$$\mathbb{P}(L_\lambda^*(r, s) - \omega_\lambda^*(r, s) \geq \delta) \leq 2\mathbb{P}(\tilde{L}_\lambda^*(r, s) - \omega_\lambda^*(r, s) \geq \delta).$$

If we consider the high dimensional behavior when $n, d \rightarrow \infty$,

$$\mathbb{P}(\lim_{n, d \rightarrow \infty} L_\lambda^*(r, s) - \lim_{n, d \rightarrow \infty} \omega_\lambda^*(r, s) \geq \delta) \leq 2\mathbb{P}(\lim_{n, d \rightarrow \infty} \tilde{L}_\lambda^*(r, s) - \lim_{n, d \rightarrow \infty} \omega_\lambda^*(r, s) \geq \delta). \quad (\text{S8})$$

By the argument shown in supplementary material Section S.4 for square loss function, we have that,

$$\lim_{n, d \rightarrow \infty} |\inf_{s^2 \leq r} \tilde{L}_\lambda(r, s) - \inf_{s^2 \leq r} \omega_\lambda^{(d)}(r, s)| \rightarrow 0.$$

Hence the right side of the (S8), becomes zero and, $\mathbb{P}(\lim_{n, d \rightarrow \infty} L_\lambda^*(r, s) - \lim_{n, d \rightarrow \infty} \omega_\lambda^*(r, s) \geq \delta) = 0$. This implies, $\mathbb{P}(\lim_{n, d \rightarrow \infty} L_\lambda^*(r, s) - \lim_{n, d \rightarrow \infty} \omega_\lambda^*(r, s) \leq \delta) = 1$. Combining (S7) with the above final result, we get S5. \square

Hence to observe the asymptotic behavior of the global training loss, we use $\omega_\lambda^{(d)}(r, s)$ as a candidate. Since we already have an expression for $\omega_\lambda^{(d)}(r, s)$ in (21), first we minimize it to find $\omega_\lambda^*(r, s)$ and finally consider the high dimensional behavior by sending $n, d \rightarrow \infty$.

$$\text{S.4} \quad \lim_{n, d \rightarrow \infty} |\inf_{s^2 \leq r} \tilde{L}_\lambda(r, s) - \inf_{s^2 \leq r} \omega_\lambda^{(d)}(r, s)| = 0$$

$\omega_\lambda^{(d)}(r, s)$ can be written as $\omega_\lambda^{(d)}(r, s) = \frac{\lambda r}{2} + \frac{1}{d} \sum_{i=1}^n \frac{(1-v_i)^2}{2}$ with the square loss function $l(v_i) = \frac{(1-v_i)^2}{2}$. Then,

$$\frac{\sum_{i=1}^n (v_i - 1)^2}{d} = \frac{1}{d} \left(\sum_{i=1}^n v_i^2 - 2 \sum_{i=1}^n v_i + \sum_{i=1}^n 1 \right). \quad (\text{S9})$$

Note that $\sum_{i=1}^n 1/d = \alpha$ and $\frac{2 \sum_{i=1}^n v_i}{d} = \frac{2 \sum_{i=1}^n y_i \sigma(\frac{\mathbf{x}_i^T \boldsymbol{\beta}}{\sqrt{d}} + b)}{d}$, where

$$\frac{\sum_{i=1}^n \sigma(\frac{\mathbf{x}_i^T \boldsymbol{\beta}}{\sqrt{d}} + b)}{d} \leq \frac{\sum_{i=1}^n |\frac{\mathbf{x}_i^T \boldsymbol{\beta}}{\sqrt{d}} + b|}{d} = \frac{\sum_{i=1}^n |\mathbf{x}_i^T \boldsymbol{\beta} + b\sqrt{d}|}{d^{1.5}} \leq \frac{\sum_{i=1}^n |\mathbf{x}_i^T \boldsymbol{\beta}|}{d^{1.5}} + \alpha|b|$$

For the first term we substitute the teacher model in (2),

$$\begin{aligned} \frac{\sum_{i=1}^n |\mathbf{x}_i^T \boldsymbol{\beta}|}{d^{1.5}} &= \frac{1}{d^{1.5}} \sum_{i=1}^n \left| \left(\frac{\boldsymbol{\eta} y_i}{\sqrt{d}} + \epsilon_i \right)^T \boldsymbol{\beta} \right| = \frac{1}{d^{1.5}} \sum_{i=1}^n \left| \frac{y_i \boldsymbol{\eta}^T \boldsymbol{\beta}}{\sqrt{d}} + \epsilon_i^T \boldsymbol{\beta} \right| \\ &\leq \frac{s}{d} \sum_{i=1}^n |y_i| + \frac{1}{d^{1.5}} \sum_{i=1}^n |\epsilon_i^T \boldsymbol{\beta}| = s\alpha + \frac{1}{d^{1.5}} \sum_{i=1}^n |\epsilon_i^T \boldsymbol{\beta}| \rightarrow s\alpha + \alpha \sqrt{\frac{2r}{\pi}} \end{aligned}$$

by the law of large numbers. Next we work on the third term in (S9).

$$\frac{\sum_{i=1}^n v_i^2}{d} = \frac{\sum_{i=1}^n (y_i \sigma(\frac{\mathbf{x}_i^T \boldsymbol{\beta}}{\sqrt{d}} + b))^2}{d} \leq \frac{\sum_{i=1}^n (\mathbf{x}_i^T \boldsymbol{\beta} + b\sqrt{d})^2}{d^2} \leq 2 \frac{\sum_{i=1}^n (\mathbf{x}_i^T \boldsymbol{\beta})^2}{d^2} + 2\alpha b^2.$$

The first term can be simplified using the teacher model as,

$$\begin{aligned} \frac{\sum_{i=1}^n (\mathbf{x}_i^T \boldsymbol{\beta})^2}{d^2} &= \frac{1}{d^2} \sum_{i=1}^n \left(\left(\frac{\boldsymbol{\eta} y_i}{\sqrt{d}} + \epsilon_i \right)^T \boldsymbol{\beta} \right)^2 \leq \frac{2}{d^2} \sum_{i=1}^n \left(\frac{y_i \boldsymbol{\eta}^T \boldsymbol{\beta}}{\sqrt{d}} \right)^2 + \frac{2}{d^2} \sum_{i=1}^n (\epsilon_i^T \boldsymbol{\beta})^2 \\ &= \frac{2s^2\alpha}{d} + \frac{2}{d^2} \sum_{i=1}^n (\epsilon_i^T \boldsymbol{\beta})^2 \rightarrow 2r\alpha \end{aligned}$$

Thus the third term in (S9) goes to $2r\alpha$ when $d \rightarrow \infty$ for a fixed value of α . Hence $\omega_\lambda^{(d)}(r, s)$ is bounded. This, in particular, implies that

$$|\inf_{s^2 \leq r} \tilde{L}_\lambda(r, s) - \inf_{s^2 \leq r} \omega_\lambda^{(d)}(r, s)| \leq \sup_{s^2 \leq r} |\tilde{L}_\lambda(r, s) - \omega_\lambda^{(d)}(r, s)|. \quad (\text{S10})$$

Indeed, if both $\tilde{L}_\lambda(r, s)$ and $\omega_\lambda^{(d)}(r, s)$ are bounded functions, then the above inequality follows, and in case only $\omega_\lambda^{(d)}(r, s)$ is bounded, then the right side of (S10) is infinite. Taking the limits $n, d \rightarrow \infty$ in (S10) gives

$$\begin{aligned} \lim_{n, d \rightarrow \infty} |\inf_{s^2 \leq r} \tilde{L}_\lambda(r, s) - \inf_{s^2 \leq r} \omega_\lambda^{(d)}(r, s)| &\leq \lim_{n, d \rightarrow \infty} \sup_{s^2 \leq r} |\tilde{L}_\lambda(r, s) - \omega_\lambda^{(d)}(r, s)| \\ &= \lim_{n, d \rightarrow \infty} \sup_{s^2 \leq r} \left| \max_{\substack{\mathbf{u}, \\ u_i y_i > 0}} \left\{ \frac{1}{d} \sum_{i=1}^n \left(\frac{u_i(s + y_i b + \sqrt{r} g_i)}{2} - \tilde{l}(u_i) \right) \right. \right. \\ &\quad + \min_{\boldsymbol{\beta}} \left\{ \frac{\|\mathbf{u}\|_2 \mathbf{h}^T \boldsymbol{\beta}}{2d\sqrt{d}} + \frac{1}{d} \sum_{i=1}^n \frac{u_i y_i}{2\sqrt{d}} |\mathbf{x}_i^T \boldsymbol{\beta} + b\sqrt{d}| \right\} \Big\} - \max_{\substack{\mathbf{u}, \\ u_i y_i > 0}} \left\{ \frac{1}{d} \sum_{i=1}^n \left(\frac{u_i(s + y_i b + \sqrt{r} g_i)}{2} - \tilde{l}(u_i) \right) \right. \\ &\quad \left. \left. + \left(\frac{s \boldsymbol{\eta}^T \mathbf{h}}{d\sqrt{d}} - \sqrt{\frac{r - s^2}{d}} \right) \frac{\|\mathbf{u}\|_2}{2} + \min_{\boldsymbol{\beta}} \frac{1}{d} \sum_{i=1}^n \frac{u_i y_i}{2\sqrt{d}} |\mathbf{x}_i^T \boldsymbol{\beta} + b\sqrt{d}| \right\} \right| \end{aligned}$$

Simplifying further,

$$\begin{aligned} \lim_{n, d \rightarrow \infty} |\inf_{s^2 \leq r} \tilde{L}_\lambda(r, s) - \inf_{s^2 \leq r} \omega_\lambda^{(d)}(r, s)| &\leq \lim_{n, d \rightarrow \infty} \sup_{s^2 \leq r} \max_{\substack{\mathbf{u}, \\ u_i y_i > 0}} \left| \min_{\boldsymbol{\beta}} \left\{ \frac{\|\mathbf{u}\|_2 \mathbf{h}^T \boldsymbol{\beta}}{2d\sqrt{d}} + \frac{1}{d} \sum_{i=1}^n \frac{u_i y_i}{2\sqrt{d}} |\mathbf{x}_i^T \boldsymbol{\beta} + b\sqrt{d}| \right\} \right. \\ &\quad \left. - \left(\frac{s \boldsymbol{\eta}^T \mathbf{h}}{d\sqrt{d}} - \sqrt{\frac{r - s^2}{d}} \right) \frac{\|\mathbf{u}\|_2}{2} - \min_{\boldsymbol{\beta}} \frac{1}{d} \sum_{i=1}^n \frac{u_i y_i}{2\sqrt{d}} |\mathbf{x}_i^T \boldsymbol{\beta} + b\sqrt{d}| \right| \\ &= \lim_{n, d \rightarrow \infty} \sup_{s^2 \leq r} \max_{\substack{\mathbf{u}, \\ u_i y_i > 0}} \left| \min_{\boldsymbol{\beta}} \left\{ \frac{\|\mathbf{u}\|_2 \mathbf{h}^T \boldsymbol{\beta}}{2d\sqrt{d}} + \frac{1}{d} \sum_{i=1}^n \frac{u_i y_i}{2\sqrt{d}} |\mathbf{x}_i^T \boldsymbol{\beta} + b\sqrt{d}| \right\} - \left(\frac{s \boldsymbol{\eta}^T \mathbf{h}}{d\sqrt{d}} - \sqrt{\frac{r - s^2}{d}} \right) \frac{\|\mathbf{u}\|_2}{2} \right|. \end{aligned}$$

We see that when $n, d \rightarrow \infty$, the last expression reaches 0 given that $\|\mathbf{u}\|_2/d$ is bounded. The boundedness of $\|\mathbf{u}\|_2/d$ follows from boundedness of $\omega_\lambda^{(d)}(r, s)$ since

$$\frac{\|\mathbf{u}\|_2^2}{d^2} = \frac{\sum_{i=1}^n (v_i - 1)^2}{d^2} \leq \omega_\lambda^{(d)}(r, s).$$

Thus, $\lim_{n, d \rightarrow \infty} |\inf_{s^2 \leq r} \tilde{L}_\lambda(r, s) - \inf_{s^2 \leq r} \omega_\lambda^{(d)}(r, s)| \rightarrow 0$ under the square loss.

S.5 Finding derivatives of (21) with respect to r, s , and b

Derivative with respect to r

Differentiating (21) with respect to r , and setting it to be zero, we have

$$\frac{d\omega_\lambda^{(d)}(r, s)}{dr} = \frac{\lambda}{2} + \frac{\alpha}{n} \sum_{i=1}^n l'(v_i) \frac{dv_i}{dr} = 0 \implies \sum_{i=1}^n l'(v_i) \frac{dv_i}{dr} = -\frac{\lambda n}{2\alpha}. \quad (\text{S11})$$

Differentiating (20) with respect to r yields, $\gamma l''(v_i) \frac{dv_i}{dr} + l'(v_i) \frac{d\gamma}{dr} + \frac{dv_i}{dr} = \frac{g_i}{4\sqrt{r}}$. Rearranging the terms will yield,

$$l''(v_i) \frac{dv_i}{dr} = \frac{1}{\gamma} \left(\frac{g_i}{4\sqrt{r}} - \frac{dv_i}{dr} - l'(v_i) \frac{d\gamma}{dr} \right). \quad (\text{S12})$$

Differentiating (22) with respect to r ,

$$8\alpha\gamma \|l'(v)\|_2^2 \frac{d\gamma}{dr} + 8\alpha\gamma^2 \sum_{i=1}^n l'(v_i) l''(v_i) \frac{dv_i}{dr} = \left(\sqrt{r-s^2} - \frac{s\boldsymbol{\eta}^T \mathbf{h}}{d} \right) \frac{n}{\sqrt{r-s^2}}.$$

Next we substitute (S11) and (S12) in the above expression. After some algebra we have

$$\frac{\alpha}{\sqrt{r}} \frac{1}{n} \sum_{i=1}^n g_i l'(v_i) = -2\lambda + \frac{1}{2\gamma} - \frac{s\boldsymbol{\eta}^T \mathbf{h}}{2\gamma d \sqrt{r-s^2}}. \quad (\text{S13})$$

For easy computation, define $w_i = (s + y_i b + \sqrt{r} g_i)$ and w_i follows a normal distribution with mean $s + y_i b$ and standard deviation \sqrt{r} conditioned on y_i for each $1 \leq i \leq n$. Now we rewrite (20) as $2v_i + 2\gamma l'(v_i) = w_i$ and obtain $l'(v_i)$ as below for all $v_i \in \mathbb{R}$,

$$l'(v_i) = \frac{w_i - 2v_i}{2\gamma}. \quad (\text{S14})$$

Then we update the relationship in (S13) using the above substitution to get (23).

Derivative with respect to s

First we differentiate (21) with respect to s and make it equal to 0 to have

$$\frac{d\omega_\lambda^{(d)}(r, s)}{ds} = \frac{\alpha}{n} \sum_{i=1}^n l'(v_i) \frac{dv_i}{ds} = 0 \implies \sum_{i=1}^n l'(v_i) \frac{dv_i}{ds} = 0. \quad (\text{S15})$$

Differentiating (20) with respect to s gives us, $l'(v_i) \frac{d\gamma}{ds} + l''(v_i) \gamma \frac{dv_i}{ds} + \frac{dv_i}{ds} = \frac{1}{2}$. Rearranging the terms will yield,

$$l''(v_i) \frac{dv_i}{ds} = \frac{1}{\gamma} \left(\frac{1}{2} - \frac{dv_i}{ds} - l'(v_i) \frac{d\gamma}{ds} \right). \quad (\text{S16})$$

Differentiating (22) with respect to s gives the following expression.

$$8\alpha\gamma \|l'(v)\|_2^2 \frac{d\gamma}{ds} + 8\alpha\gamma^2 \sum_{i=1}^n l'(v_i) l''(v_i) \frac{dv_i}{ds} = 2n \left(\sqrt{r-s^2} - \frac{s\boldsymbol{\eta}^T \mathbf{h}}{d} \right) \left(-\frac{s}{\sqrt{r-s^2}} - \frac{\boldsymbol{\eta}^T \mathbf{h}}{d} \right).$$

We let $G = \sqrt{r-s^2} \frac{\boldsymbol{\eta}^T \mathbf{h}}{d} - \frac{s^2 \boldsymbol{\eta}^T \mathbf{h}}{d\sqrt{r-s^2}} - s \left(\frac{\boldsymbol{\eta}^T \mathbf{h}}{d} \right)^2$ and write the above expression as,

$$8\alpha\gamma \|l'(v)\|_2^2 \frac{d\gamma}{ds} + 8\alpha\gamma^2 \sum_{i=1}^n l'(v_i) l''(v_i) \frac{dv_i}{ds} = -2n(s + G).$$

Simplifying the above expression using (S16) and (S15) yields, $-2\alpha\gamma \frac{1}{n} \sum_{i=1}^n l'(v_i) = s + G$. Replacing $l'(v_i)$ by (S14) results in the expression in (24).

Derivative with respect to b

Finally, we follow the same procedure for the bias term b by differentiating (21) with respect to b and make it equal to 0 to get the following relationship.

$$\frac{d\omega_{\lambda}^{(d)}(r, s)}{db} = \frac{\alpha}{n} \sum_{i=1}^n l'(v_i) \frac{dv_i}{db} = 0 \implies \sum_{i=1}^n l'(v_i) \frac{dv_i}{db} = 0. \quad (\text{S17})$$

Differentiating (20) with respect to b gives, $\frac{d\gamma}{db} l'(v_i) + \gamma l''(v_i) \frac{dv_i}{db} + \frac{dv_i}{db} = \frac{y_i}{2}$. Rearranging the terms will yield,

$$l''(v_i) \frac{dv_i}{db} = \frac{1}{\gamma} \left(\frac{y_i}{2} - \frac{dv_i}{db} - l'(v_i) \frac{d\gamma}{db} \right). \quad (\text{S18})$$

Differentiating (22) with respect to b gives the following expression.

$$8\alpha\gamma \frac{d\gamma}{db} \|l'(v)\|_2^2 + 8\alpha\gamma^2 \sum_{i=1}^n l'(v_i) l''(v_i) \frac{dv_i}{db} = 0.$$

Simplifying the above expression using (S18) and (S17) yields, $\frac{1}{2} \sum_{i=1}^n y_i l'(v_i) = 0$. We end up with the relationship as shown in (25) after replacing $l'(v_i)$ by (S14).

S.6 Application of Theorem 2 for Square Loss in empirical risk minimization procedure

In this section we workout the fixed point equations in Theorem 2 for square loss $l(v_i) = \frac{1}{2}(1 - v_i)^2$ and plot the curves for generalization error given in Theorem 1.

For all $v_i \in \mathbb{R}$ we have $l'(v_i) = v_i - 1$, and by (30) we get

$$\gamma(v_i - 1) + v_i = \frac{1}{2}w_i \implies v_i = \frac{w_i + 2\gamma}{2(\gamma + 1)}. \quad (\text{S19})$$

From this we compute

$$w_i - 2v_i = \frac{\gamma(w_i - 2)}{\gamma + 1}, \quad (\text{S20})$$

which we use later to simplify the equations introduced in Theorem 2.

Recall that $w_i = s + y_i b + \sqrt{r}g_i \sim \mathcal{N}(s + y_i b, r)$ and $y_i = \pm 1$ with probabilities ρ_1 and ρ_{-1} , respectively, and $g_i \sim \mathcal{N}(0, 1)$. Also, applying the law of large numbers, we simplify (31) as follows:

$$\begin{aligned} \lim_{n, d \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n g_i(w_i - 2v_i) \right) &= \lim_{n, d \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n g_i \frac{\gamma(w_i - 2)}{\gamma + 1} \right) \\ &= \frac{\gamma^*}{1 + \gamma^*} \lim_{n, d \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n g_i(w_i - 2) \right) \\ &= \frac{\gamma^*}{1 + \gamma^*} \lim_{n, d \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n g_i(s + y_i b + \sqrt{r}g_i - 2) \right) \\ &= \frac{\gamma^*}{1 + \gamma^*} \lim_{n, d \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n g_i s + g_i y_i b + \sqrt{r}g_i^2 - 2g_i \right) \\ &= \frac{\gamma^*}{1 + \gamma^*} (0 + 0 + \sqrt{r^*} - 0) = \frac{\gamma^* \sqrt{r^*}}{1 + \gamma^*}. \end{aligned}$$

Then from (31) we get

$$\frac{\alpha}{\sqrt{r^*}} \frac{\gamma^* \sqrt{r^*}}{1 + \gamma^*} = -4\lambda\gamma^* + 1.$$

Rearranging the terms we derive the formula for γ^* :

$$\gamma^* = \frac{-(\alpha + 4\lambda - 1) \pm \sqrt{(\alpha + 4\lambda - 1)^2 + 16\lambda}}{8\lambda}. \quad (\text{S21})$$

Next we compute the limits in (33):

$$\begin{aligned} \lim_{n,d \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n y_i (w_i - 2v_i) \right) &= \lim_{n,d \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n y_i \frac{\gamma(w_i - 2)}{\gamma + 1} \right) \\ &= \frac{\gamma^*}{\gamma^* + 1} \lim_{n,d \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n y_i (w_i - 2) \right) \\ &= \frac{\gamma^*}{\gamma^* + 1} (s^* (2\rho_1 - 1) + b - 2(2\rho_1 - 1)) = 0, \end{aligned}$$

yielding

$$b = (2 - s^*)(2\rho_1 - 1). \quad (\text{S22})$$

Similarly, (32) can be rewritten as

$$\begin{aligned} \lim_{n,d \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n (w_i - 2v_i) \right) &= \lim_{n,d \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n \frac{\gamma(w_i - 2)}{\gamma + 1} \right) \\ &= \frac{\gamma^*}{\gamma^* + 1} \lim_{n,d \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n (w_i - 2) \right) \\ &= \frac{\gamma^*}{\gamma^* + 1} (s^* + b(1 \cdot \rho_1 + (-1) \cdot \rho_{-1}) - 2) \\ &= \frac{\gamma^*}{\gamma^* + 1} (s^* + b(2\rho_1 - 1) - 2). \end{aligned}$$

Combining this with (S22) we get

$$\begin{aligned} \frac{-\alpha\gamma^*}{1 + \gamma^*} (s^* + b(2\rho_1 - 1) - 2) &= s^*, \\ \frac{-\alpha\gamma^*}{1 + \gamma^*} (s^* + (2 - s^*)(2\rho_1 - 1)(2\rho_1 - 1) - 2) &= s^*, \\ \frac{-\alpha\gamma^*}{1 + \gamma^*} (s^* + (2 - s^*)(2\rho_1 - 1)^2 - 2) &= s^*, \\ \frac{-\alpha\gamma^*}{1 + \gamma^*} (2 - s^*)((2\rho_1 - 1)^2 - 1) &= s^*, \\ \frac{4\alpha\gamma^*\rho_1\rho_{-1}}{1 + \gamma^*} (2 - s^*) &= s^*, \end{aligned}$$

which gives

$$s^* = \frac{8\alpha\gamma^*\rho_1\rho_{-1}}{1 + \gamma^* + 4\alpha\gamma^*\rho_1\rho_{-1}}. \quad (\text{S23})$$

Hence b in (S22) simplifies to

$$b = (2 - s^*)(2\rho_1 - 1) = \frac{2(1 + \gamma^*)(2\rho_1 - 1)}{1 + \gamma^* + 4\alpha\gamma^*\rho_1\rho_{-1}}. \quad (\text{S24})$$

So far we have found γ^*, s^* and b in terms of the known quantities. Finally we do the limit computation in (34),

$$\begin{aligned} \lim_{n,d \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n (w_i - 2v_i)^2 \right) &= \lim_{n,d \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n \left(\frac{\gamma(w_i - 2)}{\gamma + 1} \right)^2 \right) \\ &= \left(\frac{\gamma^*}{\gamma^* + 1} \right)^2 \lim_{n,d \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n (w_i - 2)^2 \right) \\ &= \left(\frac{\gamma^*}{\gamma^* + 1} \right)^2 \lim_{n,d \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n w_i^2 - 4w_i + 4 \right) \\ &= \left(\frac{\gamma^*}{\gamma^* + 1} \right)^2 \lim_{n,d \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n (s + y_i b + \sqrt{r} g_i)^2 - 4w_i + 4 \right) \\ &= \left(\frac{\gamma^*}{\gamma^* + 1} \right)^2 ((s^*)^2 + b^2 + r^* + 2s^*b(2\rho_1 - 1) - 4s^* - 4b(2\rho_1 - 1) + 4). \end{aligned}$$

Hence (34) simplifies to,

$$\alpha \left(\frac{\gamma^*}{\gamma^* + 1} \right)^2 ((s^*)^2 - b^2 + r^* - 4(s^* - 1)) = r^* - (s^*)^2. \quad (\text{S25})$$

Solving this for r^* yields

$$r^* = \frac{\alpha(\gamma^*)^2((s^* - 2)^2 - b^2) + (\gamma^* + 1)^2(s^*)^2}{(1 + \gamma^*)^2 - \alpha(\gamma^*)^2}. \quad (\text{S26})$$

The obtained asymptotic values of γ, r, s together with the asymptotic value of the bias term b are presented below (note that we always pick the positive value of γ):

$$\begin{aligned} \gamma^* &= \frac{-(\alpha + 4\lambda - 1) \pm \sqrt{(\alpha + 4\lambda - 1)^2 + 16\lambda}}{8\lambda}, \\ s^* &= \frac{8\alpha\gamma^*\rho_1\rho_{-1}}{1 + \gamma^* + 4\alpha\gamma^*\rho_1\rho_{-1}}, \\ b^* &= (2 - s^*)(2\rho_1 - 1), \\ r^* &= \frac{\alpha(\gamma^*)^2((s^* - 2)^2 - b^2) + (\gamma^* + 1)^2(s^*)^2}{(1 + \gamma^*)^2 - \alpha(\gamma^*)^2}. \end{aligned} \quad (\text{S27})$$

Recall that from Theorem 1

$$R^*(\hat{\beta}) = 1 - \rho_1 \Phi \left(\frac{s^* + b^*}{\sqrt{r^*}} \right) - \rho_{-1} \Phi \left(\frac{s^* - b^*}{\sqrt{r^*}} \right), \quad (\text{S28})$$

and the values of r^*, s^* and b^* are given in (S27).

References

- [1] Thrampoulidis, C., Oymak, S. and Hassibi, B. (2015). Regularized linear regression: A precise analysis of the estimation error. *Conference on Learning Theory*, 1683-1709.