Supplementary Material

We define here some notation in addition to that of Section 3 in the main text. We denote by ℓ_i the per-instance loss,

$$L^1(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(\mathbf{w}^\intercal \mathbf{x}_i), \tag{35}$$

$$\ell_i(u) = -y_i \log \sigma(u) - (1-y_i)\log(1-\sigma(u)) - \ell_i^*, \tag{36}$$

where ℓ_i^* are constants chosen such that the minimum of ℓ_i is 0, namely $\ell_i^* = -y_i \log y_i - (1-y_i)\log(1-y_i)$.

Slightly abusing notation, we write $L(\tau) = L^1(\mathbf{w}(\tau)) = L(\mathbf{W}_1(\tau), \ldots, \mathbf{W}_N(\tau))$ for the objective value at time τ.

Finally, for a full-rank matrix $\mathbf{A} \in \mathbb{R}^{d\times m}$ $(m \geq 1)$, we denote by $\mathbf{P}_\mathbf{A} \in \mathbb{R}^{d\times d}$ the matrix of projection onto the span of $\mathbf{A}$,

$$\mathbf{P}_\mathbf{A} = \begin{cases} \mathbf{I}, & m \geq d, \\ \mathbf{A}(\mathbf{A}^\intercal\mathbf{A})^{-1}\mathbf{A}^\intercal, & m < d. \end{cases} \tag{37}$$

A. Properties of the Cross-Entropy Loss

Theorem A.1 (Gradient)**.** *The gradient of the cross-entropy loss (35) takes the form*

$$\nabla L^1(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} (\sigma(\mathbf{w}^\intercal\mathbf{x}_i) - y_i)\cdot \mathbf{x}_i. \tag{38}$$

It always lies in the data span, $\nabla L^1(\mathbf{w}) \in \operatorname{span}(\mathbf{X})$.

Proof. Straightforward calculation. □

Theorem A.2 (Global minima)**.** *The global minimum of the cross-entropy loss (35) is 0 and the set of global minimisers is*

$$\left\{\mathbf{w} \in \mathbb{R}^d : \mathbf{X}^\intercal\mathbf{w} = \mathbf{X}^\intercal\mathbf{w}_*\right\}. \tag{39}$$

Proof. We know that $L^1 \geq 0$ and $L^1(\mathbf{w}_*) = 0$, so 0 is the optimal objective value, and the set of global optima consists of all $\mathbf{w}$ such that $L^1(\mathbf{w}) = 0$. The last condition is equivalent to $\forall_i : \ell_i(\mathbf{w}) = 0$, which in turn is equivalent to $\forall_i : \sigma(\mathbf{w}^\intercal\mathbf{x}_i) = \sigma(\mathbf{w}_*^\intercal\mathbf{x}_i)$. By monotonicity of σ, this is further equivalent to $\forall_i : \mathbf{w}^\intercal\mathbf{x}_i = \mathbf{w}_*^\intercal\mathbf{x}_i$, which is a restatement of (39). □

Theorem A.3 (Restricted strong convexity)**.** *Assume* $\mathbf{X}$ *is full-rank. For any sublevel set* $\mathcal{W} = \left\{\mathbf{w} : L^1(\mathbf{w}) \leq l\right\}$, *there exists* $\mu > 0$ *such that*

$$L^1(\mathbf{v}) \geq L^1(\mathbf{w}) + \nabla L^1(\mathbf{w})^\intercal(\mathbf{v}-\mathbf{w}) + \frac{\mu}{2}\|\mathbf{v}-\mathbf{w}\|^2 \tag{40}$$

for all $\mathbf{w}, \mathbf{v} \in \mathcal{W}$ *such that* $\mathbf{v} - \mathbf{w} \in \operatorname{span}(\mathbf{X})$.

Proof. Consider the 2nd-order Taylor expansion of L^1 around $\mathbf{w}$,

$$L^1(\mathbf{v}) = L^1(\mathbf{w}) + \nabla L^1(\mathbf{w})^\intercal(\mathbf{v}-\mathbf{w}) + \frac{1}{2}(\mathbf{v}-\mathbf{w})^\intercal[\nabla^2 L^1(\bar{\mathbf{w}})](\mathbf{v}-\mathbf{w}), \tag{41}$$

where $\nabla^2 L^1(\bar{\mathbf{w}})$ is the Hessian of L^1 evaluated at $\bar{\mathbf{w}}$, a point lying between $\mathbf{v}$ and $\mathbf{w}$. A straightforward calculation shows that the Hessian takes the form

$$\nabla^2 L^1(\bar{\mathbf{w}}) = \mathbf{X}\mathbf{D}_{\bar{\mathbf{w}}}\mathbf{X}^\intercal, \tag{42}$$

where

$$\mathbf{D}_{\bar{\mathbf{w}}} = \operatorname{diag}[\sigma(\bar{\mathbf{w}}^\intercal\mathbf{x}_1)(1-\sigma(\bar{\mathbf{w}}^\intercal\mathbf{x}_1)), \ldots, \sigma(\bar{\mathbf{w}}^\intercal\mathbf{x}_n)(1-\sigma(\bar{\mathbf{w}}^\intercal\mathbf{x}_n))]. \tag{43}$$

We will now show that there is a constant $\omega > 0$ such that

$$\sigma(\bar{\mathbf{w}}^\intercal\mathbf{x}_i)(1-\sigma(\bar{\mathbf{w}}^\intercal\mathbf{x}_i)) \geq \omega \tag{44}$$

for all $\bar{\mathbf{w}} \in \mathcal{W}$ and $i \in \{1, \ldots, n\}$, so that we can claim $\mathbf{D}_{\bar{\mathbf{w}}} \succeq \omega\mathbf{I}$, or consequently $\nabla^2 L^1(\bar{\mathbf{w}}) \succeq \omega\mathbf{X}\mathbf{X}^\intercal$.

Let $\mathbf{w} \in \mathcal{W}$. The bound on $L^1(\mathbf{w})$ implies a bound on $\ell_i(\mathbf{w}^\intercal\mathbf{x}_i)$ for all i,

$$\ell_i(\mathbf{w}^\intercal\mathbf{x}_i) \leq nL^1(\mathbf{w}) \leq nl. \tag{45}$$

Because ℓ_i is convex and $\ell_i(u) \to \infty$ as $u \to \pm\infty$, we know that $\ell_i^{-1}((-\infty, nl])$ is a bounded interval, and the finite union $\cup_{i=1}^{n}\ell_i^{-1}((-\infty, nl])$ is also a bounded interval, whose size depends only on nl and the data. Hence, there exists $K > 0$ such that $\mathbf{w}^\intercal\mathbf{x}_i \in [-K, K]$ for all $\mathbf{w} \in \mathcal{W}$ and $i \in \{1, \ldots, n\}$. The existence of $\omega > 0$ satisfying (44) follows.

Now, let us apply $\nabla^2 L^1(\mathbf{w}) \succeq \omega\mathbf{X}\mathbf{X}^\intercal$ to lower-bound (41):

$$L^1(\mathbf{v}) \geq L^1(\mathbf{w}) + \nabla L^1(\mathbf{w})^\intercal(\mathbf{v}-\mathbf{w}) + \frac{\omega}{2}(\mathbf{v}-\mathbf{w})^\intercal\mathbf{X}\mathbf{X}^\intercal(\mathbf{v}-\mathbf{w}). \tag{46}$$

Consider two cases. If $n \geq d$, $\mathbf{X}\mathbf{X}^\intercal$ is full-rank and $\mathbf{X}\mathbf{X}^\intercal \succeq \lambda_{\min}\mathbf{I}$ holds, where $\lambda_{\min} > 0$ is the smallest eigenvalue of $\mathbf{X}\mathbf{X}^\intercal$. Combined with (46), this proves the claim for $n \geq d$ and $\mu = \omega\lambda_{\min}$.

If $n < d$, $\mathbf{X}^\intercal\mathbf{X}$ is full rank. We can use the assumption $\mathbf{v} - \mathbf{w} \in \operatorname{span}(\mathbf{X})$ to deduce

$$\begin{aligned}\|\mathbf{v}-\mathbf{w}\|^2 &= \|\mathbf{P}_\mathbf{X}(\mathbf{v}-\mathbf{w})\|^2 \\ &= (\mathbf{v}-\mathbf{w})^\intercal\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal(\mathbf{v}-\mathbf{w}) \\ &\leq \lambda_{\max}(\mathbf{v}-\mathbf{w})^\intercal\mathbf{X}\mathbf{X}^\intercal(\mathbf{v}-\mathbf{w}),\end{aligned} \tag{47}$$

where $\lambda_{\max} > 0$ is the largest eigenvalue of $(\mathbf{X}^\intercal\mathbf{X})^{-1}$. Combined with (46), this proves the claim for $n < d$ and $\mu = \omega/\lambda_{\max}$. □

Corollary A.1 (Restricted Polyak-Lojasiewicz)**.** *Assume* $\mathbf{X}$ *is full-rank. For any sublevel set* $\mathcal{W} = \{\mathbf{w} : L^1(\mathbf{w}) \leq l\}$, *there exists* $c > 0$ *such that*

$$cL^1(\mathbf{w}) \leq \frac{1}{2}\left\|\nabla L^1(\mathbf{w})\right\|^2 \tag{48}$$

for all $\mathbf{w} \in \mathcal{W}$.

Proof. Let $\mathbf{w} \in \mathcal{W}$. (If $\mathcal{W}$ is empty, the claim is trivially true.) Theorem A.3 applied to $\mathcal{W}$ implies that for some $\mu > 0$,

$$L^1(\mathbf{v}) \geq L^1(\mathbf{w}) + \nabla L^1(\mathbf{w})^\intercal(\mathbf{v}-\mathbf{w}) + \frac{\mu}{2}\|\mathbf{v}-\mathbf{w}\|^2 \tag{49}$$

for all $\mathbf{v} \in \mathcal{W} \cap \mathcal{V}$ where $\mathcal{V} = \{\mathbf{v} : \mathbf{v} - \mathbf{w} \in \operatorname{span}(\mathbf{X})\}$. Taking $\min_{\mathbf{v}\in\mathcal{W}\cap\mathcal{V}}$ on both sides, then relaxing part of the constraint on the right-hand side yields

$$\begin{aligned}&\min_{\mathbf{v}\in\mathcal{W}\cap\mathcal{V}} L^1(\mathbf{v}) \\ &\quad\geq \min_{\mathbf{v}\in\mathcal{W}\cap\mathcal{V}} L^1(\mathbf{w}) + \nabla L^1(\mathbf{w})^\intercal(\mathbf{v}-\mathbf{w}) + \frac{\mu}{2}\|\mathbf{v}-\mathbf{w}\|^2 \\ &\quad\geq \min_{\mathbf{v}\in\mathcal{V}} L^1(\mathbf{w}) + \nabla L^1(\mathbf{w})^\intercal(\mathbf{v}-\mathbf{w}) + \frac{\mu}{2}\|\mathbf{v}-\mathbf{w}\|^2.\end{aligned} \tag{50}$$

Now, the minimum on the left-hand side is equal to 0 and is attained at $\mathbf{v} = \mathbf{w} + \mathbf{P}_\mathbf{X}(\mathbf{w}_* - \mathbf{w})$, as can be seen from Theorem A.2. For the right-hand side, we can substitute $\mathbf{v} = \mathbf{w} + \mathbf{X}\mathbf{a}$ for $\mathbf{a} \in \mathbb{R}^n$ and find the unconstrained minimum with respect to $\mathbf{a}$. We get

$$\begin{aligned}0 &\geq L^1(\mathbf{w}) - \frac{1}{2\mu}\nabla L^1(\mathbf{w})^\intercal\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\nabla L^1(\mathbf{w}) \\ &\geq L^1(\mathbf{w}) - \frac{\lambda_{\max}}{2\mu}\left\|\nabla L^1(\mathbf{w})\right\|^2,\end{aligned} \tag{51}$$

where $\lambda_{\max} > 0$ is the largest eigenvalue of $\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$. This yields the result with $c = \mu/\lambda_{\max}$. □

B. Proof of Theorem 1

We will prove a supporting lemma, and then the theorem.

Lemma B.1. *Assume the student is a directly parameterised linear classifier* $(N = 1)$ *initialised at zero,* $\mathbf{w}(0) = \mathbf{0}$. *Then,* $\mathbf{w}(\tau) \in \operatorname{span}(\mathbf{X})$ *for* $\tau \in [0, \infty)$.

Proof. Let $\mathbf{q} \in \mathbb{R}^d$ be any vector orthogonal to the span of $\mathbf{X}$. It suffices to show that $\mathbf{q}^\intercal\mathbf{w}(\tau) = 0$. For that, notice that $\mathbf{q}^\intercal\mathbf{w}(0) = 0$ and

$$\frac{\mathrm{d}}{\mathrm{d}\tau}(\mathbf{q}^\intercal\mathbf{w}(\tau)) = -\mathbf{q}^\intercal\nabla L^1(\mathbf{w}(\tau)) = 0, \tag{52}$$

where the last equality follows from the fact that $\nabla L^1(\mathbf{w}(\tau)) \in \operatorname{span}(\mathbf{X})$ (Theorem A.1). The claim follows. □

Theorem 1. *Assume the student is a directly parameterised linear classifier* $(N = 1)$ *with weight vector initialised at zero,* $\mathbf{w}(0) = \mathbf{0}$. *Then, the student's weight vector fulfills almost surely*

$$\mathbf{w}(t) \to \hat{\mathbf{w}}, \tag{5}$$

for $t \to \infty$, *with*

$$\hat{\mathbf{w}} = \begin{cases} \mathbf{w}_*, & n \geq d, \\ \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{w}_*, & n < d. \end{cases} \tag{6}$$

Proof. Recall the time-derivative of L,

$$L'(\tau) = -\left\|\nabla L^1(\mathbf{w}(\tau))\right\|^2. \tag{53}$$

The data matrix $\mathbf{X}$ is almost surely (wrt. $\mathbf{X} \sim P_\mathbf{x}^n$) full-rank, we can therefore apply Corollary A.1 to $\mathcal{W} = \{\mathbf{w} : L^1(\mathbf{w}) \leq L^1(\mathbf{0})\}$ and $\mathbf{w}(\tau)$ to lower-bound the gradient norm on the right-hand side of (53). We obtain $L'(\tau) \leq -cL(\tau)$ for some $c > 0$ and all $\tau \in [0, \infty)$, or equivalently,

$$(\log L(\tau))' \leq -c. \tag{54}$$

Integrating over $[0, t]$ yields $L(t) \leq L(0) \cdot e^{-ct}$, which proves global convergence in the objective: $L(t) \to 0$ as $t \to \infty$.

Now invoke Theorem A.3 with $\mathcal{W}$ as above, $\mathbf{v} = \mathbf{w}(t)$ and $\mathbf{w} = \hat{\mathbf{w}}$ (we know that both $\mathbf{w}(\tau), \hat{\mathbf{w}} \in \mathcal{W} \cap \operatorname{span}(\mathbf{X})$, partly by Lemma B.1):

$$L(t) \geq \frac{\mu}{2}\|\mathbf{w}(t) - \hat{\mathbf{w}}\|^2. \tag{55}$$

Since $L(t) \to 0$ as $t \to \infty$, the theorem follows. □

C. Proof of Theorem 2

Theorem 2. *Let $\hat{\mathbf{w}}$ be defined as in Theorem 1. Assume the student is a deep linear network, initialized such that for some $\epsilon > 0$,*

$$\|\mathbf{w}(0)\| < \min\left\{\|\hat{\mathbf{w}}\|, \epsilon^N\left(\epsilon^2\|\hat{\mathbf{w}}\|^{-\frac{2}{N}} + \|\hat{\mathbf{w}}\|^{2-\frac{2}{N}}\right)^{-\frac{N}{2}}\right\}, \tag{11}$$

$$L^1(\mathbf{w}(0)) < L^1(\mathbf{0}), \tag{12}$$

$$\mathbf{W}_{j+1}(0)^\intercal \mathbf{W}_{j+1}(0) = \mathbf{W}_j(0)\mathbf{W}_j(0)^\intercal \tag{13}$$

for $j = 1, \dots, N-1$. Then, for $n \geq d$, student's weight vector fulfills almost surely

$$\mathbf{w}(t) \to \hat{\mathbf{w}}, \tag{14}$$

and for $n < d$,

$$\|\mathbf{w}(t) - \hat{\mathbf{w}}\| \leq \epsilon, \tag{15}$$

for all t large enough.

For the proof, we will need a result by (Arora et al., 2018), which characterises the induced flow on $\mathbf{w}(\tau)$ when running gradient descent on the component matrices $\mathbf{W}_i$.

Lemma C.1 ((Arora et al., 2018, Claim 2)). *If the balancedness condition (13) holds, then*

$$\frac{\partial \mathbf{w}(\tau)}{\partial \tau} = -\|\mathbf{w}(\tau)\|^{\frac{2(N-1)}{N}}\left(\nabla L^1(\mathbf{w}(\tau)) + (N-1)\cdot \mathbf{P}_{\mathbf{w}(\tau)}\nabla L^1(\mathbf{w}(\tau))\right). \tag{56}$$

Proof of Theorem 2. Similarly to the case $N = 1$, we start by looking at the time-derivative of L,

$$\begin{aligned} L'(\tau) &= \nabla L^1(\mathbf{w}(\tau))^\intercal \left(\frac{\partial \mathbf{w}(\tau)}{\partial \tau}\right) \\ &= -\|\mathbf{w}(\tau)\|^{\frac{2(N-1)}{N}}\left(\left\|\nabla L^1(\mathbf{w}(\tau))\right\|^2 + (N-1)\cdot\left\|\mathbf{P}_{\mathbf{w}(\tau)}\nabla L^1(\mathbf{w}(\tau))\right\|^2\right) \\ &\leq -\|\mathbf{w}(\tau)\|^{\frac{2(N-1)}{N}}\cdot\left\|\nabla L^1(\mathbf{w}(\tau))\right\|^2. \end{aligned} \tag{57}$$

It is non-positive, so $\mathbf{w}(\tau)$ stays within the $L(0)$-sublevel set throughout optimisation,

$$\mathbf{w}(\tau) \in \mathcal{W} = \left\{\mathbf{w} : L^1(\mathbf{w}) \leq L(0)\right\}. \tag{58}$$

Also, $\mathcal{W}$ is convex and by Assumption (12) it does not contain $\mathbf{0}$. We can therefore take $\delta > 0$ to be the distance between $\mathcal{W}$ and $\mathbf{0}$, and it follows that $\|\mathbf{w}(\tau)\| \geq \delta$ for $\tau \in [0, \infty)$.

Now, noting that $\mathbf{X}$ is almost surely full-rank, apply Corollary A.1 to $\mathcal{W}$ and $\mathbf{w}(\tau)$ to upper-bound the right-hand side of (57),

$$L'(\tau) \leq -c\delta^{\frac{2(N-1)}{N}} L(\tau). \tag{59}$$

Letting $\tilde{c} = c\delta^{\frac{2(N-1)}{N}}$, we get $(\log L(\tau))' \leq -\tilde{c}$ and consequently $L(t) \leq L(0)\cdot e^{-\tilde{c}t}$. This proves convergence in the objective, $L(t) \to 0$ as $t \to \infty$.

To prove convergence in parameters, we decompose the 'error' $\mathbf{w}(\tau) - \hat{\mathbf{w}}$ into orthogonal components and bound each of them separately,

$$\|\mathbf{w}(\tau) - \hat{\mathbf{w}}\|^2 = \|\mathbf{P}_\mathbf{X}(\mathbf{w}(\tau) - \hat{\mathbf{w}})\|^2 + \|\mathbf{P}_\mathbf{Q}(\mathbf{w}(\tau) - \hat{\mathbf{w}})\|^2, \tag{60}$$

where the columns of $\mathbf{Q} \in \mathbb{R}^{d\times(d-n)}$ orthogonally complement those of $\mathbf{X}$. If $n \geq d$, we simply bound the first term and disregard the second one.

To bound the first term, invoke Theorem A.3 with $\mathcal{W}$, $\mathbf{v} = \mathbf{P}_\mathbf{X}\mathbf{w}(\tau)$ and $\mathbf{w} = \mathbf{P}_\mathbf{X}\hat{\mathbf{w}}$. One can check that $L^1(\mathbf{P}_\mathbf{X}\mathbf{u}) = L^1(\mathbf{u})$ for all $\mathbf{u} \in \mathbb{R}^d$, so $\mathbf{P}_\mathbf{X}\mathbf{w}(\tau) \in \mathcal{W}$ and our use of the theorem is legal. We obtain

$$L(\tau) \geq \frac{\mu}{2}\|\mathbf{P}_\mathbf{X}(\mathbf{w}(\tau) - \hat{\mathbf{w}})\|^2. \tag{61}$$

Since $L(\tau) \to 0$, it follows that

$$\|\mathbf{P}_\mathbf{X}(\mathbf{w}(\tau) - \hat{\mathbf{w}})\|^2 \to 0 \tag{62}$$

as $\tau \to \infty$.

For the second term, notice that $\hat{\mathbf{w}} \in \operatorname{span}(\mathbf{X})$, so $\mathbf{P}_\mathbf{Q}\hat{\mathbf{w}}$ vanishes and we are left with $\|\mathbf{P}_\mathbf{Q}\mathbf{w}(\tau)\|^2$. Denote this quantity $q(\tau)$. Its time derivative is

$$\begin{aligned} q'(\tau) &= 2(\mathbf{P}_\mathbf{Q}\mathbf{w}(\tau))^\intercal\left(\frac{\partial \mathbf{w}(\tau)}{\partial \tau}\right) \\ &= -2\|\mathbf{w}(\tau)\|^{\frac{2(N-1)}{N}}\left(\mathbf{w}(\tau)^\intercal\mathbf{P}_\mathbf{Q}\nabla L^1(\mathbf{w}(\tau)) + \frac{(N-1)}{\|\mathbf{w}(\tau)\|^2}\cdot\mathbf{w}(\tau)^\intercal\mathbf{P}_\mathbf{Q}\mathbf{w}(\tau)\cdot\mathbf{w}(\tau)^\intercal\nabla L^1(\mathbf{w}(\tau))\right) \\ &= -2q(\tau)(N-1)\|\mathbf{w}(\tau)\|^{-2/N}\mathbf{w}(\tau)^\intercal\nabla L^1(\mathbf{w}(\tau)), \end{aligned} \tag{63}$$

where we have used the fact that $\nabla L^1(\mathbf{w}(\tau)) \in \operatorname{span}(\mathbf{X})$ (Theorem A.1) and $\mathbf{Q}$ is orthogonal to $\mathbf{X}$. Rearranging, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\frac{\log q(\tau)}{2(N-1)}\right) = -\|\mathbf{w}(\tau)\|^{-2/N}\cdot\mathbf{w}(\tau)^\intercal\nabla L^1(\mathbf{w}(\tau)). \tag{64}$$

It turns out that the right-hand side expression is integrable in yet another way, namely

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\frac{1}{2N}\log\|\mathbf{w}(\tau)\|^2\right) = -\|\mathbf{w}(\tau)\|^{-2/N}\cdot\mathbf{w}(\tau)^\intercal\nabla L^1(\mathbf{w}(\tau)). \tag{65}$$

Equating the two and integrating over $[0, t]$ yields

$$\log \frac{q(t)}{q(0)} = \frac{N-1}{N} \cdot \log \frac{\|\mathbf{w}(t)\|^2}{\|\mathbf{w}(0)\|^2}, \tag{66}$$

which implies

$$\frac{q(t)}{\|\mathbf{w}(t)\|^2} \leq \left(\frac{\|\mathbf{w}(0)\|}{\|\mathbf{w}(t)\|}\right)^{2/N}, \tag{67}$$

because $q(0) \leq \|\mathbf{w}(0)\|^2$.

We now bound the norm of $\mathbf{w}(t)$. Starting from an orthogonal decomposition similar to (60) and applying (62) with (67), we get

$$\begin{aligned} \|\mathbf{w}(t)\|^2 &= \|\mathbf{P_X}\mathbf{w}(t)\|^2 + \|\mathbf{P_Q}\mathbf{w}(t)\|^2 \\ \limsup_{t\to\infty} \|\mathbf{w}(t)\|^2 &\leq \|\hat{\mathbf{w}}\|^2 + \|\mathbf{w}(0)\|^{\frac{2}{N}} \limsup_{t\to\infty} \|\mathbf{w}(t)\|^{2-\frac{2}{N}}. \end{aligned} \tag{68}$$

Denote $\nu := \limsup_{t\to\infty} \|\mathbf{w}(t)\|$. By the same orthogonal decomposition, we also know that $\nu^2 \geq \limsup_{t\to\infty} \|\mathbf{P_X}\mathbf{w}(t)\|^2 = \|\hat{\mathbf{w}}\|^2 > 0$, so we can divide both sides above by ν^2,

$$1 \leq \frac{\|\hat{\mathbf{w}}\|^2}{\nu^2} + \frac{\|\mathbf{w}(0)\|^{2/N}}{\nu^{2/N}} =: f(\nu). \tag{69}$$

On the right-hand side, we now have a decreasing function of ν that goes to zero as $\nu \to \infty$. However, evaluated at our specific ν, it is lower-bounded by 1, implying an implicit upper bound for ν.

How do we find this bound? Suppose we find some constant K such that $f(K) \leq 1$. Then, because f is decreasing, it must be the case that $\nu \leq K$. One such candidate for K is

$$K = \|\hat{\mathbf{w}}\| \cdot \left(1 - \frac{\|\mathbf{w}(0)\|^{2/N}}{\|\hat{\mathbf{w}}\|^{2/N}}\right)^{\frac{-N}{2(N-1)}}. \tag{70}$$

(Here we have used condition (11): $\|\mathbf{w}(0)\| < \|\hat{\mathbf{w}}\|$.) To check that indeed $f(K) \leq 1$, start from the inequality

$$\begin{aligned} (\|\hat{\mathbf{w}}\|/K)^{\frac{2(N-1)}{N}} + \frac{\|\mathbf{w}(0)\|^{2/N}}{\|\hat{\mathbf{w}}\|^{2/N}} &= 1 \\ \leq \left(1 - \frac{\|\mathbf{w}(0)\|^{2/N}}{\|\hat{\mathbf{w}}\|^{2/N}}\right)^{\frac{-1}{N-1}} &= (\|\hat{\mathbf{w}}\|/K)^{-\frac{2}{N}}. \end{aligned} \tag{71}$$

Taking the leftmost and rightmost expression and multiplying by $(\|\hat{\mathbf{w}}\|/K)^{2/N}$ yields

$$f(K) = \frac{\|\hat{\mathbf{w}}\|^2}{K^2} + \frac{\|\mathbf{w}(0)\|^{2/N}}{K^{2/N}} \leq 1. \tag{72}$$

Hence,

$$\limsup_{t\to\infty} \|\mathbf{w}(t)\| \leq \|\hat{\mathbf{w}}\| \cdot \left(1 - \frac{\|\mathbf{w}(0)\|^{2/N}}{\|\hat{\mathbf{w}}\|^{2/N}}\right)^{\frac{-N}{2(N-1)}}. \tag{73}$$

Finally, let us turn back to our original goal of bounding $\|\mathbf{w}(\tau) - \hat{\mathbf{w}}\|^2$. With (60), (62), (67) and (73), we now know that

$$\begin{aligned} &\limsup_{t\to\infty} \|\mathbf{w}(\tau) - \hat{\mathbf{w}}\|^2 \quad (74) \\ &\leq \|\mathbf{w}(0)\|^{\frac{2}{N}} \|\hat{\mathbf{w}}\|^{\frac{2(N-1)}{N}} \left(1 - \frac{\|\mathbf{w}(0)\|^{\frac{2}{N}}}{\|\hat{\mathbf{w}}\|^{\frac{2}{N}}}\right)^{-1} \quad (75) \\ &= \frac{\|\hat{\mathbf{w}}\|^{2+2/N}}{\|\hat{\mathbf{w}}\|^{2/N} - \|\mathbf{w}(0)\|^{2/N}} - \|\hat{\mathbf{w}}\|^2. \quad (76) \end{aligned}$$

Hence, if we initialise close enough to zero, as specified by condition (11), we can ensure that

$$\limsup_{t\to\infty} \|\mathbf{w}(\tau) - \hat{\mathbf{w}}\|^2 < \epsilon^2. \tag{77}$$

This concludes the proof. □

D. Theorem 3 for Approximate Distillation

We extend Theorem 3 to the setting where the student learns the solution $\hat{\mathbf{w}} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{w}_*$ only ϵ-approximately, as is the case for deep linear networks initialised as in Theorem 2. When $n \geq d$, the teacher's weight vector is recovered exactly and the transfer risk is zero, even when the student is deep. The following theorem therefore only covers the case $n < d$.

Theorem D.1 (Risk bound for approximate distillation)**.** *Let $n < d$. For any training set $\mathbf{X} \in \mathbb{R}^{d\times n}$, let $\hat{h}_\mathbf{X}(\mathbf{x}) = \mathbb{1}\{\hat{\mathbf{w}}_\epsilon^\intercal \mathbf{x} \geq 0\}$ be a linear classifier whose weight vector is ϵ-close to the distillation solution $\hat{\mathbf{w}}$, i.e. $\|\hat{\mathbf{w}}_\epsilon - \hat{\mathbf{w}}\| \leq \epsilon$, where ϵ is a positive constant such that $\epsilon \leq \frac{1}{2}\|\hat{\mathbf{w}}\|$. Define $\delta := \sqrt{\frac{2\pi\epsilon}{\|\hat{\mathbf{w}}\|}}$. Then, it holds for any $\beta \in [0, \pi/2 - \delta]$ that*

$$\mathop{\mathbb{E}}_{\mathbf{X}\sim P_\mathbf{x}^{\otimes n}} \left[R\big(\hat{h}_\mathbf{X} \big| P_\mathbf{x}, \mathbf{w}_*\big)\right] \leq p(\beta) + p(\pi/2 - \delta - \beta)^n. \tag{78}$$

The result is very similar to Theorem 3 in the main text, the only difference is the constant δ which compensates for the imprecision in learning $\hat{\mathbf{w}}$ by pushing the bound up (recall that p is decreasing). However, as ϵ goes to zero, so does δ and we recover the original bound.

For the proof, we start with a tool for controlling the angle between $\hat{\mathbf{w}}$ and $\hat{\mathbf{w}}_\epsilon$. Recall that the angle is defined as

$$\alpha(\mathbf{w}, \mathbf{v}) = \cos^{-1}\left(\frac{\mathbf{w}^\intercal\mathbf{v}}{\|\mathbf{w}\| \cdot \|\mathbf{v}\|}\right) \tag{79}$$

for $\mathbf{w}, \mathbf{v} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$.

Lemma D.1. *Let $\mathbf{w}, \mathbf{v} \in \mathbb{R}^d$ be such that $\|\mathbf{w} - \mathbf{v}\| \leq \epsilon$, where $\epsilon \leq \frac{1}{2}\|\mathbf{w}\|$. Then $\alpha(\mathbf{w}, \mathbf{v}) \leq \sqrt{\frac{2\pi\epsilon}{\|\mathbf{w}\|}}$.*

Proof of Lemma D.1. The first step is to lower-bound the inner product $\mathbf{w}^\intercal \mathbf{v}$. To that end, we expand and rearrange $\|\mathbf{w} - \mathbf{v}\|^2 \leq \epsilon^2$ to obtain

$$2\mathbf{w}^\intercal \mathbf{v} \geq \|\mathbf{w}\|^2 + \|\mathbf{v}\|^2 - \epsilon^2. \tag{80}$$

Now use the triangle relation $\|\mathbf{v}\| \geq \|\mathbf{w}\| - \epsilon$ squared to lower-bound the right-hand side of (80) and get

$$2\mathbf{w}^\intercal \mathbf{v} \geq 2\|\mathbf{w}\|^2 - 2\epsilon\|\mathbf{w}\|, \tag{81}$$

which implies

$$\frac{\mathbf{w}^\intercal \mathbf{v}}{\|\mathbf{w}\| \cdot \|\mathbf{v}\|} \geq \frac{\|\mathbf{w}\| - \epsilon}{\|\mathbf{v}\|} \geq \frac{\|\mathbf{w}\| - \epsilon}{\|\mathbf{w}\| + \epsilon} \geq 1 - \frac{2\epsilon}{\|\mathbf{w}\|}. \tag{82}$$

Thus,

$$1 - \frac{2\epsilon}{\|\mathbf{w}\|} \leq \frac{\mathbf{w}^\intercal \mathbf{v}}{\|\mathbf{w}\| \cdot \|\mathbf{v}\|} = \cos(\alpha(\mathbf{w}, \mathbf{v})). \tag{83}$$

The left-hand side is by assumption non-negative, so we have $\alpha(\mathbf{w}, \mathbf{v}) \in [-\pi/2, \pi/2]$. On this domain,

$$\cos x \leq 1 - \frac{x^2}{\pi}, \tag{84}$$

which lets us deduce

$$1 - \frac{2\epsilon}{\|\mathbf{w}\|} \leq 1 - \frac{\alpha(\mathbf{w}, \mathbf{v})^2}{\pi}. \tag{85}$$

Rearranging yields the result. □

Proof of Theorem D.1. We decompose the expected risk as follows:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{X} \sim P_{\mathbf{x}}^n}\left[R(\hat{h}_{\mathbf{X}} | P_{\mathbf{x}}, \mathbf{w}_*)\right] &= \mathop{\mathbb{P}}_{\substack{\mathbf{X} \sim P_{\mathbf{x}}^n \\ \mathbf{x} \sim P_{\mathbf{x}}}}[\mathbf{w}_*^\intercal \mathbf{x} \cdot \hat{\mathbf{w}}_\epsilon^\intercal \mathbf{x} < 0] = \\
&= \int_{\mathbf{x}:\bar{\alpha}(\mathbf{w}_*, \mathbf{x}) \geq \beta} \mathop{\mathbb{P}}_{\mathbf{X} \sim P_{\mathbf{x}}^n}[\mathbf{w}_*^\intercal \mathbf{x} \cdot \hat{\mathbf{w}}_\epsilon^\intercal \mathbf{x} < 0 | \mathbf{x}] \, \mathrm{d}P_{\mathbf{x}} \\
&+ \int_{\mathbf{x}:\bar{\alpha}(\mathbf{w}_*, \mathbf{x}) < \beta, \, \mathbf{w}_*^\intercal \mathbf{x} > 0} \mathop{\mathbb{P}}_{\mathbf{X} \sim P_{\mathbf{x}}^n}[\hat{\mathbf{w}}_\epsilon^\intercal \mathbf{x} < 0 | \mathbf{x}] \, \mathrm{d}P_{\mathbf{x}} \\
&+ \int_{\mathbf{x}:\bar{\alpha}(\mathbf{w}_*, \mathbf{x}) < \beta, \, \mathbf{w}_*^\intercal \mathbf{x} < 0} \mathop{\mathbb{P}}_{\mathbf{X} \sim P_{\mathbf{x}}^n}[\hat{\mathbf{w}}_\epsilon^\intercal \mathbf{x} > 0 | \mathbf{x}] \, \mathrm{d}P_{\mathbf{x}}.
\end{aligned} \tag{86}$$

Let us fix some $\mathbf{x}$ for which $\bar{\alpha}(\mathbf{w}_*, \mathbf{x}) < \beta$ and $\mathbf{w}_*^\intercal \mathbf{x} > 0$; for this $\mathbf{x}$ we have $\alpha(\mathbf{w}_*, \mathbf{x}) = \bar{\alpha}(\mathbf{w}_*, \mathbf{x})$. Consider the situation where $\bar{\alpha}(\mathbf{w}_*, \mathbf{x}_i) < \pi/2 - \beta - \delta$ for some i. Then by the triangle inequality, Lemma D.1 and Lemma 1,

$$\begin{aligned}
\alpha(\hat{\mathbf{w}}_\epsilon, \mathbf{x}) &\leq \alpha(\hat{\mathbf{w}}_\epsilon, \hat{\mathbf{w}}) + \alpha(\mathbf{w}_*, \hat{\mathbf{w}}) + \alpha(\mathbf{w}_*, \mathbf{x}) & (87) \\
&\leq \delta + \bar{\alpha}(\mathbf{w}_*, \mathbf{x}_i) + \bar{\alpha}(\mathbf{w}_*, \mathbf{x}) & (88) \\
&< \pi/2, & (89)
\end{aligned}$$

which implies $\hat{\mathbf{w}}_\epsilon^\intercal \mathbf{x} > 0$, i.e. a correct prediction (same as the teacher's). Conversely, an error can occur only if $\bar{\alpha}(\mathbf{w}_*, \mathbf{x}_i) \geq \pi/2 - \delta - \beta$ for all i. Because $\mathbf{x}_i$ are independent, we have

$$\begin{aligned}
&\mathop{\mathbb{P}}_{\mathbf{X} \sim P_{\mathbf{x}}^n}[\hat{\mathbf{w}}_\epsilon^\intercal \mathbf{x} < 0 | \, \mathbf{x} : \bar{\alpha}(\mathbf{w}_*, \mathbf{x}) < \beta, \, \mathbf{w}_*^\intercal \mathbf{x} > 0] \\
&\qquad \leq \mathop{\mathbb{P}}_{\mathbf{X} \sim P_{\mathbf{x}}^n}[\forall_i : \bar{\alpha}(\mathbf{w}_*, \mathbf{x}_i) \geq \pi/2 - \delta - \beta] \\
&\qquad = p(\pi/2 - \delta - \beta)^n.
\end{aligned} \tag{90}$$

By a symmetric argument, one can show that

$$\begin{aligned}
&\mathop{\mathbb{P}}_{\mathbf{X} \sim P_{\mathbf{x}}^n}[\hat{\mathbf{w}}_\epsilon^\intercal \mathbf{x} > 0 | \, \mathbf{x} : \bar{\alpha}(\mathbf{w}_*, \mathbf{x}) < \beta, \, \mathbf{w}_*^\intercal \mathbf{x} < 0] \\
&\qquad \leq p(\pi/2 - \delta - \beta)^n.
\end{aligned} \tag{91}$$

Combining (86), (90) and (91) yields the result. □