A Proofs

Proof of Lemma 2.1. Recall the property of the activation function $\sigma(z) = \sigma'(z)z$. Let us prove for any $0 \leqslant t \leqslant s \leqslant L$, and any $l \in [k_{s+1}]$

$$\sum_{i\in[k_t],j\in[k_{t+1}]} \frac{\partial O_l^{s+1}}{\partial W_{ij}^t} W_{ij}^t = O_l^{s+1}(x). \tag{A.1}$$

We prove this statement via induction on the non-negative gap $s - t$. Starting with $s - t = 0$, we have

$$\frac{\partial O_l^{t+1}}{\partial W_{il}^t} = \frac{\partial O_l^{t+1}}{\partial N_l^{t+1}} \frac{\partial N_l^{t+1}}{\partial W_{il}^t} = \sigma'(N_l^{t+1}(x))O_i^t(x),$$
$$\frac{\partial O_l^{t+1}}{\partial W_{ij}^t} = 0, \quad \text{for } j \neq l,$$

and, therefore,

$$\sum_{i\in[k_t],j\in[k_{t+1}]} \frac{\partial O_l^{t+1}}{\partial W_{ij}^t} W_{ij}^t = \sum_{i\in[k_t]} \sigma'(N_l^{t+1}(x))O_i^t(x)W_{il}^t = \sigma'(N_l^{t+1}(x))N_l^{t+1}(x) = O_l^{t+1}(x). \tag{A.2}$$

This solves the base case when $s - t = 0$.

Let us assume for general $s - t \leqslant h$ the induction hypothesis ($h \geqslant 0$), and let us prove it for $s - t = h + 1$. Due to chain-rule in the back-propagation updates

$$\frac{\partial O_l^{s+1}}{\partial W_{ij}^t} = \frac{\partial O_l^{s+1}}{\partial N_l^{s+1}} \sum_{k\in[k_s]} \frac{\partial N_l^{s+1}}{\partial O_k^s} \frac{\partial O_k^s}{\partial W_{ij}^t}. \tag{A.3}$$

Using the induction on $\frac{\partial O_k^s}{\partial W_{ij}^t}$ as $(s-1) - t = h$

$$\sum_{i\in[k_t],j\in[k_{t+1}]} \frac{\partial O_k^s}{\partial W_{ij}^t} W_{ij}^t = O_k^s(x), \tag{A.4}$$

and, therefore,

$$\begin{aligned}
&\sum_{i\in[k_t],j\in[k_{t+1}]} \frac{\partial O_l^{s+1}}{\partial W_{ij}^t} W_{ij}^t \\
&= \sum_{i\in[k_t],j\in[k_{t+1}]} \frac{\partial O_l^{s+1}}{\partial N_l^{s+1}} \sum_{k\in[k_s]} \frac{\partial N_l^{s+1}}{\partial O_k^s} \frac{\partial O_k^s}{\partial W_{ij}^t} W_{ij}^t \\
&= \frac{\partial O_l^{s+1}}{\partial N_l^{s+1}} \sum_{k\in[k_s]} \frac{\partial N_l^{s+1}}{\partial O_k^s} \sum_{i\in[k_t],j\in[k_{t+1}]} \frac{\partial O_k^s}{\partial W_{ij}^t} W_{ij}^t \\
&= \sigma'(N_l^{s+1}(x)) \sum_{k\in[k_s]} W_{kl}^s O_k^s(x) = O_l^{s+1}(x).
\end{aligned}$$

This completes the induction argument. In other words, we have proved for any t, s that $t \leqslant s$, and l is any hidden unit in layer s

$$\sum_{i,j\in\dim(W^t)} \frac{\partial O_l^{s+1}}{\partial W_{ij}^t} W_{ij}^t = O_l^{s+1}(x). \tag{A.5}$$

Remark that in the case when there are hard-coded zero weights, the proof still goes through exactly. The reason is, for the base case $s = t$,

$$\sum_{i\in[k_t],j\in[k_{t+1}]} \frac{\partial O_l^{t+1}}{\partial W_{ij}^t} W_{ij}^t = \sum_{i\in[k_t]} \sigma'(N_l^{t+1}(x))O_i^t(x)W_{il}^t \mathbf{1}(W_{il}^t \neq 0) = \sigma'(N_l^{t+1}(x))N_l^{t+1}(x) = O_l^{t+1}(x).$$

and for the induction step,

$$\sum_{i\in[k_t],j\in[k_{t+1}]} \frac{\partial O_l^{s+1}}{\partial W_{ij}^t} W_{ij}^t = \sigma'(N_l^{s+1}(x)) \sum_{k\in[k_s]} W_{kl}^s O_k^s(x) \mathbf{1}(W_{kl}^s \neq 0) = O_l^{s+1}(x).$$

□

Proof of Corollary 2.1. Observe that $\partial \ell(f,Y)/\partial f = -y$ if $yf < 1$, and $\partial \ell(f,Y)/\partial f = 0$ if $yf \geqslant 1$. When the output layer has only one unit, we find

$$\langle \nabla_\theta \widehat{L}(\theta), \theta \rangle = (L+1)\widehat{\mathbb{E}}\left[\frac{\partial \ell(f_\theta(X),Y)}{\partial f_\theta(X)} f_\theta(X)\right] = (L+1)\widehat{\mathbb{E}}\left[-Y f_\theta(X) \mathbf{1}_{Y f_\theta(X)<1}\right] \ .$$

For a stationary point θ, we have $\nabla_\theta \widehat{L}(\theta) = \mathbf{0}$, which implies the LHS of the above equation is 0. Now recall that the second condition that θ separates the data implies implies $-Y f_\theta(X) < 0$ for any point in the data set. In this case, the RHS equals zero if and only if $Y f_\theta(X) \geqslant 1$. □

Proof of Corollary 2.2. The proof follows from applying Lemma 2.1

$$0 = \theta^T \nabla_\theta \widehat{L}(\theta) = (L+1)\widehat{\mathbb{E}}\left[(Y - X^T \prod_{t=0}^{L} W^t) X^T \prod_{t=0}^{L} W^t\right] \ ,$$

which means $\langle w(\theta), \mathbf{X}^T\mathbf{X} w(\theta) - \mathbf{X}^T\mathbf{Y}\rangle = 0$. □

Proof of Theorem 3.2 (spectral norm). The proof follows from a peeling argument from the right hand side. Recall that $O^t \in \mathbb{R}^{1\times k_t}$, $W^L \in \mathbb{R}^{k_L \times 1}$ and $|O^L W^L| \leqslant \|W^L\|_\sigma \|O^L\|_2$ so one has

$$\begin{aligned}
\frac{1}{(L+1)^2}\|\theta\|_{\mathrm{fr}}^2 &= \mathbb{E}\left[|O^L W^L D^{L+1}|^2\right] \\
&\leqslant \mathbb{E}\left[\|W^L\|_\sigma^2 \cdot \|O^L\|_2^2 \cdot |D^{L+1}(X)|^2\right] \\
&= \mathbb{E}\left[|D^{L+1}(X)|^2 \cdot \|W^L\|_\sigma^2 \cdot \|O^{L-1} W^{L-1} D^L\|_2^2\right] \\
&\leqslant \mathbb{E}\left[|D^{L+1}(X)|^2 \cdot \|W^L\|_\sigma^2 \cdot \|O^{L-1} W^{L-1}\|_2^2 \cdot \|D^L\|_\sigma^2\right] \\
&\leqslant \mathbb{E}\left[\|D^L\|_\sigma^2 |D^{L+1}(X)|^2 \cdot \|W^L\|_\sigma^2 \|W^{L-1}\|_\sigma^2 \cdot \|O^{L-1}\|_2^2\right] \\
&\leqslant \mathbb{E}\left[\|D^L\|_\sigma^2 \|D^{L+1}(X)\|_\sigma^2 \|O^{L-1}\|_2^2\right] \cdot \|W^{L-1}\|_\sigma^2 \|W^L\|_\sigma^2 \\
&\ldots \quad \text{repeat the process to bound } \|O^{L-1}\|_2 \\
&\leqslant \mathbb{E}\left(\|X\|^2 \prod_{t=1}^{L+1} \|D^t(X)\|_\sigma^2\right) \prod_{t=0}^{L} \|W^t\|_\sigma^2 = \|\theta\|_\sigma^2.
\end{aligned}$$

□

Proof of Theorem 3.2 (group norm). The proof still follows a peeling argument from the right. We have

$$\begin{aligned}
\frac{1}{(L+1)^2}\|\theta\|_{\mathrm{fr}}^2 &= \mathbb{E}\left[|O^L W^L D^{L+1}|^2\right] \\
&\leqslant \mathbb{E}\left[\|W^L\|_{p,q}^2 \cdot \|O^L\|_{p*}^2 \cdot |D^{L+1}(X)|^2\right] \quad \text{use (A.6)} \\
&= \mathbb{E}\left[|D^{L+1}(X)|^2 \cdot \|W^L\|_{p,q}^2 \cdot \|O^{L-1} W^{L-1} D^L\|_{p*}^2\right] \\
&\leqslant \mathbb{E}\left[|D^{L+1}(X)|^2 \cdot \|W^L\|_{p,q}^2 \cdot \|O^{L-1} W^{L-1}\|_q^2 \cdot \|D^L\|_{q\to p*}^2\right] \\
&\leqslant \mathbb{E}\left[\|D^L\|_{q\to p*}^2 \|D^{L+1}(X)\|_{p,q}^2 \cdot \|W^L\|_{p,q}^2 \|W^{L-1}\|_{p,q}^2 \cdot \|O^{L-1}\|_{p*}^2\right] \quad \text{use (A.8)} \\
&= \mathbb{E}\left[\|D^L\|_{q\to p*}^2 \|D^{L+1}(X)\|_{p,q}^2 \cdot \|O^{L-1}\|_{p*}^2\right] \cdot \|W^{L-1}\|_{p,q}^2 \|W^L\|_{p,q}^2 \\
&\leqslant \ldots \quad \text{repeat the process to bound } \|O^{L-1}\|_{p*} \\
&\leqslant \mathbb{E}\left(\|X\|_{p*}^2 \prod_{t=1}^{L+1} \|D^t(X)\|_{q\to p*}^2\right) \prod_{t=0}^{L} \|W^t\|_{p,q}^2 = \|\theta\|_{p,q}^2
\end{aligned}$$

In the proof of the first inequality we used Holder's inequality

$$\langle w, v\rangle \leqslant \|w\|_p \|v\|_{p^*} \tag{A.6}$$

where $\frac{1}{p} + \frac{1}{p^*} = 1$. Let's prove for $v \in \mathbb{R}^n$, $M \in \mathbb{R}^{n\times m}$, we have

$$\|v^T M\|_q \leqslant \|v\|_{p^*} \|M\|_{p,q}. \tag{A.7}$$

Denote each column of M as $M_{\cdot j}$, for $1 \leqslant j \leqslant m$,

$$\|v^T M\|_q = \left(\sum_{j=1}^m |v^T M_{\cdot j}|^q\right)^{1/q} \leqslant \left(\sum_{j=1}^m \|v\|_{p^*}^q \|M_{\cdot j}\|_p^q\right)^{1/q} = \|v\|_{p^*}\|M\|_{p,q}. \tag{A.8}$$

□

Proof of Theorem 3.2 (path norm). The proof is due to Holder's inequality. For any $x \in \mathbb{R}^p$

$$\begin{aligned}
&\left|\sum_{i_0,i_1,\ldots,i_L} x_{i_0} W^0_{i_0 i_1} D^1_{i_1}(x) W^1_{i_1 i_2} \cdots D^L_{i_L}(x) W^L_{i_L} D^{L+1}(x)\right| \\
&\leqslant \left(\sum_{i_0,i_1,\ldots,i_L} |x_{i_0} D^1_{i_1}(x)\cdots D^L_{i_L}(x) D^{L+1}(x)|^{q^*}\right)^{1/q^*} \cdot \left(\sum_{i_0,i_1,\ldots,i_L} |W^0_{i_0 i_1} W^1_{i_1 i_2} W^2_{i_2 i_3}\cdots W^L_{i_L}|^q\right)^{1/q}.
\end{aligned}$$

Therefore we have

$$\begin{aligned}
\frac{1}{(L+1)^2}\|\theta\|_{\mathrm{fr}}^2 &= \mathbb{E}\left|\sum_{i_0,i_1,\ldots,i_L} X_{i_0} W^0_{i_0 i_1} D^1_{i_1}(X) W^1_{i_1 i_2}\cdots W^L_{i_L} D^{L+1}_{i_L}(X)\right|^2 \\
&\leqslant \left(\sum_{i_0,i_1,\ldots,i_L} |W^0_{i_0 i_1} W^1_{i_1 i_2} W^2_{i_2 i_3}\cdots W^L_{i_L}|^q\right)^{2/q} \cdot \mathbb{E}\left(\sum_{i_0,i_1,\ldots,i_L} |X_{i_0} D^1_{i_1}(X)\cdots D^L_{i_L}(X) D^{L+1}(X)|^{q^*}\right)^{2/q^*},
\end{aligned}$$

which gives

$$\frac{1}{L+1}\|\theta\|_{\mathrm{fr}} \leqslant \left[\mathbb{E}\left(\sum_{i_0,i_1,\ldots,i_L} |X_{i_0}\prod_{t=1}^{L+1} D^t_{i_t}(X)|^{q^*}\right)^{2/q^*}\right]^{1/2} \cdot \left(\sum_{i_0,i_1,\ldots,i_L}\prod_{t=0}^{L} |W^t_{i_t i_{t+1}}|^q\right)^{1/q} = |||\pi(\theta)|||_q.$$

□

Proof of Theorem 3.2 (matrix-induced norm). The proof follows from the recursive use of the inequality,

$$\|M\|_{p\to q}\|v\|_p \geqslant \|v^T M\|_q.$$

We have

$$\begin{aligned}
\|\theta\|_{\mathrm{fr}}^2 &= \mathbb{E}\left[|O^L W^L D^{L+1}|^2\right] \\
&\leqslant \mathbb{E}\left[\|W^L\|_{p\to q}^2 \cdot \|O^L\|_p^2 \cdot |D^{L+1}(X)|^2\right] \\
&\leqslant \mathbb{E}\left[|D^{L+1}(X)|^2 \cdot \|W^L\|_{p\to q}^2 \cdot \|O^{L-1} W^{L-1} D^L\|_p^2\right] \\
&\leqslant \mathbb{E}\left[|D^{L+1}(X)|^2 \cdot \|W^L\|_{p\to q}^2 \cdot \|O^{L-1} W^{L-1}\|_q^2 \cdot \|D^L\|_{q\to p}^2\right] \\
&\leqslant \mathbb{E}\left[\|D^L\|_{q\to p}^2 \|D^{L+1}(X)\|_{q\to p}^2 \cdot \|W^L\|_{p\to q}^2 \|W^{L-1}\|_{p\to q}^2 \cdot \|O^{L-1}\|_p^2\right] \\
&\leqslant \ldots \quad \text{repeat the process to bound } \|O^{L-1}\|_p \\
&\leqslant \mathbb{E}\left(\|X\|_p^2 \prod_{t=1}^{L+1} \|D^t(X)\|_{q\to p}^2\right)\prod_{t=0}^{L} \|W^t\|_{p\to q}^2 = |||\theta|||_{p\to q}^2,
\end{aligned}$$

where third to last line is because $D^{L+1}(X) \in \mathbb{R}^1$, $|D^{L+1}(X)| = \|D^{L+1}(X)\|_{q\to p}$. □

Proof of Theorem 3.2 (chain of induced norm). The proof follows from a different strategy of peeling the terms from the right hand side, as follows,

$$\begin{aligned}
\|\theta\|_{\mathrm{fr}}^2 &= \mathbb{E}\left[|O^L W^L D^{L+1}|^2\right] \\
&\leqslant \mathbb{E}\left[\|W^L\|_{p_L\to p_{L+1}}^2 \cdot \|O^L\|_{p_L}^2 \cdot |D^{L+1}(X)|^2\right] \\
&\leqslant \mathbb{E}\left[|D^{L+1}(X)|^2 \cdot \|W^L\|_{p_L\to p_{L+1}}^2 \cdot \|O^{L-1}W^{L-1}D^L\|_{p_L}^2\right] \\
&\leqslant \mathbb{E}\left[|D^{L+1}(X)|^2 \cdot \|W^L\|_{p_L\to p_{L+1}}^2 \cdot \|O^{L-1}W^{L-1}\|_{p_L}\|D^L\|_{p_L\to p_L}^2\right] \\
&\leqslant \mathbb{E}\left[\|D^L\|_{p_L\to p_L}^2|D^{L+1}(X)|^2 \cdot \|W^L\|_{p_L\to p_{L+1}}^2\|W^{L-1}\|_{p_{L-1}\to p_L}^2 \cdot \|O^{L-1}\|_{p_{L-1}}^2\right] \\
&\leqslant \mathbb{E}\left(\|X\|_{p_0}^2 \prod_{t=1}^{L+1}\|D^t(X)\|_{p_t\to p_t}^2\right)\prod_{t=0}^{L}\|W^t\|_{p_t\to p_{t+1}}^2 = |||\theta|||_P^2.
\end{aligned}$$

□

Proof of Lemma 4.1.

$$\begin{aligned}
\frac{d}{dr}\|r\theta\|_{\mathrm{fr}}^2 &= \mathbb{E}\left[2\langle\theta, \nabla_\theta f_{r\theta}(X)\rangle f_{r\theta}(X)\right] \\
&= \mathbb{E}\left[\frac{2(L+1)}{r} f_{r\theta}(X) f_{r\theta}(X)\right] \quad \text{use Lemma 2.1} \\
&= \frac{2(L+1)}{r}\|r\theta\|_{\mathrm{fr}}^2
\end{aligned}$$

The last claim can be proved through solving the simple ODE. □

Proof of Lemma 4.2. Let us first construct $\theta' \in \Theta_{L+1}$ that realizes $\lambda f_{\theta_1} + (1-\lambda) f_{\theta_2}$. The idea is very simple: we put θ_1 and θ_2 networks side-by-side, then construct an additional output layer with weights λ, $1-\lambda$ on the output of f_{θ_1} and f_{θ_2}, and the final output layer is passed through $\sigma(x) = x$. One can easily see that our key Lemma 2.1 still holds for this network: the interaction weights between f_{θ_1} and f_{θ_2} are always hard-coded as 0. Therefore we have constructed a $\theta' \in \Theta_{L+1}$ that realizes $\lambda f_{\theta_1} + (1-\lambda) f_{\theta_2}$.

Now recall that

$$\begin{aligned}
\frac{1}{L+2}\|\theta'\|_{\mathrm{fr}} &= \left(\mathbb{E}\, f_{\theta'}^2\right)^{1/2} \\
&= \left(\mathbb{E}(\lambda f_{\theta_1} + (1-\lambda) f_{\theta_2})^2\right)^{1/2} \\
&\leqslant \lambda\left(\mathbb{E}\, f_{\theta_1}^2\right)^{1/2} + (1-\lambda)\left(\mathbb{E}\, f_{\theta_2}^2\right)^{1/2} \leqslant 1
\end{aligned}$$

because $\mathbb{E}[f_{\theta_1} f_{\theta_2}] \leqslant \left(\mathbb{E}\, f_{\theta_1}^2\right)^{1/2}\left(\mathbb{E}\, f_{\theta_2}^2\right)^{1/2}$. □

Proof of Theorem 4.1. Due to Eqn. (3.2), one has

$$\begin{aligned}
\frac{1}{(L+1)^2}\|\theta\|_{\mathrm{fr}}^2 &= \mathbb{E}\left[v(\theta, X)^T X X^T v(\theta, X)\right] \\
&= v(\theta)^T \mathbb{E}\left[X X^T\right] v(\theta)
\end{aligned}$$

because in the linear case $v(\theta, X) = W^0 D^1(x) W^1 D^2(x) \cdots D^L(x) W^L D^{L+1}(x) = \prod_{t=0}^{L} W^t =: v(\theta) \in \mathbb{R}^p$. There-

fore

$$\begin{aligned}
\mathcal{R}_N\left(B_{\mathrm{fr}}(\gamma)\right) &= \mathop{\mathbb{E}}_{\epsilon} \sup_{\theta\in B_{\mathrm{fr}}(\gamma)} \frac{1}{N}\sum_{i=1}^{N}\epsilon_i f_\theta(X_i) \\
&= \mathop{\mathbb{E}}_{\epsilon} \sup_{\theta\in B_{\mathrm{fr}}(\gamma)} \frac{1}{N}\sum_{i=1}^{N}\epsilon_i X_i^T v(\theta) \\
&= \mathop{\mathbb{E}}_{\epsilon} \sup_{\theta\in B_{\mathrm{fr}}(\gamma)} \frac{1}{N}\left\langle \sum_{i=1}^{N}\epsilon_i X_i, v(\theta)\right\rangle \\
&\leqslant \gamma \mathop{\mathbb{E}}_{\epsilon} \frac{1}{N}\left\| \sum_{i=1}^{N}\epsilon_i X_i\right\|_{[\mathbb{E}(XX^T)]^{-1}} \\
&\leqslant \gamma \frac{1}{\sqrt{N}} \sqrt{\frac{1}{N}\mathop{\mathbb{E}}_{\epsilon}\left\| \sum_{i=1}^{N}\epsilon_i X_i\right\|^2_{[\mathbb{E}(XX^T)]^{-1}}} \\
&= \gamma \frac{1}{\sqrt{N}} \sqrt{\left\langle \frac{1}{N}\sum_{i=1}^{N} X_i X_i^T, [\mathbb{E}(XX^T)]^{-1}\right\rangle}.
\end{aligned}$$

Therefore

$$\mathbb{E}\,\mathcal{R}_N\left(B_{\mathrm{fr}}(\gamma)\right) \leqslant \gamma \frac{1}{\sqrt{N}} \sqrt{\mathbb{E}\left\langle \frac{1}{N}\sum_{i=1}^{N} X_i X_i^T, [\mathbb{E}(XX^T)]^{-1}\right\rangle} = \gamma\sqrt{\frac{p}{N}}.$$

□

Proof of Proposition 4.1. If $\mathcal{G}\subseteq\mathcal{F}$ then one has the lower bound $\mathcal{R}_N(\mathcal{G}) \leqslant \mathcal{R}_N(\mathcal{F})$ on the empirical Rademacher complexity of $\mathcal{F}$. One can also obtain an upper bound by examining how the sub-space of functions $\mathcal{G}$ approximates $\mathcal{F}$. For each $f\in\mathcal{F}$ consider the closest point $g_f\in\mathcal{G}$ to f,

$$g_f := \mathop{\arg\min}_{g\in\mathcal{F}_\sigma} \|f-g\|_\infty \ .$$

Then the empirical Rademacher complexity $\mathcal{R}_N(\mathcal{F})$ is upper-bounded in terms of $\mathcal{R}_N(\mathcal{G})$ by

$$\begin{aligned}
\mathcal{R}_N(\mathcal{F}) &= \mathop{\mathbb{E}}_{\epsilon} \sup_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N}\epsilon_i f(X_i) \ , \\
&\leqslant \mathop{\mathbb{E}}_{\epsilon} \sup_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N}\epsilon_i [f(X_i)-g_f(X_i)] + \mathcal{R}_N(\mathcal{G}) \ , \\
&\leqslant \sup_{f\in\mathcal{F}}\inf_{g\in\mathcal{G}} \|f-g\|_\infty + \mathcal{R}_N(\mathcal{G}) \ .
\end{aligned}$$

Therefore, taking expectation values over the data gives,

$$\mathbb{E}\,\mathcal{R}_N(\mathcal{F}) \leqslant \sup_{f\in\mathcal{F}}\inf_{g\in\mathcal{G}} \|f-g\|_\infty + \mathbb{E}\,\mathcal{R}_N(\mathcal{G}) \ .$$

Setting $\mathcal{F}=\mathcal{F}_{\mathrm{fr}}(1)$ without loss of generality, we obtain the required result by appropriate choice of $\mathcal{G}\subseteq\mathcal{F}_{\mathrm{fr}}(1)$. Setting $\mathcal{G}=\mathcal{F}_\sigma(r)$ with $r=1/[\widehat{\mathbb{E}}\|X\|^2]^{1/2}$ gives (Remark 4.1, Theorem 1.1 in [3]),

$$\mathbb{E}\,\mathcal{R}_N(\mathcal{F}_{\mathrm{fr}}(1)) \leqslant \sup_{f\in\mathcal{F}_{\mathrm{fr}}(1)}\inf_{g\in\mathcal{F}_\sigma(r)} \|f-g\|_\infty + \frac{\mathrm{Polylog}}{\sqrt{N}} \ .$$

Setting $\mathcal{G}=\mathcal{F}_{p,q}(r)$ with $r=1/(k^{[1/p^*-1/q]_+})^L \max_i \|X_i\|_{p^*}$ gives (Remark 4.2, Theorem 1 in [15])

$$\mathbb{E}\,\mathcal{R}_N(\mathcal{F}_{\mathrm{fr}}(1)) \leqslant \sup_{f\in\mathcal{F}_{\mathrm{fr}}(1)}\inf_{g\in\mathcal{F}_{p,q}(r)} \|f-g\|_\infty + \frac{2^L\mathrm{Polylog}}{\sqrt{N}} \ .$$

Setting $\mathcal{G} = \mathcal{F}_{\pi,1}(r)$ with $r = 1/\max_i \|X_i\|_\infty$ gives (Remark 4.3, Corollary in [15])

$$\mathbb{E}\,\mathcal{R}_N(\mathcal{F}_{\rm fr}(1)) \leqslant \sup_{f\in\mathcal{F}_{\rm fr}(1)} \inf_{g\in\mathcal{F}_{\pi,1}(r)} \|f-g\|_\infty + \frac{2^L \mathrm{Polylog}}{\sqrt{N}} \quad .$$

In all cases data-dependent pre-factors exactly cancel out and moreover the first term is in function space, not in parameter space.

□

A.1 Invariance of natural gradient

Consider the continuous-time analog of natural gradient flow,

$$d\theta_t = -\mathbf{I}(\theta_t)^{-1}\nabla_\theta L(\theta_t)dt, \tag{A.9}$$

where $\theta \in \mathbb{R}^p$. Consider a differentiable transformation from one parametrization to another $\theta \mapsto \xi \in \mathbb{R}^q$ denoted by $\xi(\theta) : \mathbb{R}^p \to \mathbb{R}^q$. Denote the Jacobian $\mathbf{J}_\xi(\theta) = \frac{\partial(\xi_1,\xi_2,\ldots,\xi_q)}{\partial(\theta_1,\theta_2,\ldots,\theta_p)} \in \mathbb{R}^{q\times p}$. Define the loss function $\tilde{L} : \xi \to \mathbb{R}$ that satisfies

$$L(\theta) = \tilde{L}(\xi(\theta)) = \tilde{L}\circ\xi(\theta),$$

and denote $\tilde{\mathbf{I}}(\xi)$ as the Fisher Information on ξ associated with $\tilde{L}$. Consider also the natural gradient flow on the ξ parametrization,

$$d\xi_t = -\tilde{\mathbf{I}}(\xi_t)^{-1}\nabla_\xi \tilde{L}(\xi_t)dt. \tag{A.10}$$

Intuitively, one can show that the natural gradient flow is "invariant" to the specific parametrization of the problem.

Lemma A.1 (Parametrization invariance)**.** *Denote $\theta \in \mathbb{R}^p$, and the differentiable transformation from one parametrization to another $\theta \mapsto \xi \in \mathbb{R}^q$ as $\xi(\theta) : \mathbb{R}^p \to \mathbb{R}^q$. Assume $\mathbf{I}(\theta)$, $\tilde{\mathbf{I}}(\xi)$ are invertible, and consider two natural gradient flows $\{\theta_t, t>0\}$ and $\{\xi_t, t>0\}$ defined in Eqn.* (A.9) *and* (A.10) *on θ and ξ respectively.*

(1) Re-parametrization: if $q = p$, and assume $\mathbf{J}_\xi(\theta)$ is invertible, then natural gradient flow on the two parameterizations satisfies,

$$\xi(\theta_t) = \xi_t, \quad \forall t,$$

if the initial locations θ_0, ξ_0 are equivalent in the sense $\xi(\theta_0) = \xi_0$.

(2) Over-parametrization: If $q > p$ and $\xi_t = \xi(\theta_t)$ at some fixed time t, then the infinitesimal change satisfies

$$\xi(\theta_{t+dt}) - \xi(\theta_t) = M_t(\xi_{t+dt} - \xi_t), \quad M_t \text{ has eigenvalues either } 0 \text{ or } 1$$

where $M_t = \mathbf{I}(\xi_t)^{-1/2}(I_q - U_\perp U_\perp^T)\mathbf{I}(\xi_t)^{1/2}$, and $U_\perp$ denotes the null space of $\mathbf{I}(\xi)^{1/2}\mathbf{J}_\xi(\theta)$.

Proof of Lemma A.1. From basic calculus, one has

$$\begin{aligned}\nabla_\theta L(\theta) &= \mathbf{J}_\xi(\theta)^T\nabla_\xi\tilde{L}(\xi)\\ \mathbf{I}(\theta) &= \mathbf{J}_\xi(\theta)^T\tilde{\mathbf{I}}(\xi)\mathbf{J}_\xi(\theta)\end{aligned}$$

Therefore, plugging in the above expression into the natural gradient flow in θ

$$\begin{aligned}d\theta_t &= -\mathbf{I}(\theta_t)^{-1}\nabla_\theta L(\theta_t)dt\\ &= -[\mathbf{J}_\xi(\theta_t)^T\tilde{\mathbf{I}}(\xi(\theta_t))\mathbf{J}_\xi(\theta_t)]^{-1}\mathbf{J}_\xi(\theta_t)^T\nabla_\xi\tilde{L}(\xi(\theta_t))dt.\end{aligned}$$

In the re-parametrization case, $\mathbf{J}_\xi(\theta)$ is invertible, and assuming $\xi_t = \xi(\theta_t)$,

$$\begin{aligned}d\theta_t &= -[\mathbf{J}_\xi(\theta_t)^T\tilde{\mathbf{I}}(\xi(\theta_t))\mathbf{J}_\xi(\theta_t)]^{-1}\mathbf{J}_\xi(\theta_t)^T\nabla_\xi\tilde{L}(\xi(\theta_t))dt\\ &= -\mathbf{J}_\xi(\theta_t)^{-1}\tilde{\mathbf{I}}(\xi(\theta_t))^{-1}\nabla_\xi\tilde{L}(\xi(\theta_t))dt\\ \mathbf{J}_\xi(\theta_t)d\theta_t &= -\tilde{\mathbf{I}}(\xi(\theta_t))^{-1}\nabla_\xi\tilde{L}(\xi(\theta_t))dt\\ d\xi(\theta_t) &= -\tilde{\mathbf{I}}(\xi(\theta_t))^{-1}\nabla_\xi\tilde{L}(\xi(\theta_t))dt = -\tilde{\mathbf{I}}(\xi_t)^{-1}\nabla_\xi\tilde{L}(\xi_t)dt.\end{aligned}$$

What we have shown is that under $\xi_t = \xi(\theta_t)$, $\xi(\theta_{t+dt}) = \xi_{t+dt}$. Therefore, if $\xi_0 = \xi(\theta_0)$, we have that $\xi_t = \xi(\theta_t)$.

In the over-parametrization case, $\mathbf{J}_\xi(\theta) \in \mathbb{R}^{q\times p}$ is a non-square matrix. For simplicity of derivation, abbreviate $B := \mathbf{J}_\xi(\theta) \in \mathbb{R}^{q\times p}$. We have

$$\begin{aligned} d\theta_t = \theta_{t+dt} - \theta_t &= -\mathbf{I}(\theta_t)^{-1}\nabla_\theta L(\theta_t)dt \\ &= -[B^T\tilde{\mathbf{I}}(\xi)B]^{-1}B^T\nabla_\xi\tilde{L}(\xi(\theta_t))dt \\ B(\theta_{t+dt} - \theta_t) &= -B\left[B^T\tilde{\mathbf{I}}(\xi)B\right]^{-1}B^T\tilde{L}(\xi(\theta_t))dt. \end{aligned}$$

Via the Sherman-Morrison-Woodbury formula

$$\left[I_q + \frac{1}{\epsilon}\tilde{\mathbf{I}}(\xi)^{1/2}BB^T\tilde{\mathbf{I}}(\xi)^{1/2}\right]^{-1} = I_q - \tilde{\mathbf{I}}(\xi)^{1/2}B(\epsilon I_p + B^T\tilde{\mathbf{I}}(\xi)B)^{-1}B^T\tilde{\mathbf{I}}(\xi)^{1/2}$$

Denoting $\tilde{\mathbf{I}}(\xi)^{1/2}BB^T\tilde{\mathbf{I}}(\xi)^{1/2} = U\Lambda U^T$, we have that $\mathrm{rank}(\Lambda) \leqslant p < q$. Therefore, the LHS as

$$\begin{aligned} \left[I_q + \frac{1}{\epsilon}\tilde{\mathbf{I}}(\xi)^{1/2}BB^T\tilde{\mathbf{I}}(\xi)^{1/2}\right]^{-1} &= U\left[I_q + \frac{1}{\epsilon}\Lambda\right]^{-1}U^T \\ \lim_{\epsilon\to 0}\left[I_q + \frac{1}{\epsilon}\tilde{\mathbf{I}}(\xi)^{1/2}BB^T\tilde{\mathbf{I}}(\xi)^{1/2}\right]^{-1} &= U_\perp U_\perp^T \end{aligned}$$

where $U_\perp$ corresponding to the space associated with zero eigenvalue of $\tilde{\mathbf{I}}(\xi)^{1/2}BB^T\tilde{\mathbf{I}}(\xi)^{1/2}$. Therefore taking $\epsilon \to 0$, we have

$$\begin{aligned} \lim_{\epsilon\to 0}\left[I_q + \frac{1}{\epsilon}\tilde{\mathbf{I}}(\xi)^{1/2}BB^T\tilde{\mathbf{I}}(\xi)^{1/2}\right]^{-1} &= \lim_{\epsilon\to 0} I_q - \tilde{\mathbf{I}}(\xi)^{1/2}B(\epsilon I_p + B^T\tilde{\mathbf{I}}(\xi)B)^{-1}B^T\tilde{\mathbf{I}}(\xi)^{1/2} \\ \tilde{\mathbf{I}}(\xi)^{-1/2}U_\perp U_\perp^T\tilde{\mathbf{I}}(\xi)^{-1/2} &= \tilde{\mathbf{I}}(\xi)^{-1} - B(B^T\tilde{\mathbf{I}}(\xi)B)^{-1}B^T \end{aligned}$$

where only the last step uses the fact $\tilde{\mathbf{I}}(\xi)$ is invertible. Therefore

$$\begin{aligned} \xi(\theta_{t+dt}) - \xi(\theta_t) &= B(\theta_{t+dt} - \theta_t) \\ &= -B\left[B^T\mathbf{I}_n(\xi)B\right]^{-1}B^T\nabla_\xi\tilde{L}(\xi)dt \\ &= -\eta\mathbf{I}(\xi)^{-1/2}(I_d - U_\perp U_\perp^T)\mathbf{I}(\xi)^{-1/2}\nabla_\xi\tilde{L}(\xi)dt \\ &= \mathbf{I}(\xi)^{-1/2}(I_d - U_\perp U_\perp^T)\mathbf{I}(\xi)^{1/2}\left\{\mathbf{I}(\xi)^{-1}\nabla_\xi\tilde{L}(\xi)dt\right\} \\ &= M_t(\xi_{t+dt} - \xi_t). \end{aligned}$$

The above claim asserts that in the over-parametrized setting, running natural gradient in the over-parametrized space is nearly "invariant" in the following sense: if $\xi(\theta_t) = \xi_t$, then

$$\begin{aligned} \xi(\theta_{t+dt}) - \xi(\theta_t) &= M_t\left(\xi_{t+dt} - \xi_t\right) \\ M_t &= \mathbf{I}(\xi_t)^{-1/2}(I_q - U_\perp U_\perp^T)\mathbf{I}(\xi_t)^{1/2} \end{aligned}$$

and we know M_t has eigenvalue either 1 or 0. In the case when $p = q$ and $\mathbf{J}_\xi(\theta)$ has full rank, it holds that $M_t = I$ is the identity matrix, reducing the problem to the re-parametrized case.

□

B Experimental details

In the realistic K-class classification context there is no activation function on the K-dimensional output layer of the network ($\sigma_{L+1}(x) = x$) and we focus on ReLU activation $\sigma(x) = \max\{0, x\}$ for the intermediate layers. The loss function is taken to be the cross entropy $\ell(y', y) = -\langle e_y, \log g(y')\rangle$, where $e_y \in \mathbb{R}^K$ denotes the one-hot-encoded class label and $g(z)$ is the softmax function defined by,

$$g(z) = \left(\frac{\exp(z_1)}{\sum_{k=1}^K \exp(z_k)}, \ldots, \frac{\exp(z_K)}{\sum_{k=1}^K \exp(z_k)}\right)^T.$$

It can be shown that the gradient of the loss function with respect to the output of the neural network is $\nabla \ell(f, y) = -\nabla \langle e_y, \log g(f) \rangle = g(f) - e_y$, so plugging into the general expression for the Fisher-Rao norm we obtain,

$$\|\theta\|_{\mathrm{fr}}^2 = (L+1)^2 \mathbb{E}[\{\langle g(f_\theta(X)), f_\theta(X) \rangle - f_\theta(X)_Y\}^2]. \tag{B.1}$$

In practice, since we do not have access to the population density $p(x)$ of the covariates, we estimate the Fisher-Rao norm by sampling from a test set of size m, leading to our final formulas

$$\|\theta\|_{\mathrm{fr}}^2 = (L+1)^2 \frac{1}{m} \sum_{i=1}^{m} \sum_{y=1}^{K} g(f_\theta(x_i))_y [\langle g(f_\theta(x_i)), f_\theta(x_i) \rangle - f_\theta(x_i)_y]^2 \ , \tag{B.2}$$

$$\|\theta\|_{\mathrm{fr,emp}}^2 = (L+1)^2 \frac{1}{m} \sum_{i=1}^{m} [\langle g(f_\theta(x_i)), f_\theta(x_i) \rangle - f_\theta(x_i)_{y_i}]^2 \ . \tag{B.3}$$

B.1 Additional experiments and figures

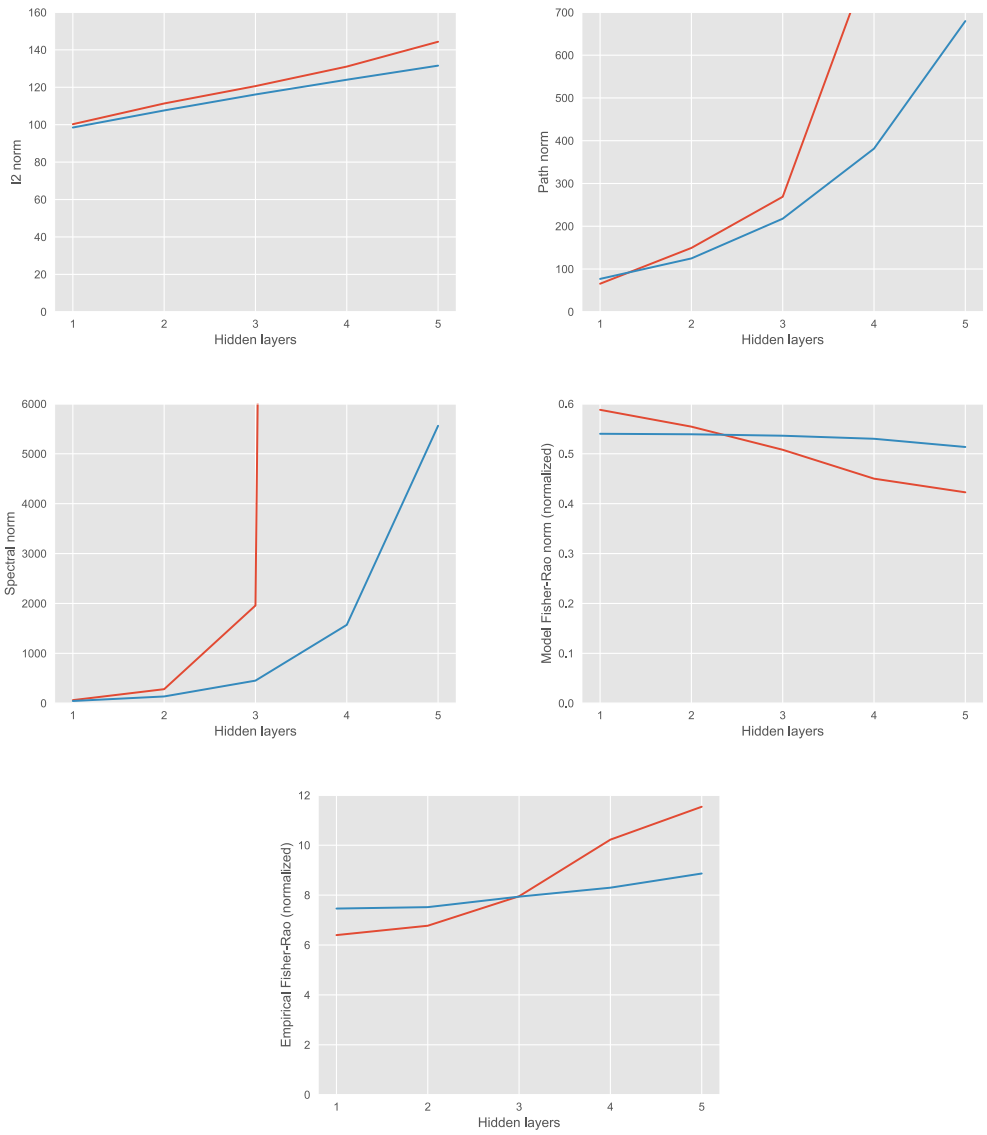

Figure 3: Dependence of different norms on depth L ($k = 500$) after optimzing with vanilla gradient descent (red) and natural gradient descent (blue). The Fisher-Rao norms are normalized by $L + 1$.

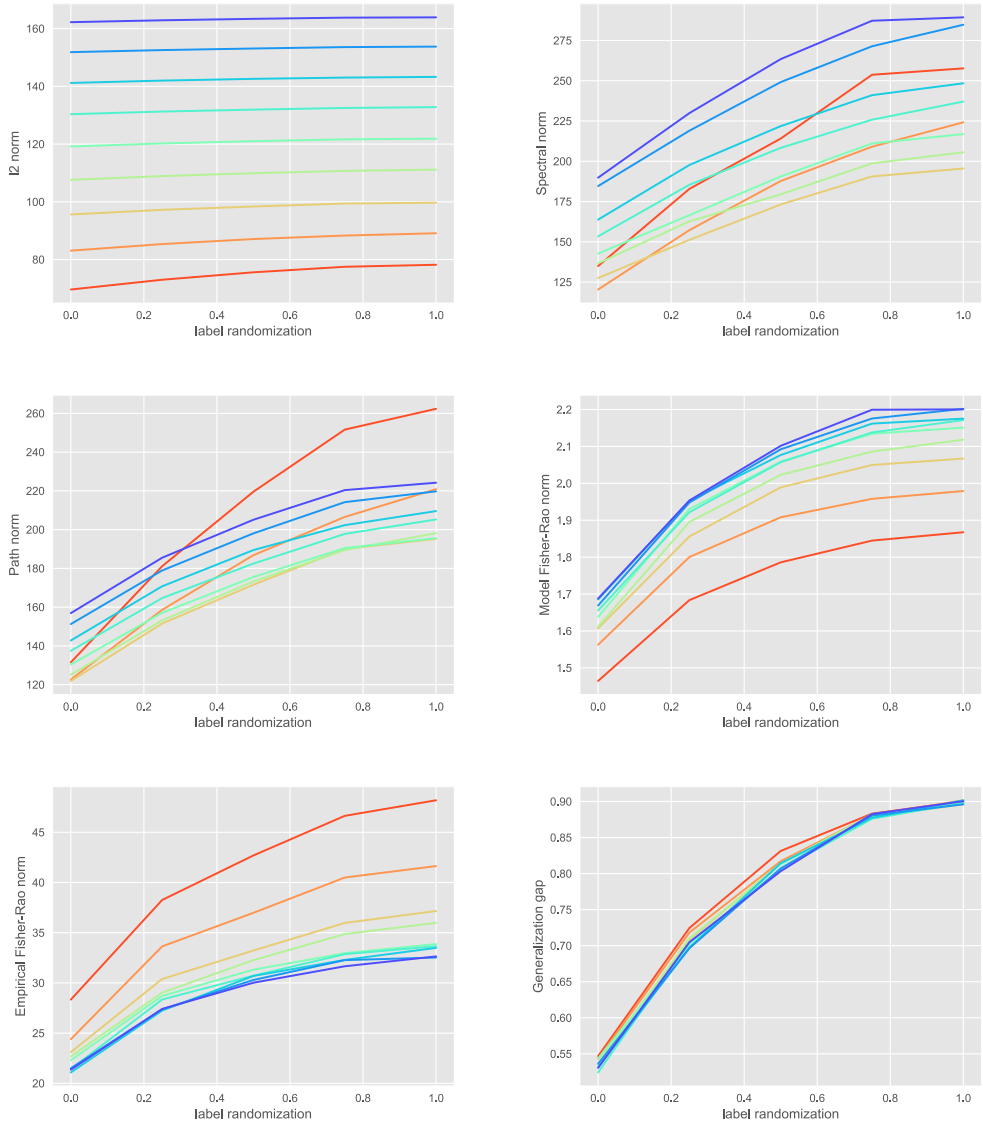

Figure 4: Dependence of capacity measures on label randomization after optimizing with natural gradient descent. The colors show the effect of varying network width from $k = 200$ (red) to $k = 1000$ (blue) in increments of 100. The natural gradient optimization clearly distinguishes the network architectures according to their Fisher-Rao norm.

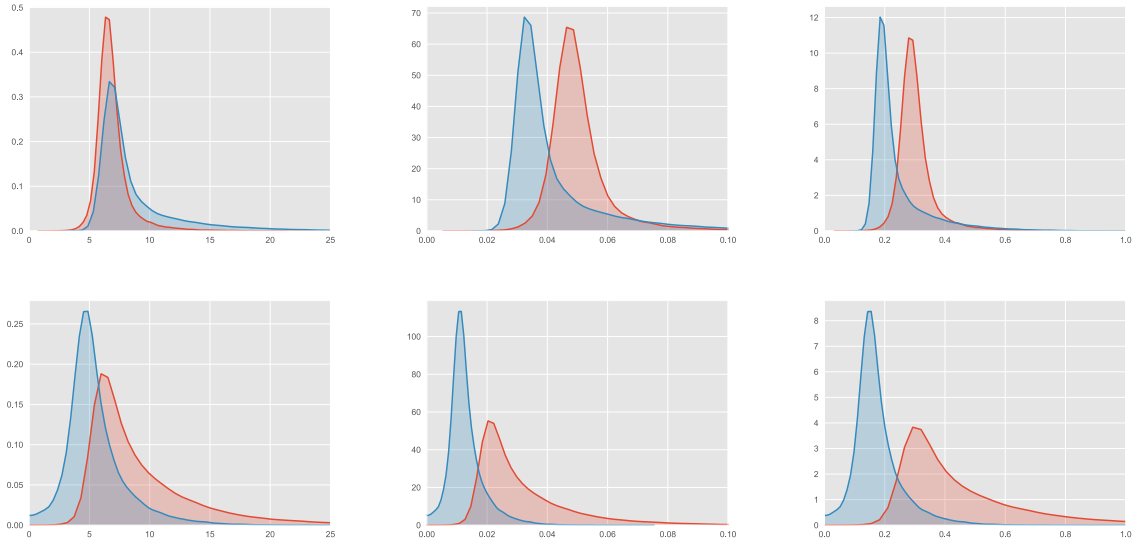

Figure 5: Distribution of margins found by natural gradient (top) and vanilla gradient (bottom) before rescaling (left) and after rescaling by spectral norm (center) and empirical Fisher-Rao norm (right).

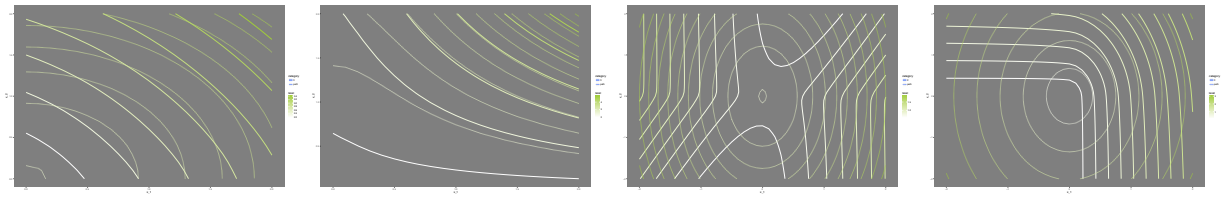

Figure 6: The levelsets of Fisher-Rao norm (solid) and path-2 norm (dotted). The color denotes the value of the norm.

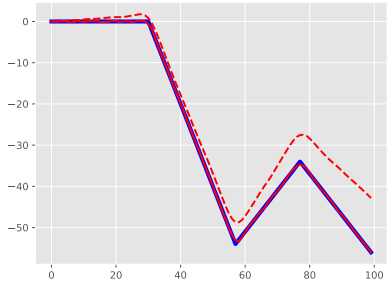

Figure 7: Reproduction of conditioning experiment from [16] after 10^4 iterations of Adam (dashed) and K-FAC (red).