CONVERGENCE OF GRADIENT FLOW UNDER OVERPARAMETERIZATION

Let $f(x;\omega,c) := \sum_{i=1}^{m} c_i\,\sigma(\omega_i^T x)$, $\forall x \in \mathbb{R}^d$.

Given $(x_j, y_j) \in \mathbb{R}^d \times \mathbb{R}$, $j = 1, 2, \dots, n$, let

$$f(\omega) := \begin{bmatrix} f(x_1;\omega,c) \\ f(x_2;\omega,c) \\ \vdots \\ f(x_n;\omega,c) \end{bmatrix} \in \mathbb{R}^n.$$

Our objective is $\frac{1}{\alpha^2} g(\alpha f(\omega))$ for a scale factor $\alpha > 0$.

Recall that $\alpha = \frac{1}{\sqrt{m}}$ yields the usual NTK scaling.

Gradient Flow Given a symmetric Xavier initialization $(\omega(0), c)$,

$$\dot{\omega}(t) = -\frac{1}{\alpha} J_{\omega(t)}^T \nabla g(\alpha f(\omega(t))), \quad t \geq 0,$$

where

$$J_\omega = \begin{bmatrix} \nabla^T f(x_1;\omega,c) \\ \vdots \\ \nabla^T f(x_n;\omega,c) \end{bmatrix} \in \mathbb{R}^{n \times md}.$$

Tangent Gradient Flow

$$T_f(u) = J_{\omega(0)}[u - \omega(0)] = \left[\nabla^T f(x_j;\omega(0),c)[\omega - \omega(0)]\right]_j$$

Note that this is the linearized model with random features $\nabla f(x;\omega(0),c)$.

For the linear model, the loss is: $\frac{1}{\alpha^2} g(\alpha T_f(u))$.

Starting from $u(0) = \omega(0)$,

$$\dot{u}(t) = -\frac{1}{\alpha} J_{\omega(0)}^T \nabla g(\alpha T_f(u(t))), \quad t \geq 0.$$

Note: GF is the continuous-time version of GD.

Assumptions :

$$\mathrm{rank}(J_{\omega(0)}) = n$$

$$\sigma_{min} = \sqrt{\lambda_{min}(J_{\omega(0)} J_{\omega(0)}^T)} > 0,$$

$$\| J_\omega - J_v \|_2 \leq L \cdot \| \omega - v \|, \quad \forall \omega, v.$$

Moreover, g is μ-sc. and ν-smooth.

Exercise : Verify that $v \mapsto \| v - y \|_2^2$ satisfies sc and smoothness assumptions, for any $y \in \mathbb{R}^n$.

THEOREM | $\kappa = \frac{\nu}{\mu}$. Let f^* be the minimizer of g.

If $\alpha > \frac{\|f^*\|}{M}$ for $M := \frac{\sigma_{min}^3}{32\kappa^{3/2} \|J_{\omega(0)}\|_2 L}$, then, for $t \geq 0$,

$$\| \alpha f(\omega(t)) - f^* \| \leq \sqrt{\kappa} \, \| \alpha f(\omega(0)) - f^* \|_2 \cdot e^{-\mu \sigma_{min}^2 t/4}.$$

Furthermore, as $\alpha \to \infty$,

$$\sup_{t \geq 0} \| \omega(t) - \omega(0) \|_2 = O(1/\alpha),$$

Notes : We will relax the full-rank assumption later in Lemma 3, and show that overparameterization ($m = \tilde{O}(n^2)$) is required to ensure $\sigma_{min} > 0$.

Above, $\| \omega(t) - \omega(0) \| \leq O(\frac{1}{\alpha})$ is satisfied without projection.

Previously, we needed projection to keep $\| \omega(t) - \omega(0) \|$ bounded.

But, we just required $m = O(\log(n))$ neurons rather than $\tilde{O}(n^2)$.

Proof Trajectory in the function space:

$$\frac{d}{dt}\alpha f(\omega(t)) = -J_{\omega(t)} J_{\omega(t)}^T \nabla g(\alpha f(\omega(t))).$$

Let $r_0 := \frac{\sigma_{min}}{8L}$. Then,

$$\|\omega - \omega(0)\|_2 < r_0 \Rightarrow J_\omega J_\omega^T \succcurlyeq \frac{\sigma_{min}^2}{4} I.$$ (Lemma 1 (next page))

Let

$$T = \inf\{t \geq 0 : \|\omega(t) - \omega(0)\|_2 > r_0\}.$$

We want to show that $T = \infty$.

From gradient flow (i.e., $\dot\omega(t) = -\frac{1}{\alpha} J_{\omega(t)}^T \nabla g(\alpha f(\omega(t)))$),

$$\|\dot\omega(t)\|_2 \leq \frac{1}{\alpha} \|J_{\omega(t)}\|_2 . \|\nabla g(\alpha f(\omega(t))) - \underbrace{\nabla g(f^*)}_{=0}\|_2$$

$$\leq \frac{2\nu}{\alpha} \|J_{\omega(0)}\|_2 . \|\alpha f(\omega(t)) - f^*\|_2 \quad \text{since } g \text{ is } \nu\text{-sm.}$$

By Lemma 2, for $t \in [0,T]$,

$$\|\omega(t) - \omega(0)\| = \left\|\int_0^t \dot\omega(s) ds\right\| \leq \int_0^t \|\dot\omega(s)\| ds \leq \int_0^t \|\alpha f(\omega(s)) - f^*\| ds \, \frac{2\nu}{\alpha} \|J_{\omega(0)}\|$$

$$\leq \frac{2\nu}{\alpha} \|J_{\omega(0)}\| . \int_0^t \sqrt{\frac{\nu}{\mu}} . \|\alpha f(\omega(0)) - f^*\| . e^{-\frac{\mu\sigma_{min}^2}{4} s} ds$$

$$\leq \frac{8\kappa^{3/2}}{\alpha\sigma_{min}^2} \|J_{\omega(0)}\| . \|\alpha f(\omega(0)) - f^*\|_2 = \frac{8\kappa^{3/2}}{\alpha\sigma_{min}^2} \|J_{\omega(0)}\| . \|f^*\|$$

Sufficiently large $\alpha \Rightarrow \frac{8\kappa^{3/2}}{\alpha\sigma_{min}^2} \|J_{\omega(0)}\| . \|f^*\| < r_0$

$$\Rightarrow T = \infty. \quad \blacksquare$$

Lemma 1 and Lemma 2 are on the following pages.

Lemma 1 Suppose $\|\omega - \omega(0)\| \leq r_0 = \frac{\sigma_{min}}{2L}$. Then,

$$\sigma_{min}(J_\omega) \geq \sigma_{min}(J_0) - L \cdot \|\omega - \omega(0)\| \geq \frac{\sigma_{min}}{2}.$$

Proof
Given u, let $A_u = J_{\omega(0)}^T u$ and $B_u = (J_\omega - J_{\omega(0)})^T u$.

Then,

$$\sigma_{min}^2(J_\omega) = \min_{\|u\|=1} u^T J_\omega J_\omega^T u$$

$$= \min_u \left[(J_{\omega(0)} + J_\omega - J_{\omega(0)})^T u \right]^T \left[(J_{\omega(0)} + J_\omega - J_{\omega(0)})^T u \right]$$

$$= \min_u \|A_u\|^2 + 2 A_u^T B_u + \|B_u\|_2^2$$

$$\overset{C.S.}{\geq} \min_u \|A_u\|^2 - 2\|A_u\| \cdot \|B_u\| + \|B_u\|^2$$

$$= \min_u (\|A_u\| - \|B_u\|)^2 \geq \min_u (\sigma_{min} - \beta L)^2 \|u\|^2$$

$$= \frac{\sigma_{min}^2}{4}. \quad \blacksquare$$

Lemma 2

Suppose that g is μ-sc and ν-smooth, z^* is the global minimizer of g. Let $Q(t)$ be a continuous linear operator
auto adjoint
s.t.

$$\inf_{t \in [0,\tau]} \lambda_{min}(Q(t)) \geq \lambda > 0.$$

Then, solutions on $[0,\tau]$ to

$$\dot{z}(t) = -Q(t)\nabla g(z(t))$$

satisfy, for $t \in [0,\tau]$,

$$\|z(t) - z^*\| \leq \sqrt{\frac{\nu}{\mu}} \cdot \|z(0) - z^*\| \cdot e^{-\mu \lambda t}$$

Pf: From μ-sc. of g, (Łojasiewicz Lemma)

$$\bar{g}(z) := g(z) - g(z^*) \leq \frac{1}{2\mu} \cdot \|\nabla g(z)\|^2.$$

Then,

$$\frac{d}{dt}\bar{g}(z(t)) \overset{\text{chain rule}}{=} -\nabla g^T(z(t))\, Q(t)\, \nabla g(z(t))$$

$$\leq -\lambda \|\nabla g(z(t))\|^2$$

$$\leq -2\mu\lambda \cdot [g(z(t)) - g(z^*)] = \bar{g}(z(t))$$

By Grönwall's lemma,

$$\bar{g}(z(t)) \leq \bar{g}(z(0)) \cdot e^{-2\mu\lambda}$$

Using μ-sc and ν-smooth properties of g above,

$$\frac{\mu}{2}\|z - z^*\|_2^2 \leq \bar{g}(z) \leq \frac{\nu}{2}\|z - z^*\|_2^2$$

$$\frac{\mu}{2}\|z(t) - z^*\|_2^2 \leq e^{-2\mu\lambda} \cdot \frac{\nu}{2}\|z(0) - z^*\|_2^2$$

$$\Rightarrow \|z(t) - z^*\| \leq e^{-\mu\lambda} \cdot \frac{\nu}{\mu} \cdot \|z(0) - z^*\|. \quad \blacksquare$$

Remark (An implication of the full-rank assumption)

$$T_f(w) = \begin{bmatrix} \nabla_w^T f(x_1; w(0), c)\,[w - w(0)] \\ \vdots \\ \nabla_w^T f(x_n; w(0), c)\,[w - w(0)] \end{bmatrix}$$ is the linearized regressor.

Consider

$$\min_{w \in \mathbb{R}^{md}} \frac{1}{2} \| T_f(w) - y \|_2^2 = \min_w \frac{1}{2} \| J_{w(0)} w - y_0 \|_2^2$$

where

$$y_0 = y + \underbrace{J_{w(0)} w(0)}_{=0 \text{ for ReLU}}.$$

Then, the normal equations:

$$J_{w(0)}^T J_{w(0)} w = J_{w(0)}^T y_0. \qquad (+)$$

$J_{w(0)}$ is full-rank $\Rightarrow$ $J_{w(0)} = \sum_{i=1}^n s_i u_i v_i^T$ SVD.

$$J_{w(0)}^+ = \sum_{i=1}^n s_i^{-1} v_i u_i^T$$

Then, from $(+)$:

$$(J_{w(0)}^+)^T J_{w(0)}^T J_{w(0)} w = (J_0^+)^T J_{w(0)}^T y_0$$

$$\Rightarrow J_{w(0)} w = \underbrace{\left[\sum_{i=1}^n u_i u_i^T \right]}_{\text{idempotent and full rank}} y_0 = y_0.$$

Choose $\hat{w} = J_{w(0)}^+ y_0$. Then, $J_{w(0)} \hat{w} = \sum_i u_i u_i^T y_0 = y_0$.

$$\Rightarrow \frac{1}{2} \| T_f(\hat{w}) - y \|_2^2 = \frac{1}{2} \| J_{w(0)} \hat{w} - y_0 \|_2^2 = 0.$$

No assumptions on y!

Remark How to ensure that $\lambda_{min}(J_{w(0)} J_{w(0)}^T) > 0$?

$\rightarrow$ Overparameterization.

$$\frac{1}{m}\left[J_{w(0)} J_{w(0)}^T\right]_{i,j} = \frac{1}{m}\nabla^T f(x_i; w(0), c)\, \nabla f(x_j; w(0), c)$$

$$= \frac{1}{m} x_i^T x_j \sum_{k=1}^{m} \sigma'(w_k^T(0) x_i)\, \sigma'(w_k^T(0) x_j)$$

$$\xrightarrow[m\to\infty]{} x_i^T x_j\, \mathbb{E}_{w_0 \sim N(0, I_d)}\left[\sigma'(w_0^T x_i)\, \sigma'(w_0^T x_j)\right].$$

$$= K_{ij}.$$

Lemma 3 Let $\|x_i\|_2 \le 1$, $|\sigma'| \le 1$.

Assume $\lambda_{min}(K) = \lambda_0 > 0$. Then,

if $m = \Omega\left(\frac{n^2}{\lambda_0^2} \log\left(\frac{n}{\delta}\right)\right)$, w.p. $\ge 1-\delta$,

$$\left\|\frac{1}{m} J_{w(0)} J_{w(0)}^T - K\right\|_2 \le \frac{\lambda_0}{4}, \text{ and}$$

$$\lambda_{min}\left(\frac{1}{m} J_{w(0)} J_{w(0)}^T\right) \ge \frac{3\lambda_0}{4}.$$

Proof: $\forall i, j \in \{1, \dots, n\}$,

$$\left|\frac{1}{m}\left[J_{w(0)} J_{w(0)}^T\right]_{i,j} - K_{ij}\right| \le \frac{2\sqrt{\log(n^2/\delta)}}{\sqrt{m}}$$

simultaneously w.p. $\ge 1-\delta$. Then,

$$\left\|\frac{1}{m} J_{w(0)} J_{w(0)}^T - K\right\|_2^2 \le \left\|\frac{1}{m} J_{w(0)} J_{w(0)}^T - K\right\|_F^2$$

$$\le \frac{8 \log(n/\delta)\, n^2}{m} \qquad \blacksquare$$