Lecture 13: NTK opt continued

Goal for these lectures:

* Idealized/abstract (smooth) network minimizes training error of a strongly convex loss by staying close to initialization.

Recall

$$f(w) = \begin{bmatrix} f(x_1; w) \\ \vdots \\ f(x_n; w) \end{bmatrix} \in \mathbb{R}^n, \qquad f \text{ smooth}$$

$$J_w^T = \begin{bmatrix} - \nabla f(x_1; w)^T - \\ \vdots \\ - \nabla f(x_n; w)^T - \end{bmatrix}^T \qquad \|J_w - J_v\| \le \beta \|v - w\|$$

Rem: Smooth?

if $\sigma : \mathbb{R} \to \mathbb{R}$ smooth, is $\sum_{j=1}^{m} a_j \sigma(w_j^T x)$?

$x \mapsto W_L \sigma(W_{L-1} \sigma(\dots \sigma(W_1 x) \dots))$

$\dots \quad w_L \; w_{L-1} \; w_{L-2}$

Two gradient flows:

$$\dot w(s) = -\nabla_w L(\alpha f(w(s)))$$
$$= -\alpha J_s \nabla L(\alpha f(w(s))) = -\alpha J (\alpha f(w(s)) - y)$$

"tangent flow"

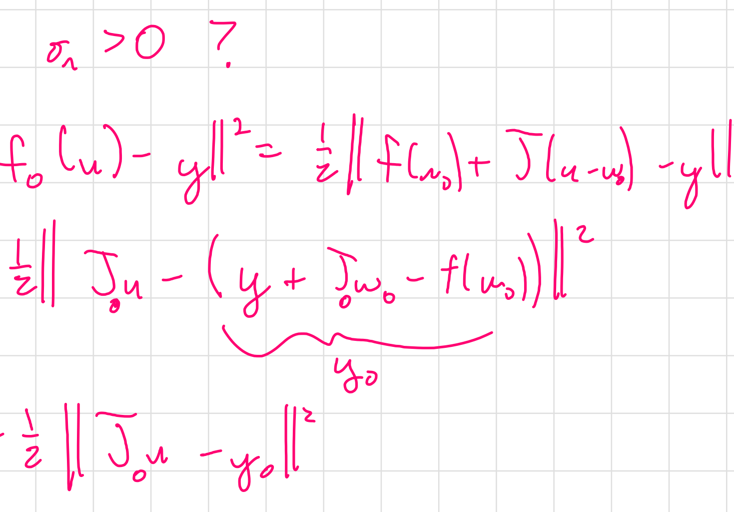
$$f_0(u) = f(w_0) + J_0^T (u - w_0)$$
$$\dot u(s) = -\nabla_u L(\alpha f_0(u(s)))$$
$$= -\alpha J_0 \nabla L(\alpha f_0(u(s))).$$

Theorem. Assume $\|J_w - J_v\| \le \beta \|w - v\|$, $\sigma_n(J_0) > 0$,

$$\alpha \ge \frac{\beta \sqrt{10^6 \sigma_{max}^3 L(\alpha f(w_0))}}{\sigma_{min}^3}$$

Then: $\max\{L(\alpha f(w(t))), L(\alpha f(u(t)))\} \le L(\alpha f(w(0))) \exp\left(\frac{-t\alpha^2 \sigma_n^2}{2}\right)$.

$$\max\{\|w(t) - w(0)\|, \|u(t) - u(0)\|\} \le \frac{3\sqrt{8\sigma_1^2(J_0) L(\alpha f(w_0))}}{\alpha \sigma_n^2(J_0)}$$



Rem. $\sigma_n > 0$?

$$\tfrac{1}{2}\|f_0(u) - y\|^2 = \tfrac{1}{2}\|f(w_0) + J(u - w_0) - y\|^2$$
$$= \tfrac{1}{2}\| J_0 u - \underbrace{(y + J w_0 - f(w_0))}_{y_0} \|^2$$
$$= \tfrac{1}{2}\|J_0 u - y_0\|^2$$
$$\xrightarrow{\nabla} J_0^T J_0 u = J_0^T y_0$$
$$\sigma_n(J_0) > 0 \longrightarrow u = (J_0^T J_0)^{-1} J_0^T y_0$$

Proof plan.

Define "convenient" radius $B := \frac{\sigma_n}{2\beta}$,

and earliest exit time: $T := \inf\{t > 0 : \|w_t - w_0\| > B\}$.

* Consider $t \in [0, T]$.
  * Then $\sigma_n(J_t) \ge \frac{\sigma_n(J_0)}{2}$
    $\Rightarrow$ strongly convex along GF path in $[0, T]$.
  * This implies quickly minimize $L(\alpha f(w(t)))$
  * Because $\nabla L(\alpha f(w_x)) = (\alpha f(w_x) - y)$,
    small loss $\Rightarrow$ small gradient norm
    $\Rightarrow$ iterates don't move much
  * We'll show $T = \infty$
    (stay in NTK; rare setting).

Lemma. If $\|w - w(0)\| \le B = \frac{\sigma_n}{2\beta}$:

① $\sigma_n(J_w) \ge \frac{\sigma_n}{2}$ & ② $\sigma_1(J_w) \le \frac{3}{2}\sigma_1$

Pf. ② $\sigma_1(J_w) = \|J_w\|_2 = \|J_w - J_0 + J_0\|_2$

$$\le \|J_w - J_0\| + \|J_0\| \le \beta \cdot B + \sigma_1$$
$$= \tfrac{3}{2}\sigma_1.$$

① $\sigma_n(J_w)^2 = \lambda_{min}(J_w^T J_w)$ ($\in \mathbb{R}^{n \times n}$)

$$= \min_{\|u\|=1} u^T J_w^T J_w u = \min_{\|u\|=1} u^T (J_w - J_0 + J_0)^T (J_w - J_0 + J_0) u$$
$$= \min_{\|u\|=1} \|u^T (J_w - J_0)\|^2 + 2u^T (J_w - J_0) J_0 u + \|J_0 u\|^2$$
$$\ge \min_{\|u\|=1} \|u^T (J_w - J_0)\|^2 - 2\|u^T (J_w - J_0)\| \cdot \|J_0 u\| + \|J_0 u\|^2$$
$$= \min_{\|u\|} \Big(\underbrace{\|u^T (J_w - J_0)\|}_{\le \beta \cdot B = \frac{\sigma_n}{2}} - \underbrace{\|J_0 u\|}_{\sigma_n} \Big)^2$$
$$\ge \left(\sigma_n - \frac{\sigma_n}{2}\right)^2 = \left(\frac{\sigma_n}{2}\right)^2.$$

$t \in [0, T] \Rightarrow \|w(t) - w(0)\| \le B$

$\Rightarrow \sigma_n(J_t) \ge \frac{\sigma_n}{2}$, $\sigma_1(J_t) \le \frac{3\sigma_1}{2}$.

Next: Loss quickly $\downarrow 0$ for $u(t)$ & $w(t)$ via abstract lemma.

Lemma Consider $\dot z(t) = -Q(t) \nabla L(z(t))$
where $\inf_{t \in [0,T]} \lambda_{min}(Q(t)) \ge \lambda > 0$.

$\forall t \in [0, T]$ $L(z(t)) \le L(z(0)) \exp(-2t\lambda)$.

Example. $z(t) = \alpha f(w(t)) \Rightarrow \dot z(t) = \frac{d}{dt} \alpha f(w(t))$

$$= \alpha J_t^T \dot w(t)$$
$$= -\alpha^2 \underbrace{J_t^T J_t}_{Q(t)} \nabla L(\alpha f(w(t)))$$

$\forall t \in [0, T]$

$$\Rightarrow L(\alpha f(w(t))) \le L(\alpha f(w(0))) \exp\left(-\alpha^2 \frac{\sigma_1^2}{2} t\right).$$

Pf. $\frac{d}{dt} L(z(t)) = \langle \nabla L(z(t)), \dot z(t) \rangle$

$$= \langle z(t) - y, -Q(t) \nabla L(z(t)) \rangle$$
$$= -(z(t) - y)^T Q(t) (z(t) - y)$$
$$\le -\lambda \|z(t) - y\|^2 = -2\lambda L(z(t))$$

Grönwall.

$$\Rightarrow L(z(t)) \le L(z(0)) \exp(-2t\lambda).$$

Still need to show T large.

Lemma: $\dot v(t) = -S(t) \nabla L(g(v(t)))$

Suppose $\forall t \in [0, T]$ $A \ge \sigma_1(S(t)) \ge \sigma_n(S(t)) \ge B$.

Then

$$\|v(t) - v(0)\| \le \frac{\sigma_n}{\sigma_1^2} \sqrt{2 L(g(v(0)))}$$

Why true: we've shown $L(\alpha f(w(t))) \downarrow 0$ exponentially fast,

$$\|\dot w(t)\| = \|J_t\| \cdot \|\nabla L(\alpha f(w(t)))\|$$
$$\ge \|J_t\| \cdot \underbrace{\|\alpha f(w(t)) - y\|}_{= \sqrt{2 L(\alpha f(w))}}$$

$\Rightarrow$ must also go down exponentially fast.