Lec 12

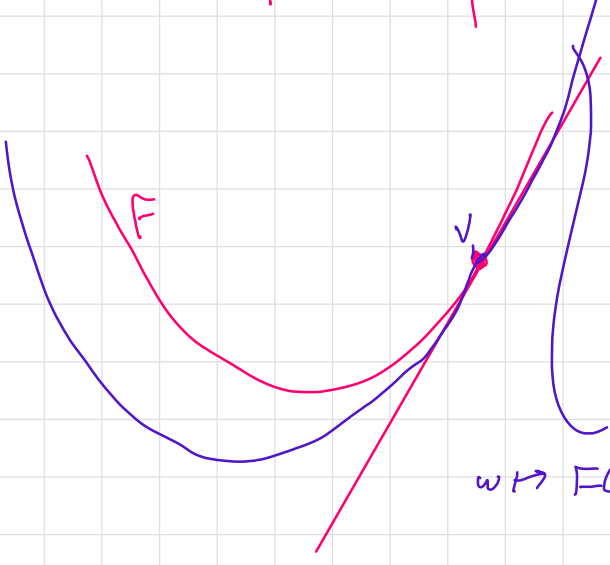
end of semi-classical; NTK optional

Ans:

* 15 line coding pro

* "α" discussion :- NTK notes is bad.

	β-smooth GD	exists GF	"Lyapunov" / Potential
opt stationary points (somewhere in first t iterates	$\frac{2\beta}{t} \cdot OC()$	$\frac{1}{t}$	$F(w_t)$.
Minimize convex F (in fact, also minimize $\|w_t - z\|^2$)	$\frac{\beta}{2t}$	$\frac{1}{t}$	$\frac{1}{2}\|w_t - z\|^2$
λ-strongly-convex (a) minimize F (b) minimize $\frac{1}{2}\|w_t - \bar{w}\|^2$	$\exp(-\frac{\lambda}{\beta}t)$	$\exp(-\lambda t)$	(a) $F(w_t)$ (b) $\frac{1}{2}\|w_t - \bar{w}\|^2$



Remark ("units" of GD & GF)

"arc length" of GD&GF paths?

assume $\|\nabla F(w)\|_2 \le 1$

arc length for GD: $\sum_{i<t} \eta_i \|\nabla F(w_i)\| \approx \frac{t}{\beta}$ ($\eta \le \frac{1}{\beta}$)

for GF: $\int_0^t \|\nabla F(w(s))\| ds \le t$

Theorem. F is λ-sc, then $\bar{w}$ with $F(\bar{w}) = \inf_w F(w)$, & if GF exists, then

(a) $F(w(t)) - F(\bar{w}) \le (F(w(0)) - F(\bar{w}))\exp(-2\lambda t)$,

(b) $\frac{1}{2}\|w(t) - \bar{w}\|^2 \le \frac{1}{2}\|w(0) - \bar{w}\|^2 \exp(-2\lambda t)$.

Proof. not doing (a); for (b):

$$\frac{1}{2}\|w(t) - \bar{w}\|^2 - \frac{1}{2}\|w(0) - \bar{w}\|^2 = \int_0^t \frac{d}{ds}\frac{1}{2}\|w(s) - \bar{w}\|^2 ds$$

$$= \int_0^t \langle w(s) - \bar{w}, \dot{w}(s) \rangle ds \quad (\frac{d}{ds} w(s))$$

$$= \int_0^t \langle w(s) - \bar{w}, -\nabla F(w(s)) \rangle ds$$

$$= -\int_0^t \langle w(s) - \bar{w}, \nabla F(w(s)) - \nabla F(\bar{w}) \rangle ds \quad (\nabla F(\bar{w}) = 0)$$

≥ 0 convex; $\ge \lambda \|w(s) - \bar{w}\|^2$

$$\le -\int_0^t \lambda \|w(s) - \bar{w}\|^2 ds$$

$\le \dots ?$

define $u(s) := \frac{1}{2}\|w(s) - \bar{w}\|^2$

$$\frac{d}{ds} u(s) \le -2\lambda u(s). \quad (q(s))$$

By Grönwall's inequality

$$u(t) \le u(0) \exp\left(\int_0^t q(s) ds\right)$$

$$= u(0) \exp\left(-\int_0^t 2\lambda ds\right)$$

$$= u(0) \exp(-2\lambda t).$$

standard tool, will use with NTK.

Remark (on SGD).

* With β-smooth, had $+\eta^2 \|\nabla F(w)\|^2$ term, used $\|\nabla F(w_i)\|^2 \le 2\beta(F(w_i) - F(w_{i+1}))$

* For stochastic, in most theory papers, use $\eta = \frac{1}{\sqrt{t}}$, thus $\eta^2 = \frac{1}{t}$, which eats the $\frac{1}{\eta^2}\|\nabla F(w)\|^2$.

* Theory does capture that stochastic gradients are cheap but doesn't capture statistical benefits. } active area

Standard NTK proof (via strong convexity).

Remarks (on approach)

* This proof style is most common (> 99%).
* Strong convexity ⇒ what is λ?
  λ depends on strong convexity of loss and λ_{min}(kernel gram matrix)
  $= \lambda_{min}\left(\left[\left\langle \nabla f(x_i; w_0), \nabla f(x_j; w_0) \right\rangle_{ij} \right] \right)$

(Recall with $\frac{1}{2}\|Xw - y\|^2 \to \sigma_{min}(X)$.)

* Proof plan: inductively maintain stay near initialization ⇒ proof is like linear case
* rate will be $\exp(-\lambda t)$
* Proof here is loosely based on Chizat-Bach ("Lazy training...")

Notation

Bake training set into predictor

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

$f: \mathbb{R}^p \to \mathbb{R}^n$

~~loss~~ "unnormalized empirical risk"

$$L(v) = \frac{1}{2}\|v - y\|^2$$

for us $L(\alpha f(w))$

$$= \frac{1}{2}\|\alpha f(w) - y\|^2$$

captures "zooming" whereby NTK with more parameters becomes linear.

Gradient

$$\dot{w}(t) = -\nabla_w L(\alpha f(w(t))) = -\begin{bmatrix} \nabla f(x_1; w(t))^T \\ \vdots \\ \nabla f(x_n; w(t))^T \end{bmatrix}^T \alpha \underbrace{(\alpha f(w_t) - y)}_{\nabla L(\alpha f(w_t))}$$

J_t^T — interpret as "data using features at time t"

"tangent flow"

$$u(0) = w(0)$$

$$\dot{u}(t) = -\alpha J_0^T \nabla L(\alpha f(u(t)))$$

One point: we'll show $\|u(t) - w(t)\|$ small to control strong convexity.

* $\operatorname{rank}(\underbrace{J_0 J_0^T}_{\text{kernel gram matrix}}) = n$; $\sigma_n(J_0) > 0$
* Smoothness: $\|J_w - J_v\| \le \beta \|w - v\|$

⇒ if $\sigma_n(J_t J_t^T) \approx \sigma_n(J_0 J_0^T)$ if $\|w_t - w_0\|$.

where $J_v^T = \begin{bmatrix} \nabla f(x_1; v)^T \\ \vdots \\ \nabla f(x_n; v)^T \end{bmatrix}^T$

Office hours

we used $\nabla f(\bar{w}) = 0$

optimality condition for unconstrained

constrained: $-\nabla f(\bar{w}) = N_S(\bar{w})$