Lecture 11: Still semi-classical convex opt

Announcements

* Shifting to hybrid Thursdays.
  Loomis — 1017 ; Sibel 1105 ≈ 10/14

Recap: Optimization

GD: $w' := w - \eta \nabla F(w)$

β-smooth: $\|\nabla F(w) - \nabla F(v)\| \le \beta \|w - v\|$

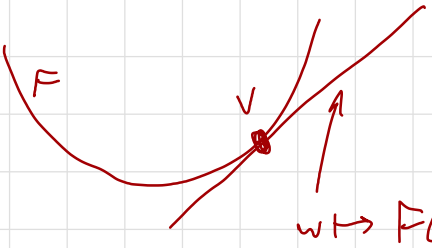
$$\Rightarrow F(w) \le F(v) + \langle \nabla F(v), w - v \rangle + \frac{\beta}{2}\|w - v\|^2$$

$$\Rightarrow \min_{i \le t} \|\nabla F(w_i)\|^2 \le \frac{2\beta}{t}\left(F(w_0) - F(w_t)\right)$$

"apx stationary point"

F convex

$$F(w) \ge F(v) + \langle \nabla F(v), w - v \rangle$$



Theorem. F β-smooth, convex, $\eta = 1/\beta$, $\forall z \in \mathbb{R}^d$

$$F(w_t) \le F(z) + \frac{\beta}{2t}\left(\|w_0 - z\|^2 - \|w_t - z\|^2\right).$$

↑ recap.

"self-regularization"
(see book §3, or margin lectures)

Remark/reminder

* Doing smooth & strong convexity analyses because both proof schemes pervasive for NTK opt.
* Not doing SGD because theory story is still unclear.

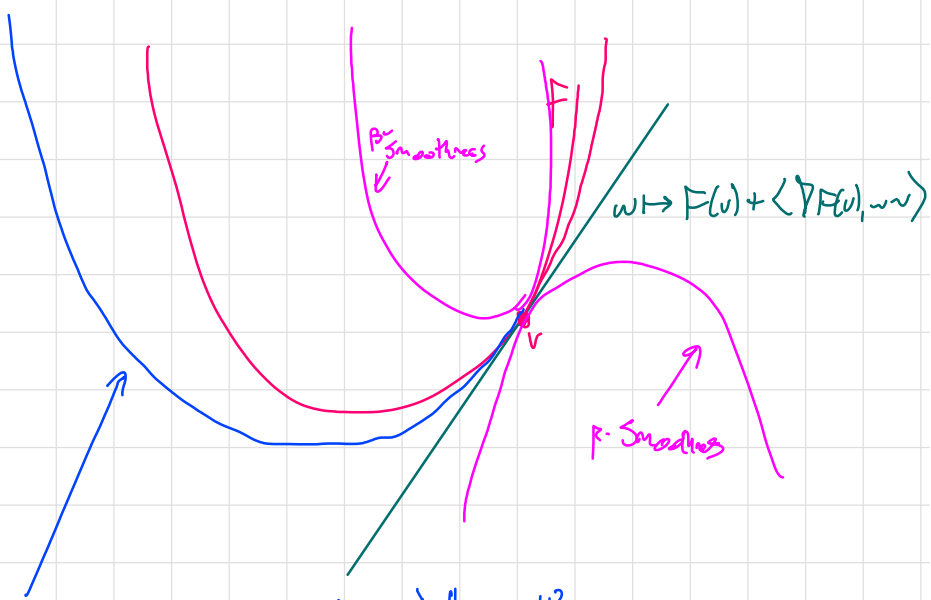
Strong convexity.

F is "λ-sc" (λ-strongly-convex)

if

$\forall w, v$

$$F(w) \geq F(v) + \langle \nabla F(v), w - v \rangle + \frac{\lambda}{2}\|w - v\|^2.$$



$$w \mapsto F(v) + \langle \nabla F(v), w - v \rangle + \frac{\lambda}{2}\|w - v\|^2$$

Exercise: which functions have $\beta = \lambda = 1$?

Answer: $w \mapsto \frac{1}{2}\|w\|^2 + a^T w + b$.

Ex.

Regularize convex $F \to F_\lambda(w) = F(w) + \frac{\lambda}{2}\|w\|^2$

$\Rightarrow F_\lambda$ is λ-sc.

Unnormalized least squares

$$w \mapsto \frac{1}{2}\|Xw - y\|^2$$

if $\sigma_{min}(X) > 0$

$\sigma_{min}(X)$-sc, $\sigma_{max}(X)$-smooth

(min might be 0 (not sc!))

Recall β-smooth $\Rightarrow$ $F\left(w - \frac{1}{\beta}\nabla F(w)\right) \leq F(w) - \frac{1}{2\beta}\|\nabla F(w)\|^2$

Lemma. F is λ-sc,

$$F\left(w - \frac{1}{\lambda}\nabla F(w)\right) \geq F(w) - \frac{1}{2\lambda}\|\nabla F(w)\|^2$$

and $\inf_v F(v) \geq F(w) - \frac{1}{2\lambda}\|\nabla F(w)\|^2$.

or $F(w) \leq \frac{1}{2\lambda}\|\nabla F(w)\|^2 + \inf_v F(v)$.

Proof.

$$Q_w(v) := F(w) + \langle \nabla F(w), v - w \rangle + \frac{\lambda}{2}\|w - v\|^2$$

$\nabla F(w) + \lambda(w - v) = 0$

$v = w - \frac{1}{\lambda}\nabla F(w)$

$$\inf_v Q_w(v) = Q_w\left(w - \frac{1}{\lambda}\nabla F(w)\right)$$

$$= F(w) + \left\langle \nabla F(w), -\frac{1}{\lambda}\nabla F(w) \right\rangle + \frac{\lambda}{2}\left\|\frac{1}{\lambda}\nabla F(w)\right\|^2$$

$$= F(w) - \frac{1}{2\lambda}\|\nabla F(w)\|^2.$$

$$\inf_v F(v) \geq \inf_v Q_w(v) = F(w) - \frac{1}{2\lambda}\|\nabla F(w)\|^2$$

Reminder. If F is β-smooth,
$\Rightarrow$ convergence rate $1/t$.

Theorem. If F is β-smooth & λ-sc,
$F(\bar{w}) = \inf_v F(w)$, and

$$F(w_t) - F(\bar{w}) \leq \left(F(w_0) - F(\bar{w})\right)\exp\left(-\frac{\lambda}{\beta}t\right),$$

$$\|w_t - \bar{w}\|^2 \leq \|w_0 - \bar{w}\|^2 \exp\left(-\frac{\lambda}{\beta}t\right).$$

Remark. See Nesterov's book for simultaneous proof & faster rate.

Proof. See notes for last part,

$$F(w_{i+1}) - F(\bar{w}) \leq F(w_i) - F(\bar{w}) - \frac{1}{2\beta}\|\nabla F(w_i)\|^2$$

$$\leq F(w_i) - F(\bar{w}) - \frac{\lambda}{\beta}\left(F(w_i) - F(\bar{w})\right)$$

$$= \left(1 - \frac{\lambda}{\beta}\right)\left(F(w_i) - F(\bar{w})\right),$$

unroll

$$\Rightarrow \left(F(w_t) - F(\bar{w})\right) \leq \left(F(w_0) - F(\bar{w})\right)\left(1 - \frac{\lambda}{\beta}\right)^t$$

$$\leq \left(F(w_0) - F(\bar{w})\right)\exp\left(-\frac{\lambda}{\beta}t\right).$$

Remark/PSA.

$F \geq 0$, $F_\lambda(w) = F(w) + \frac{\lambda}{2}\|w\|^2$.

Consider $F_\lambda(v) \leq F_\lambda(0)$.

$$\frac{\lambda}{2}\|v\|^2 \leq F_\lambda(v) \leq F_\lambda(0) = F(0) + \frac{\lambda}{2}\|0\|^2$$

$$= F(0).$$

$$\Rightarrow \|v\| \leq \sqrt{\frac{2F(0)}{\lambda}}$$

$$\left\{v \in \mathbb{R}^p : F_\lambda(v) \leq F_\lambda(0)\right\} \subseteq \left\{v \in \mathbb{R}^p : \|v\| \leq \sqrt{\frac{2F(0)}{\lambda}}\right\}.$$

i.e., regularized $\Rightarrow$ constrained.

GF (gradient flow). $\frac{f(w_{k+1}) - f(w_k)}{\eta} = \nabla f(w_k)$

$$\dot{w}(t) = \frac{dw(t)}{dt} = -\nabla F(w(t)).$$ ("infinitessimal gradient descent")

Why: * simplifies many proofs
* seems to often not change story.

Remark: "works" for nonsmooth, see later lectures and/or "Clarke differential"

Stationary points:

$$F(w(t)) - F(w(0)) = \int_0^t \frac{d}{ds} F(w(s))\, ds \quad \text{(FTC)}$$

chain rule

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

$\dot{w}(s)$ ODE

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

$$= \int_0^t -\|\nabla F(w(s))\|^2\, ds$$

$$\le \int_0^t -\inf_{r \in [0,t]} \|\nabla F(w(r))\|^2\, ds$$

$$= -t \cdot \inf_{r \in [0,t]} \|\nabla F(w(r))\|^2$$

Theorem. $\inf_{r \in [0,t]} \|\nabla F(w(r))\|^2 \le \frac{1}{t}\left(F(w(0)) - F(w(t))\right).$

Now let's use convexity.

$$\tfrac{1}{2}\|w(t) - z\|^2 - \tfrac{1}{2}\|w(0) - z\|^2 \quad \text{(FTC)}$$

$$= \int_0^t \frac{d}{ds} \tfrac{1}{2}\|w(s) - z\|^2\, ds \quad \text{(chain rule)}$$

$$= \int_0^t \langle w(s) - z, \dot{w}(s) \rangle\, ds \quad \text{(characterization of } \dot{w}(s))$$

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

$$\le \int_0^t \left(F(z) - F(w(s))\right) ds \quad (F(w(r)) \text{ non-increasing})$$

$$\le \int_0^t \left(F(s) - F(w(t))\right) ds$$

$$= t\left(F(z) - F(w(t))\right). \quad \text{Rearrange:}$$

Theorem. F is convex, $z \in \mathbb{R}^d$ arbitrary,

$$F(w(t)) \le F(z) + \frac{1}{2t}\left(\|w(0) - z\|^2 - \|w(t) - z\|^2\right).$$

OFFICE HOURS

$VV^T \neq I$ if $k < d$

$$\begin{bmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{bmatrix} \in \mathbb{R}^{d \times k}$$

$$V^T V = \begin{bmatrix} - v_1^T - \\ \vdots \\ - v_n^T - \end{bmatrix} \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix} \in \mathbb{R}^{k \times k}$$

$$V = \begin{bmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{bmatrix} \in \mathbb{R}^{d \times k}$$

$k \leq d$

orthonormal means

$$v_i^T v_j = \begin{cases} 1 & i = j, \\ 0 & i \neq j. \end{cases}$$

F continuously differentiable in an open set containing $\{v \in \mathbb{R}^d : F(v) \leq F(\omega(0))\}$.