Convex Optimization for Machine Learning

1. Optimization problems in ML

- **Minimize expected risk:**

$$\min_{w \in \mathbb{R}^d} \left\{ R(w) := \mathbb{E}_{z \sim \mathcal{D}} [\ell(w, z)] \right\}$$

$\mathcal{D}$: data distribution over z (ex: $z = (x, y)$)

w: parameters

ℓ: loss function

e.g.: • linear models

• $\ell(w, z) = \ell(w, (x, y)) = \hat{\ell}(y, \langle w, \phi(x) \rangle)$

• $\hat{\ell}$: square loss : $\ell(w, z) = \frac{1}{2}(y - \langle w, \phi(x) \rangle)^2$

logistic/hinge loss

• NN $w = (W_1, \dots, W_L)$, $\ell(w, z)$ is non-convex in w.

- **Empirical risk minimization** (ERM)

Samples $z_i \sim \mathcal{D}$ i.i.d. $i = 1, \dots, n$

$$\hat{R}(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w, z_i)$$

(Regularized) ERM:

$$\min_{w} \hat{R}(w) + \frac{\lambda}{2} \|w\|^2 \quad \text{or} \quad \min_{\|w\| \le B} \hat{R}(w)$$

Empirical vs Expected Risk:

$$R(w) = \underbrace{R(w) - \hat{R}(w)}_{\text{estimation}} + \underbrace{\hat{R}(w)}_{\text{optimization}}$$

$\Rightarrow$ No need to optimize $\hat{R}$ below the estimation error! (Bottou & Bousquet '08)

Ex: linear model, $\|\phi(x)\| \leq R$

loss function G-Lipschitz

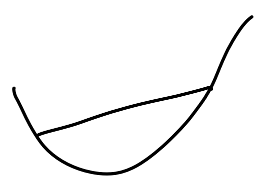
$$\mathbb{E}\left[\sup_{\|w\| \leq B} |R(w) - \hat{R}(w)|\right] \leq \frac{G \cdot R \cdot B}{\sqrt{n}}$$

In some cases, "faster rates" are possible $\sim \frac{1}{n}$

$\boxed{2.}$ Convexity, smoothness, gradient descent

Refs: Nesterov '18 "Lectures on convex optimization"

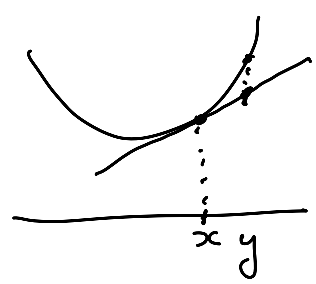
Bubeck '15 "Conv. Opt.: algorithms & complexity"



Def: (convexity)

f is convex if for any x, y:

$$f(y) \geq \underbrace{f(x) + \langle \nabla f(x), y - x \rangle}_{\text{linear approximation at } x}$$

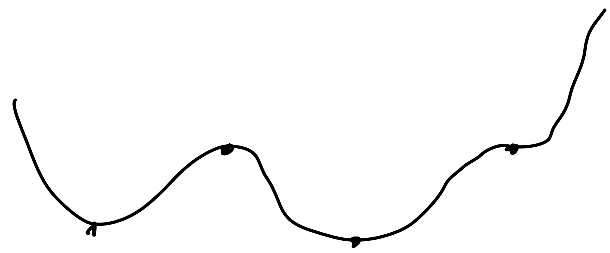


Why convexity?

→ local info ⇒ global info

Fact: x is a global minimum
$$\Longleftrightarrow$$
$$\nabla f(x) = 0$$

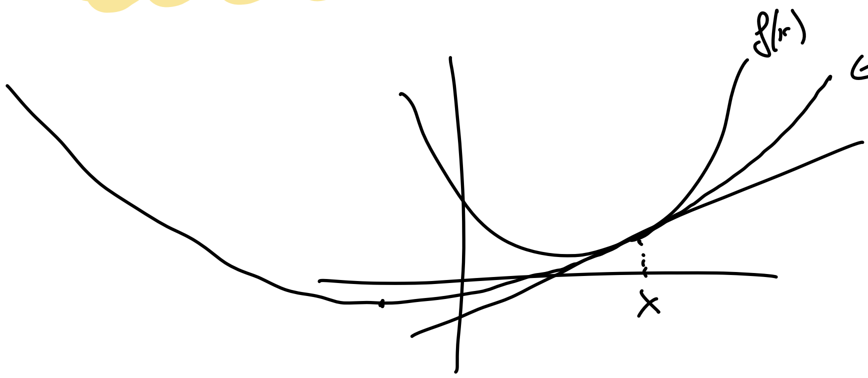
→ not true for non-convex



→ easy to bound $\underbrace{f(x) - f(x^*)}_{\text{suboptimality gap}}$ if x^* is global min

Strong convexity

Def: f is μ-strongly convex if for all x, y.
$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2$$



← quadratic lower bound on f.

Ex: • $\|Ax - b\|^2 = x^T \underline{A^T A} x$
↳ need $\lambda_{min}(A^T A) \geq \mu$

In the case of least squares, $\lambda_{min}(\Sigma) \geq \mu$
where $\Sigma = \mathbb{E}[\phi(x) \phi(x)^T]$

- ℓ^2 regularization $\frac{\mu}{2}\|w\|^2 \Rightarrow \mu$-strong conv.

Fact: Equivalent definition: (see Nesterov '18)

$$f(y) \leq f(x) + \langle \nabla f(x), y-x \rangle + \frac{1}{2\mu} \|\nabla f(x) - \nabla f(y)\|$$

In particular, if $x = x^*$

$$\|\nabla f(y)\|^2 \geq 2\mu \left(f(y) - f(x^*)\right)$$

(Łojasiewicz inequality)

(small gradient $\Rightarrow$ small suboptimality)

Fact: Convergence of gradient flow on μ-strongly cvx f:

$$\frac{d}{dt} x_t = -\nabla f(x_t)$$

x_0

$$\left(x_t = x_{t-1} - \gamma \nabla f(x_{t-1})\right)$$

We have:

$$f(x_t) - f(x^*) \leq \exp(-2\mu t) \left(f(x_0) - f(x^*)\right)$$

proof:

$$\frac{d}{dt}\left(f(x_t) - f(x^*)\right) = \frac{d}{dt}\left(f(x_t)\right)$$

$$= \langle \nabla f(x_t), \frac{d}{dt} x_t \rangle$$

$$= -\|\nabla f(x_t)\|^2$$

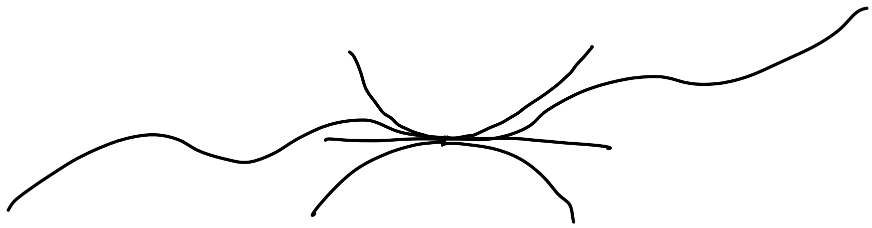
$$\leq -2\mu \left(f(x_t) - f(x^*)\right)$$

(Gronwall's Lemma) integrate this leads to the result $\square$

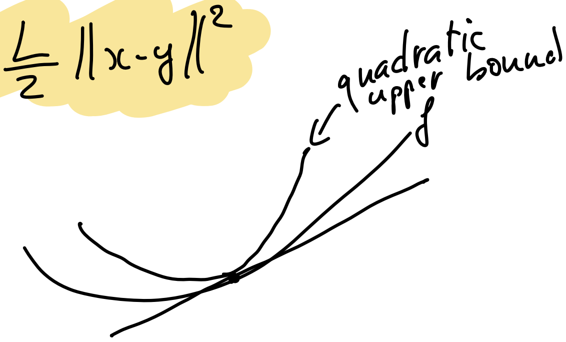
Remark: What about discrete time?
$\Rightarrow$ need smoothness!

Smoothness

Def: f is L-smooth if ∇f is L-Lipschitz, i.e.
$$\|\nabla f(x) - \nabla f(y)\| \leq L \cdot \|x-y\| \quad \text{for any } x, y.$$



Fact: If f is L-smooth, we ha
$$f(y) \leq f(x) + \langle \nabla f(x), y-x \rangle + \frac{L}{2}\|x-y\|^2$$



proof: exercise (fund. thm. of calculus)

Ex: • $\|Ax-b\|^2 \Rightarrow L = \lambda_{max}(A^TA)$

in least squares, $L = \lambda_{max}(\Sigma)$

• If loss is G-Lipschitz $\ell(\hat{y}, y)$, $\hat{y} = \langle w, \phi(x)\rangle$

$\|\phi(x)\| \leq R$

$L \leq G \cdot R^2$

Thm: Convergence of G.D. for L-smooth, μ-strongly convex f.

$$x_t = x_{t-1} - \gamma \nabla f(x_{t-1})$$

For $\gamma = \frac{1}{L}$:

$$f(x_t) - f(x^*) \leq \left(1 - \frac{\mu}{L}\right)^t \left(f(x_0) - f(x^*)\right)$$

proof:

$$f(x_t) \leq f(x_{t-1}) + \langle \nabla f(x_{t-1}), x_t - x_{t-1} \rangle + \frac{L}{2} \|x_t - x_{t-1}\|^2$$

$$= f(x_{t-1}) - \frac{1}{L} \|\nabla f(x_{t-1})\|^2 + \frac{1}{2L} \|\nabla f(x_{t-1})\|^2$$

$$= f(x_{t-1}) - \frac{1}{2L} \|\nabla f(x_{t-1})\|^2$$

$$f(x_t) - f(x^*) = f(x_{t-1}) - f(x^*) - \frac{1}{2L} \|\nabla f(x_{t-1})\|^2$$

$$\leq f(x_{t-1}) - f(x^*) - \frac{\mu}{L} \left(f(x_{t-1}) - f(x^*)\right)$$

$$= \left(1 - \frac{\mu}{L}\right) \left(f(x_{t-1}) - f(x^*)\right) \leq \dots \leq \left(1 - \frac{\mu}{L}\right)^t \left(f(x_0) - f(x^*)\right)$$

$\square$

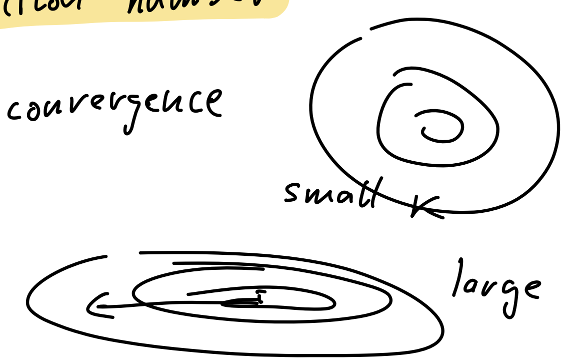
Remark:

- $\left(1 - \frac{\mu}{L}\right)^t \leq e^{-\frac{\mu}{L} t}$ → exponential / linear convergence

- $\kappa = \frac{L}{\mu} \geq 1$ is the condition number

  small κ ⇒ fast convergence

  $\left(\kappa = \frac{\lambda_{max}(\text{Hessian})}{\lambda_{min}(\text{Hessian})}\right.$ for quadratics$\left.\right)$



large κ

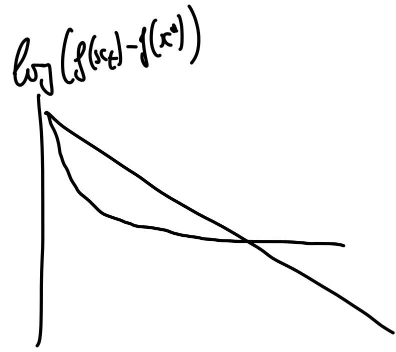
- In practice, κ may often be very large – often $\mu \approx \frac{1}{\sqrt{n}}$ or $\frac{1}{n}$

  $\Rightarrow$ can we obtain convergence rates independent of κ?

Thm: Convergence of G.D. under L-smoothness and convexity

With $\gamma = \frac{1}{L}$, we have

$$f(x_t) - f(x^*) \leq \frac{L}{2t} \|x_0 - x^*\|^2$$



proof: (Bansal & Gupta '2019)

control: $V_t(x_t) = t(f(x_t) - f(x^*)) + \frac{L}{2}\|x_t - x^*\|^2$

$\square$

- Other methods:

  - Nesterov's acceleration can achieve faster convergence rates

    $\rightarrow$ $\exp(-\frac{t}{\sqrt{\kappa}})$ instead of $\exp(-\frac{t}{\kappa})$ (str.cvx)

    or $\frac{1}{t^2}$ instead of $\frac{1}{t}$ (cvx)

This is optimal (cannot do better under these assumptions) for "first-order" methods (i.e. use only $\nabla f(x_t)$ at each t)

- **Newton's method** can converge much faster by computing Hessians

$$x_t = x_{t-1} - \gamma \, \mathrm{Hess}(x_{t-1})^{-1} \nabla f(x_{t-1})$$

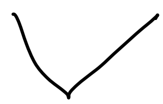
pros: break dependence on cond. number κ

cons: more costly (invert $d \times d$ matrix)

(Boyd & Vandenberghe book)

- **Proximal methods**

$$f(w) = \hat{R}(w) + \lambda \|w\|_1$$



- assume access to prox-oracle

$$\mathrm{prox}_{\|\cdot\|_1}(x) = \operatorname*{argmin}_z \|x-z\|^2 + \lambda \|z\|_1$$

- preserve similar convergence rates despite non-smooth term.

$\boxed{3.}$ Stochastic Gradient Descent

Until now: optimize $f(w) = \hat{R}(w) \; \{+ \frac{\lambda}{2}\|w\|^2\}$

treated as black-box, access gradients $\nabla f(w)$

- pro: fast convergence rates (linear when $\kappa < \infty$)
- cons: computing a single gradient $\nabla f(w)$

requires an entire pass over data (EXPENSIVE!)

<u>Q:</u> Can we bypass optimizing $\hat{R}(w)$ and directly optimize $R(w)$?

$$R(w) = \mathbb{E}_{z \sim \mathcal{D}}\left[\ell(w,z)\right]$$

optimization of $R(w)$ $\leftrightarrow$ generalization

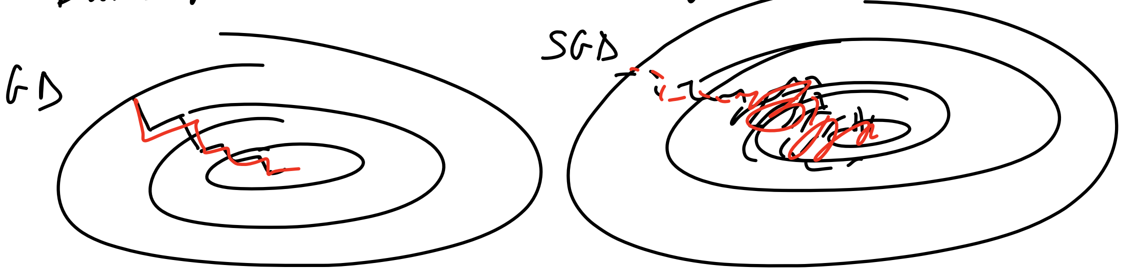
convergence rate on $R(\hat{w}) - R(w^*)$ $\leftrightarrow$ generalization bound

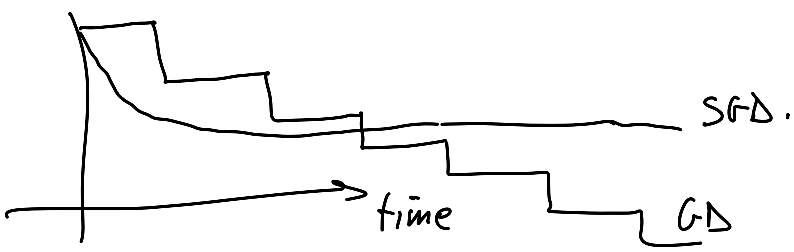
<u>SGD:</u> GD on $R(w)$, but replace $\nabla R(w)$ by stochastic gradients $\nabla_w \ell(w,z)$

$$\nabla R(w) = \nabla \mathbb{E}_z\left[\ell(w,z)\right] = \mathbb{E}_z\left[\underline{\nabla \ell(w,z)}\right]$$

<u>Remarks:</u>

- each step is much cheaper (gradient on a single datapoint)
- But: there is noise in the gradients





Theorem: Consider SGD updates:

$$w_t = w_{t-1} - \gamma \nabla \ell(w_{t-1}, z_t) \quad \text{with } z_t \sim \mathcal{D}$$

(fresh sample)

Assume:

$$\|\nabla \ell(w,z)\| \leq G \cdot R$$

$\left(\begin{array}{l} \tilde{\ell} \text{ is } G\text{-Lipschitz} \\ \|\phi(x)\| \leq R \end{array} \right.$

$\Rightarrow \ell(\cdot, z)$ is GR-Lipschitz $\Big)$

$$\|w^*\| \leq B$$

$\uparrow$

$w^* \in \arg\min R(w)$

Then, with $\gamma = \dfrac{B}{GR\sqrt{m}}$, SGD satisfies

$$\mathbb{E}\left[f(\bar{w}_m) - f(w^*) \right] \leq \frac{BGR}{\sqrt{m}}$$

$\left(f(w) = \mathbb{E}_z[\ell(w,z)] \right)$

where $\bar{w}_m = \dfrac{1}{m} \displaystyle\sum_{t=0}^{m-1} w_t \quad \left(= \dfrac{m-1}{m} \bar{w}_{m-1} + \dfrac{1}{m} w_{m-1} \right)$

proof: We have:

$$\|w_t - w^*\|^2 = \|w_{t-1} - \gamma \nabla \ell(w_{t-1}, z_t) - w^*\|^2$$

$$= \|w_{t-1} - w^*\|^2 - 2\langle w_{t-1} - w^*, \gamma \nabla \ell(w_{t-1}, z_t)\rangle + \underbrace{\|\gamma \nabla \ell(w_{t-1}, z_t)\|^2}_{\leq \gamma^2 G^2 R^2}$$

We have

$$\mathbb{E}\left[\|w_t - w^*\|^2 \mid w_{t-1}\right] \leq \|w_{t-1} - w^*\|^2 - 2\gamma \mathbb{E}\left[\langle w_{t-1} - w^*, \nabla \ell(w_{t-1}, z_t)\rangle \mid w_{t-1}\right] + \gamma^2 G^2 R^2$$

$$= \|w_{t-1} - w^*\|^2 - 2\gamma \langle w_{t-1} - w^*, \underbrace{\mathbb{E}_{z_t}\left[\nabla \ell(w_{t-1}, z_t) \mid w_{t-1}\right]}_{\nabla f(w_{t-1})}\rangle + \gamma^2 G^2 R^2$$

use convexity

$$f(w^*) \geq f(w_{t-1}) + \langle \nabla f(w_{t-1}), w^* - w_{t-1} \rangle$$

$$- \langle \nabla f(w_{t-1}), w_{t-1} - w^* \rangle \leq - \left(f(w_{t-1}) - f(w^*) \right)$$

This yields:

$$\mathbb{E}\left[\|w_t - w^*\|^2 \mid w_{t-1} \right] \leq \|w_{t-1} - w^*\|^2 - 2\gamma \left(f(w_{t-1}) - f(w^*) \right) + \gamma^2 G^2 R^2$$

Take expectation w.r.t. w_{t-1}

$$\mathbb{E}\left(\|w_t - w^*\|^2 \right) \leq \mathbb{E}\left[\|w_{t-1} - w^*\|^2 \right] - 2\gamma\, \mathbb{E}\left[f(w_{t-1}) - f(w^*) \right] + \gamma^2 G^2 R^2$$

Re-arrange:

$$2\gamma\, \mathbb{E}\left[f(w_{t-1}) - f(w^*) \right] \leq \mathbb{E}\|w_{t-1} - w^*\|^2 - \mathbb{E}\|w_t - w^*\|^2 + \gamma^2 G^2 R^2$$

Summing from $t=1$ to $t=m$,

$$2\gamma\, \mathbb{E}\left[\sum_{t=1}^{m} f(w_{t-1}) - f(w^*) \right] \leq \mathbb{E}\|w_0 - w^*\|^2 \cancel{- \mathbb{E}\|w_m - w^*\|^2} + m \cdot \gamma^2 G^2 R^2$$

$$\leq B^2 + m\gamma^2 G^2 R^2$$

$$\frac{1}{m} \mathbb{E}\left(\sum_{t=1}^{m} f(w_{t-1}) - f(w^*) \right) \leq \frac{B^2}{2\gamma m} + \frac{\gamma G^2 R^2}{2} = \frac{BGR}{\sqrt{m}}$$

$$\mathbb{E} f(\bar{w}_m) - f(w^*) \leq \left(\frac{1}{m} \mathbb{E} \sum_{t=0}^{m-1} f(w_t) \right) - f(w^*) \leq \frac{BGR}{\sqrt{m}}$$

Jensen's ineq.

Remarks: • matches estim. error w/ Rademacher compl.!

$\Rightarrow$ one-pass SGD can suffice

(not always! e.g. fast rates)

- step-size depends on n, can use $\frac{1}{\sqrt{t}}$ decreasing step-size instead
- Smoothness does not help, but it can help with mini-batches
- strongconvexity can improve rate to $\frac{1}{\mu n}$ instead of $\frac{1}{\sqrt{n}}$
- Can also use SGD to optimize

$$(*) \qquad \hat{R}(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, z_i) = \mathbb{E}_{i_t}\left[\ell(w, z_{i_t})\right]$$

by sampling $i_t \sim \mathrm{Unif}(\{1, \dots, n\})$

$\to$ but: no more generalization guarantees!

- For finite-sum problems $(*)$, there are faster algorithms! "Variance-reduction" (SAG, SAGA, SVRG, SDCA, MISO, ...)

similar rate as GD, at a fraction of cost (similar to SGD)

---

mini-batch: gradient: $g_t = \frac{1}{b}\sum_{i=1}^{b} \nabla \ell(w_{t-1}, z_t^i)$ $\quad z_t^i \sim \mathcal{D}$, $i = 1, \dots, b$

$w_t = w_{t-1} - \gamma g_t$ $\qquad$ instead of $g_t = \nabla \ell(w_{t-1}, z_t)$

replace $\|\nabla \ell(w, z)\| \leq GR$

by $\underbrace{\mathbb{E}\|g_t - \mathbb{E} g_t\|^2}_{b \text{ times smaller with mini-batch.}} \leq \sigma^2$ condition

$\sigma^2 = \frac{\sigma_0^2}{b}$

different bound: $\mathbb{E}\left[f(\bar{w}_T) - f(w^*)\right] \leq \frac{L\|w^*\|^2}{T} + \frac{\sigma_0\|w^*\|}{\underbrace{\sqrt{T \cdot b}}_{= n}}$