Online Subgradient Descent

Shuai Li

John Hopcroft Center, Shanghai Jiao Tong University

https://shuaili8.github.io

References

- References:
  - A Modern Introduction to Online Learning
    - Francesco Orabona, Associate Professor at Boston University
    - https://arxiv.org/abs/1912.13213

What is Online Learning?

A Repeated-game Example

- In each round $t = 1, \ldots, T$
  - An adversary choose a real number $y_t \in [0,1]$ and keeps it secret;
  - You try to guess the real number, choosing $x_t \in [0,1]$;
  - The adversary's number is revealed and you pay the squared difference $(x_t - y_t)^2$

What is "winning strategy"?

- Assume $y_t \overset{i.i.d.}{\sim} \mathfrak{D}$
- If we know $\mathfrak{D}$, then $x_t \coloneqq \mathrm{mean}(\mathfrak{D})$ and pay $\sigma^2 T$ in expectation
- So it's natural to measure $\mathbb{E}_Y[\sum_{t=1}^T (x_t - Y)^2] - \sigma^2 T$ $o(T)$
- Or equivalently $\frac{1}{T}\mathbb{E}_Y[\sum_{t=1}^T (x_t - Y)^2] - \sigma^2$ $\to 0$
- Minimize regret

$$\mathrm{Regret}_T \coloneqq \sum_{t=1}^{T} (x_t - y_t)^2 - \min_{x \in [0,1]} \sum_{t=1}^{T} (x - y_t)^2 \quad o(T)$$

- $\mathrm{Regret}_T(u) \coloneqq \sum_{t=1}^T (x_t - y_t)^2 - \sum_{t=1}^T (u - y_t)^2$

A Winning Strategy

- Best strategy in hindsight:

$$x_T^* := \arg\min_{x \in [0,1]} \sum_{t=1}^{T} (x - y_t)^2 = \frac{1}{T} \sum_{t=1}^{T} y_t$$

- Follow-the-leader (FTL):

$$x_t = x_{t-1}^* = \frac{1}{t-1} \sum_{i=1}^{t-1} y_i$$

FTL has $O(\ln T)$ regret

- Theorem 1.2. $y_t \in [0,1], \forall t.\ x_t = x_{t-1}^* = \frac{1}{t-1} \sum_{i=1}^{t-1} y_i$. Then

$$\text{Regret}_T = \sum_{t=1}^{T} (x_t - y_t)^2 - \min_{x \in [0,1]} \sum_{t=1}^{T} (x - y_t)^2 \leq O(\ln T)$$

- Remark
  - No parameters to tune
  - Doesn't need to maintain a complete record of the past, only a summary
  - Doesn't use gradient

Online Gradient Descent

Failure of FTL

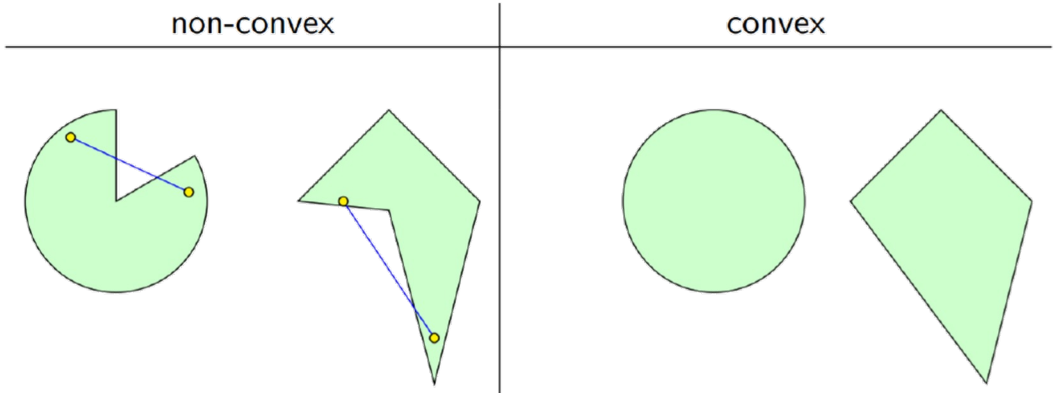
- Example. $V = [-1,1]$. $\ell_t(x) = z_t x + i_V(x)$

$i_V(x) = \begin{cases} 0, & \text{if } x \in V \\ +\infty, & \text{o.w.} \end{cases}$

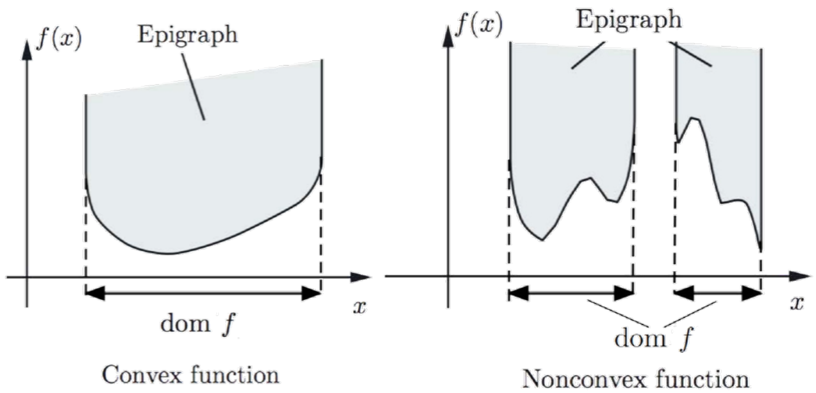
  - $z_1 = -0.5$
  - $z_t = 1, \quad \text{t} = 2,4,6, \ldots$
  - $z_t = -1, \quad \text{t} = 1,3,5, \ldots$
- FTL: $x_t = 1$ (t even); $x_t = -1$ (t odd)
  - Cumulative loss $= T$
  - But the cumulative loss for $u = 0$ is 0. Thus the regret is T

does not converge

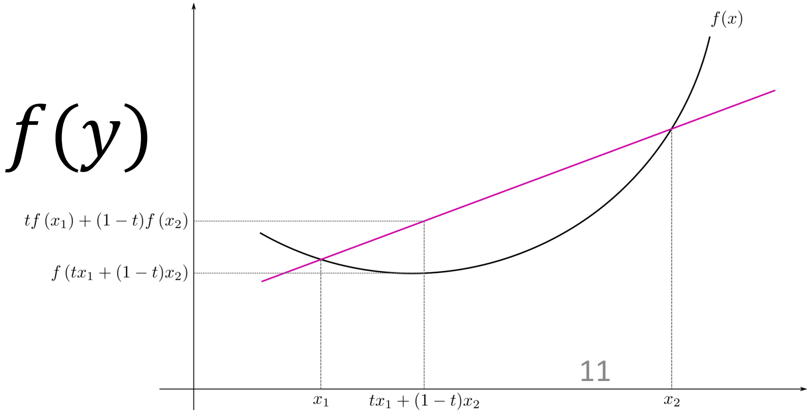
Convexity



- Definition 2.2. $V \subseteq \mathbb{R}^d$ is **convex** if $\forall x, y \in V, \lambda \in (0,1)$, there is $\lambda x + (1-\lambda)y \in V$
- Definition 2.3. $f: \mathbb{R}^d \to (-\infty, +\infty]$ is **convex** if the epigraph of the function
$$\{(x,y) \in \mathbb{R}^{d+1}: y \geq f(x)\}$$
is convex

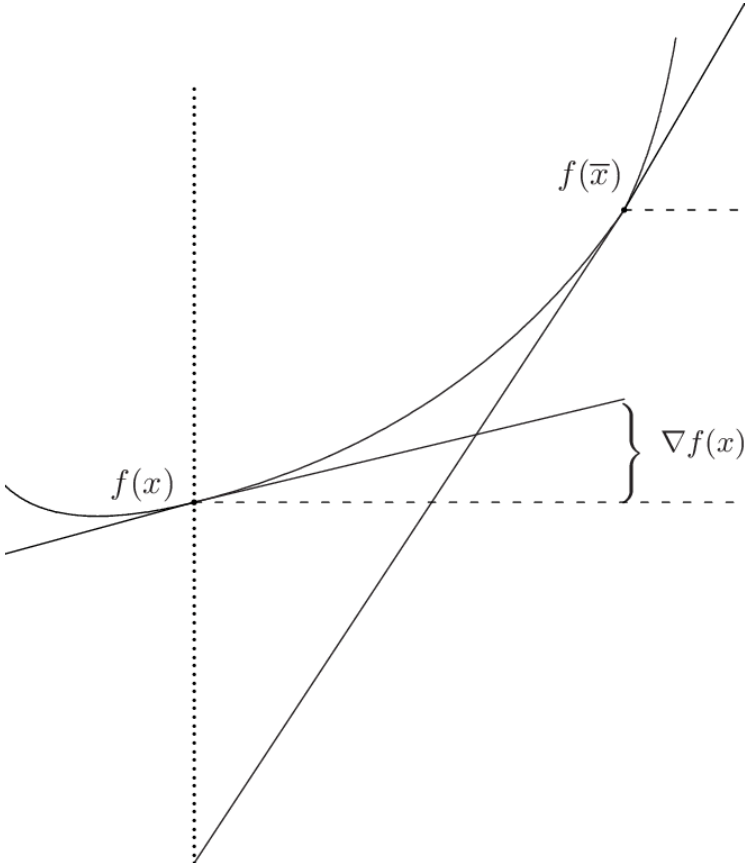


- Theorem 2.4. $f: \mathbb{R}^d \to (-\infty, +\infty]$. $\mathrm{dom}(f)$ is a convex set. Then f is convex $\Leftrightarrow \forall \lambda \in (0,1), x, y \in \mathrm{dom}(f)$
$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$$



Convexity

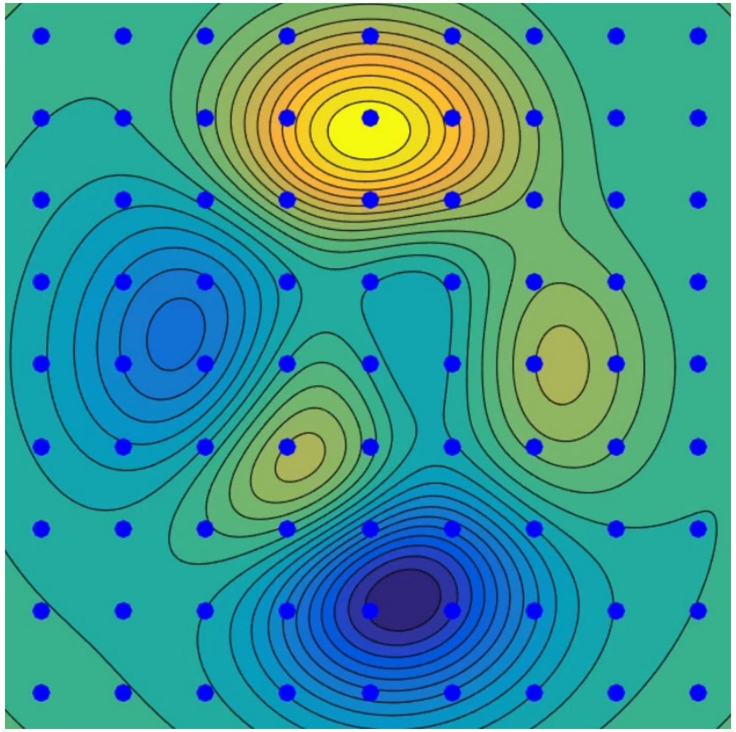
- Theorem 2.7. $f: \mathbb{R}^d \to (-\infty, +\infty]$ is convex. $x \in \operatorname{int}\operatorname{dom}(f)$. f is differentiable at x. Then
$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$

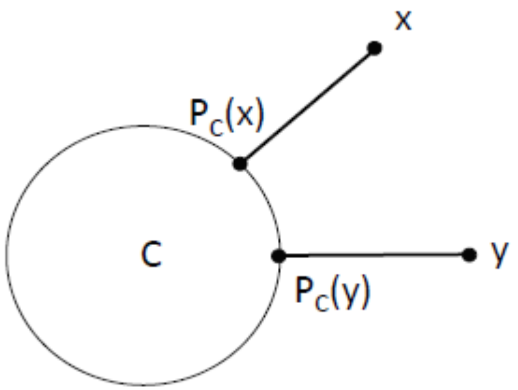


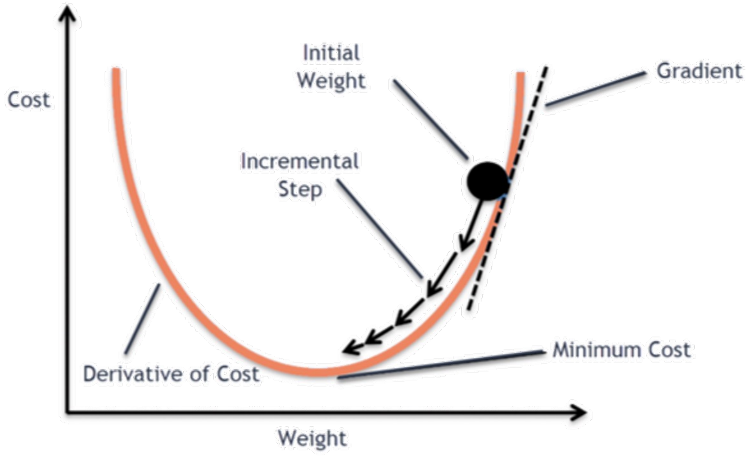
(Projected) Online Gradient Descent

- Require: Closed convex set $V \subseteq \mathbb{R}^d, x_1 \in V, \eta_1, \ldots, \eta_T > 0$
- For $t = 1{:}T$ do
  - Output x_t
  - Receive $\ell_t: \mathbb{R}^d \to (-\infty, +\infty]$ and pay $\ell_t(x_t)$
  - Set $g_t = \nabla \ell_t(x_t)$
  - $x_{t+1} = \Pi_V(x_t - \eta_t g_t) = \arg\min_{y \in V} \|x_t - \eta_t g_t - y\|_2$

Projection





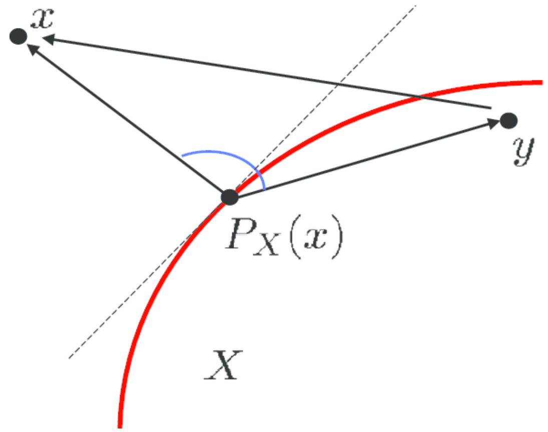


$O(\sqrt{T})$ Regret for OGD

$\max_{x,y\in V}\|x-y\|_2 \le D$

- Theorem 2.13. $\emptyset \neq V \subseteq \mathbb{R}^d$ closed convex set with diameter D. $\ell_t\colon \mathbb{R}^d \to (-\infty, +\infty]$ convex differentiable in open set containing $V, \forall t$. $x_1 \in V$. $\|\nabla \ell_t(\cdot)\|_2 \le L$. $\eta_t \equiv \frac{D}{L\sqrt{T}}$. Then $\forall u \in V$,

$$\sum_{t=1}^{T} (\ell_t(x_t) - \ell_t(u)) \le DL\sqrt{T}$$



could choose adaptive $\eta_t = \frac{D}{L\sqrt{t}}$

see Section 4.2.1

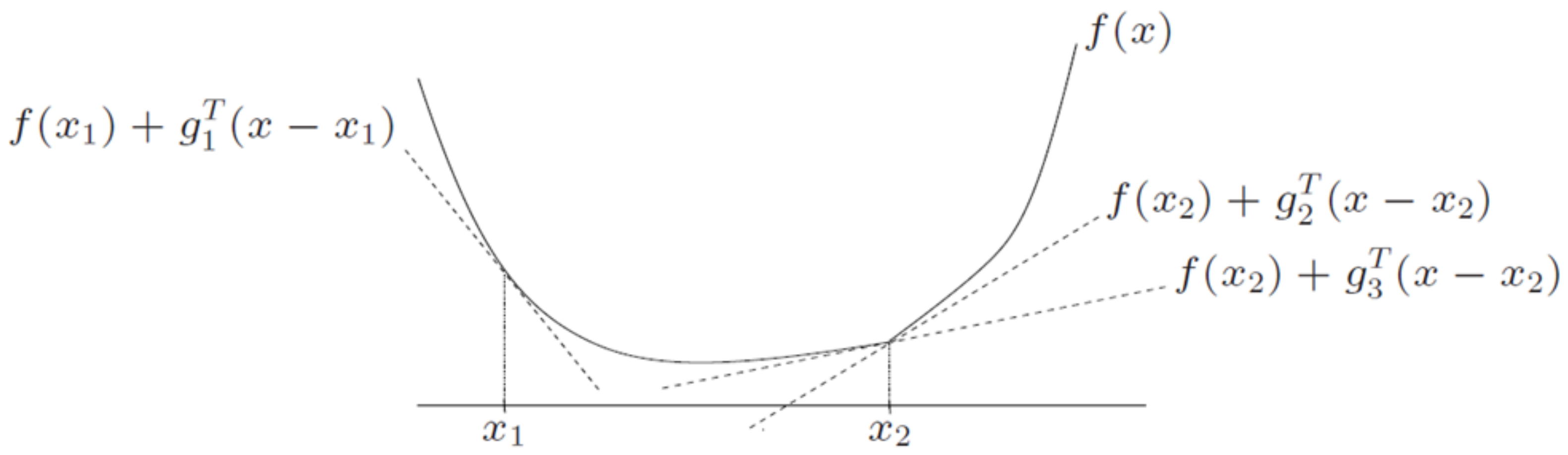
Online Subgradient Descent

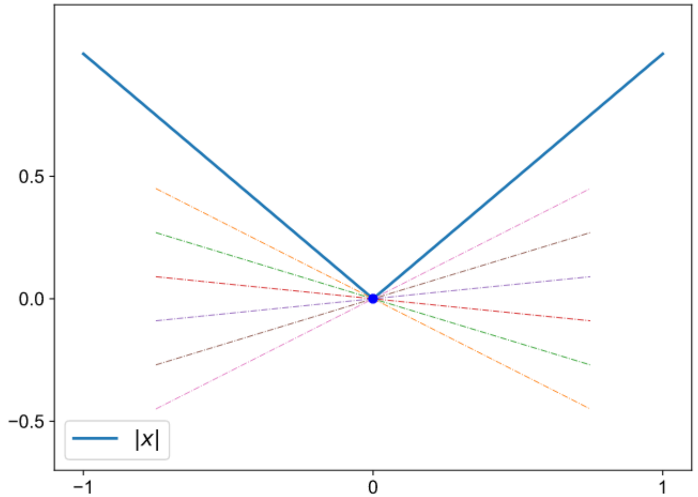
Subgradient

f is finite somewhere

- Definition 2.15. A subgradient of a proper function $f: \mathbb{R}^d \to (-\infty, +\infty]$ in $x \in \mathbb{R}^d$ is a vector $g \in \mathbb{R}^d$ satisfying

$$f(y) \geq f(x) + \langle g, y - x \rangle, \qquad \forall y \in \mathbb{R}^d$$





- $\partial f(x)$: subdifferential of f at x
  - The set of subgradients of f at x
- A proper convex function f is always subdifferentiable in $\mathrm{int\,dom}(f)$

Projected Online Subgradient Descent

- Require: Closed convex set $V \subseteq \mathbb{R}^d, x_1 \in V, \eta_1, \dots, \eta_T > 0$
- For $t = 1{:}\,T$ do
  - Output x_t
  - Receive $\ell_t: \mathbb{R}^d \to (-\infty, +\infty]$ and pay $\ell_t(x_t)$
  - Set $\cancel{g_t = \nabla \ell_t(x_t)}$ $g_t \in \partial \ell_t(x_t)$
  - $x_{t+1} = \Pi_V(x_t - \eta g_t) = \arg\min_{y \in V} \|x_t - \eta g_t - y\|_2$

- Lemma 2.23.

$$\ell_t(x_t) - \ell_t(u) \leq \langle g_t, x_t - u \rangle \leq \frac{\|x_t - u\|_2^2 - \|x_{t+1} - u\|_2^2}{2\eta_t} + \frac{\eta_t}{2}\|g_t\|_2^2$$

From Convex Losses to Linear Losses

- $\ell_t(x_t) - \ell_t(u) \leq \langle g_t, x_t - u \rangle$
- Online linear optimization!

- In each round $t = 1, \ldots, T$
  - An adversary choose a real number $y_t \in [0,1]$ and keeps it secret;
  - You try to guess the real number, choosing $x_t \in [0,1]$;
  - The adversary's number is revealed and you pay the squared difference $(x_t - y_t)^2$

- Solve it using OGD w/ $\nabla \ell_t(x) = 2(x - y_t), V = [0,1]$
- W/ the optimal learning rate, the regret would be $O(\sqrt{T})$
- worse than previous $O(\ln T)$

the OLO reduction might not always give the best possible regret

$\Omega\left(\sqrt{T}\right)$ Lower Bounds for OLO

- Theorem 5.1. $\emptyset \neq V \subseteq \mathbb{R}^d$ closed convex set with diameter D.
  $\mathcal{A}$ is any (possibly randomized) algorithm for OLO on V.
  $T > 0$ is any integer.
  $\Rightarrow \exists g_1, \dots, g_T \in \mathbb{R}^d$ with $\|g_t\|_2 \leq L$ and $u \in V$ s.t. the regret of algorithm $\mathcal{A}$ satisfies

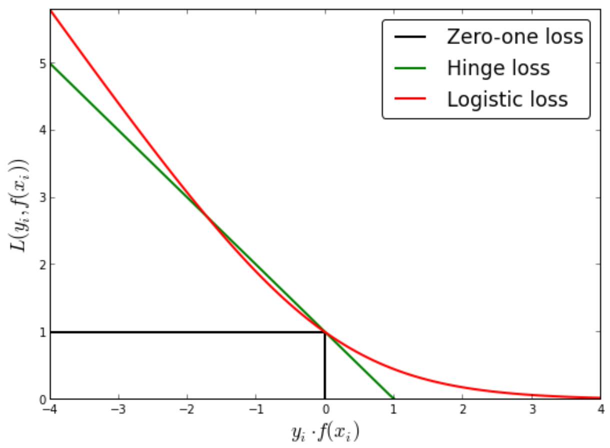
$$\mathrm{Regret}_T(u) = \sum_{t=1}^{T} \langle g_t, x_t \rangle - \sum_{t=1}^{T} \langle g_t, u \rangle \geq \frac{LD}{2}\sqrt{T} = \Omega(\sqrt{T})$$

Online-to-Batch Conversion

- Theorem 3.1. $f: \mathbb{R}^d \times X \to (-\infty, +\infty]$ is convex in the first argument. $F(x) = \mathbb{E}_{\xi\sim\rho}[f(x,\xi)]$. $\xi_1, \dots, \xi_T \overset{i.i.d.}{\sim} \rho$. $\ell_t(x) = \alpha_t f(x, \xi_t)$ w/ $\alpha_t > 0$. Run any OCO algorithm over the losses ℓ_t to construct the sequence of predictions $x_1, \dots, x_T$. Then

$$\mathbb{E}_{\xi_1,\dots,\xi_T}\left[F\left(\frac{1}{\sum_{t=1}^T \alpha_t}\sum_{t=1}^T \alpha_t x_t\right)\right] \le F(u) + \frac{\mathbb{E}[\mathrm{Regret}_T(u)]}{\sum_{t=1}^T \alpha_t}, \forall u \in \mathbb{R}^d$$

Example: Binary Classification



- Example 3.2. Input $z_i \in \mathbb{R}^d$ with norm $\leq R$, output $y_i \in \{-1,1\}$.
- Hinge loss $f\big(x,(z,y)\big) := \max(1 - y\langle z, x\rangle, 0)$. The objective is to
$$\min_x F(x) := \frac{1}{N}\sum_{i=1}^{N} \max(1 - y_i\langle z_i, x\rangle, 0)$$
- In each iteration, sample a training point uniformly
$$\ell_t(x) = \max(1 - y_t\langle z_t, x\rangle, 0)$$
- Run OSD to get $(x_1 = 0), x_2, \ldots, x_T$ w/ $\eta = \frac{1}{R\sqrt{T}}$

constant

- $\Rightarrow \mathbb{E}\left[F\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right)\right] - F(x^*) \leq R\,\frac{\|x^*\|_2^2+1}{\sqrt{T}}$

Example: Binary Classification (cont.)

- Example 3.3. Input $z_i \in \mathbb{R}^d$ with norm $\leq R$, output $y_i \in \{-1,1\}$.
- Hinge loss $f\big(x,(z,y)\big) := \max(1 - y\langle z, x\rangle, 0)$. The objective is to

$$\min_x F(x) := \frac{1}{N}\sum_{i=1}^{N} \max(1 - y_i\langle z_i, x\rangle, 0)$$

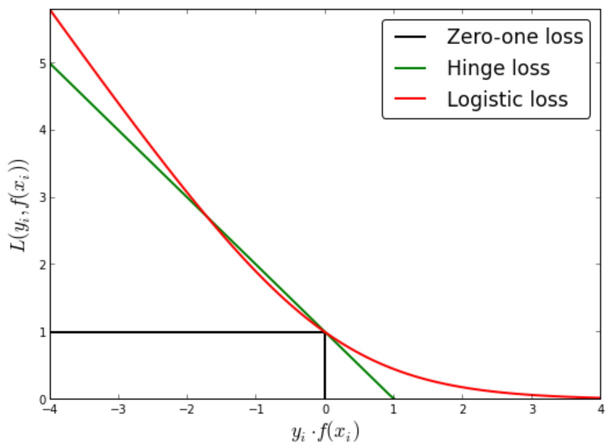
- In each iteration, sample a training point uniformly

$$\ell_t(x) = \frac{1}{R\sqrt{t}} \max(1 - y_t\langle z_t, x\rangle, 0)$$

varying learning rate

- Run OSD to get $(x_1 = 0), x_2, \ldots, x_T$ w/ $\eta = 1$
- $\Rightarrow \mathbb{E}\left[F\left(\frac{1}{\sum_{t=1}^{T}\frac{1}{R\sqrt{t}}}\sum_{t=1}^{T}\frac{1}{R\sqrt{t}}x_t\right)\right] - F(x^*) \leq \frac{1}{\sum_{t=1}^{T}\frac{1}{R\sqrt{t}}}\left(\frac{\|x^*\|_2^2}{2} + \frac{1}{2}\sum_{t=1}^{T}\|g_t\|_2^2\right)$

Example: Binary Classification 2



- Example 3.4. Input $z_i \in \mathbb{R}^d$ with norm $\leq R$, output $y_i \in \{-1,1\}$.
- Logistic loss $f\big(x,(z,y)\big) \coloneqq \ln(1+\exp(-y\langle z,x\rangle))$. The objective is to

$$\min_x F(x) \coloneqq \frac{1}{N}\sum_{i=1}^{N} \ln(1+\exp(-y_i\langle z_i,x\rangle))$$

Suppose the training set is linear separable. The minimizer of F would be infinite

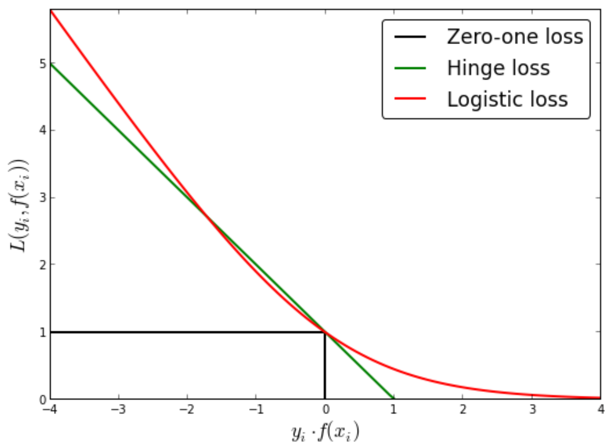
- In each iteration, sample a training point uniformly

$$\ell_t(x) = \ln(1+\exp(-y_t\langle z_t,x\rangle))$$

- Run OSD to get $(x_1 = 0), x_2, \ldots, x_T$ w/ $\eta = \frac{1}{R\sqrt{T}}$
- $\Rightarrow \mathbb{E}\left[F\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right)\right] \leq \frac{R}{2\sqrt{T}} + \min_{u\in\mathbb{R}^d} F(u) + R\frac{\|u\|_2^2}{2\sqrt{T}}$

Online learning automatically converges to the regularized value

Example: Binary Classification 3



- Example 3.13. Input $z_i \in \mathbb{R}^d$ with norm $\leq R$, output $y_i \in \{-1,1\}$.
- Hinge loss $f\big(x,(z,y)\big) \coloneqq \max(1 - y\langle z, x\rangle, 0)$. The objective is to

$$\min_{x \in \mathbb{R}^d} \mathrm{Risk}(x) \coloneqq \mathbb{E}_{(z,y)\sim\rho}[\max(1 - y\langle z, x\rangle, 0)]$$

- Draw T samples i.i.d. $\sim\rho$ w/ $\ell_t(x) = \max(1 - y_t\langle z_t, x\rangle, 0)$
- Run OSD to get $(x_1 = 0), x_2, \ldots, x_T$

High-probability result

convex

concentration

- $\mathrm{Risk}\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) \leq \frac{1}{T}\sum_{t=1}^{T} \mathrm{Risk}(x_t) \leq \frac{1}{T}\sum_{t=1}^{T} \ell_t(x_t) + \sqrt{\frac{2\ln(2/\delta)}{T}}$
- $\leq \frac{1}{T}\mathrm{Regret}_T(u) + \frac{1}{T}\sum_{t=1}^{T} \ell_t(u) + \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{T}} \leq \frac{1}{T}\mathrm{Regret}_T(u) + \mathrm{Risk}(u) + 2\sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{T}}$

Regret

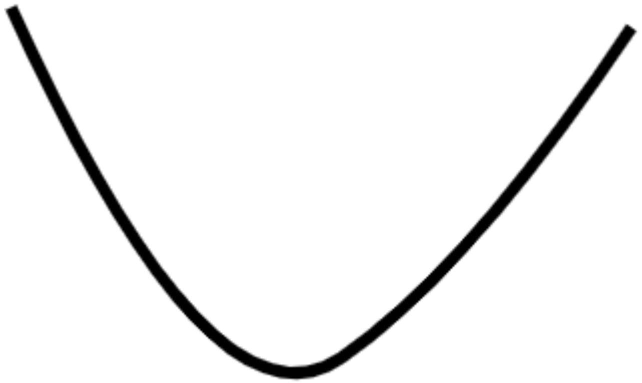
Concentration

Strong Convexity

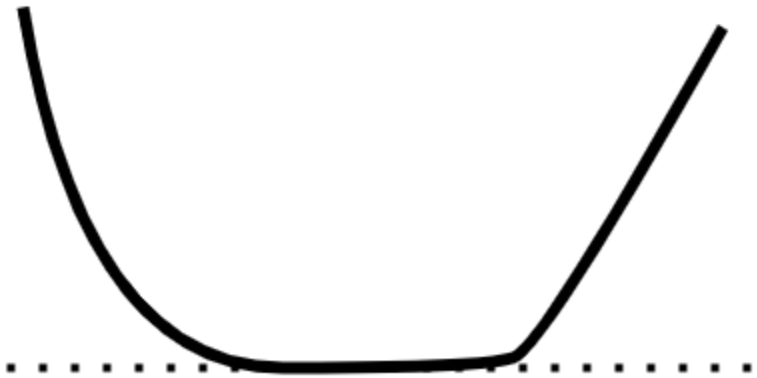
Strong Convexity

- Definition 4.1. A proper function $f: \mathbb{R}^d \to (-\infty, +\infty]$ is $\mu(\geq 0)$-strongly convex over a convex set $V \subseteq \operatorname{int} \operatorname{dom}(f)$ w.r.t. $\|\cdot\|$ if

$$f(y) \geq f(x) + \langle g, y - x \rangle + \frac{\mu}{2}\|x - y\|^2, \qquad \forall y \in \mathbb{R}^d$$



strongly convex



convex (but not strongly convex)

$O(\ln T)$ Regret of OSD for Strongly Convex Losses

- Theorem 4.7. $\emptyset \neq V \subseteq \mathbb{R}^d$ closed convex set.
  $\ell_t: \mathbb{R}^d \to (-\infty, +\infty]$ μ-strongly convex w.r.t. $\|\cdot\|_2$ over $V \subseteq \bigcap_{t=1}^{T} \operatorname{int} \operatorname{dom}(\ell_t)$.
  $x_1 \in V$. $\eta_t = \frac{1}{\mu t}$, $\|g_t\|_2 \leq L$. Then $\forall u \in V$,

$\eta_t = \frac{1}{\sum_{i=1}^{t} \mu_i}$

$$\sum_{t=1}^{T} (\ell_t(x_t) - \ell_t(u)) \leq O\left(\frac{L^2}{\mu} \ln T\right)$$

Online-to-batch Conversion

- Example 4.12. Input $z_i \in \mathbb{R}^d$ with norm $\leq R$, output $y_i \in \{-1,1\}$.
- Classic SVM:

λ-strongly convex

$$\min_x F(x) := \frac{\lambda}{2}\|x\|_2^2 + \frac{1}{N}\sum_{i=1}^{N} \max(1 - y_i\langle z_i, x\rangle, 0)$$

- $\operatorname{argmin} F \in B\left(0; \sqrt{\frac{1}{\lambda}}\right) =: V$
- $\ell_t(x) = \frac{\lambda}{2}\|x\|_2^2 + \max(1 - y_t\langle z_t, x\rangle, 0)$ w/ $\eta_t = \frac{1}{\lambda t}$

$$\mathbb{E}\left[F\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right)\right] - \min_x F(x) \leq O\left(\frac{R^2 \ln T}{\lambda T}\right)$$

Online-to-batch Conversion (cont.)

- Example 4.12. Input $z_i \in \mathbb{R}^d$ with norm $\le R$, output $y_i \in \{-1,1\}$.
- Classic SVM:

λ-strongly convex

$$\min_x F(x) := \frac{\lambda}{2}\|x\|_2^2 + \frac{1}{N}\sum_{i=1}^{N} \max(1 - y_i\langle z_i, x\rangle, 0)$$

- $\operatorname{argmin} F \in B\left(0; \sqrt{\frac{1}{\lambda}}\right) =: V$

λt-strongly convex

- $\ell_t(x) = \frac{\lambda t}{2}\|x\|_2^2 + t \max(1 - y_t\langle z_t, x\rangle, 0)$ w/ $\eta_t = \frac{2}{\lambda t(t+1)}$

$\ln T$ improvement

$$\mathbb{E}\left[F\left(\frac{1}{T(T+1)/2}\sum_{t=1}^{T} t x_t\right)\right] - \min_x F(x) \le O\left(\frac{R^2}{\lambda T}\right)$$

Summary

Shuai Li
https://shuaili8.github.io

- What is online learning
- Online gradient descent $O(\sqrt{T})$ Regret
- Online subgradient descent $O(\sqrt{T})$ Regret
  - Online-to-batch conversion
- Strongly convex losses $O(\ln T)$ regret

Questions?