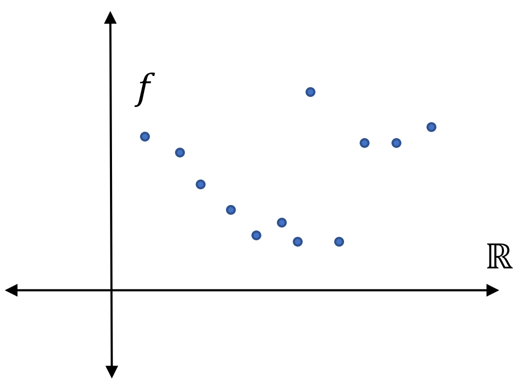
Introduction to Optimization Theory

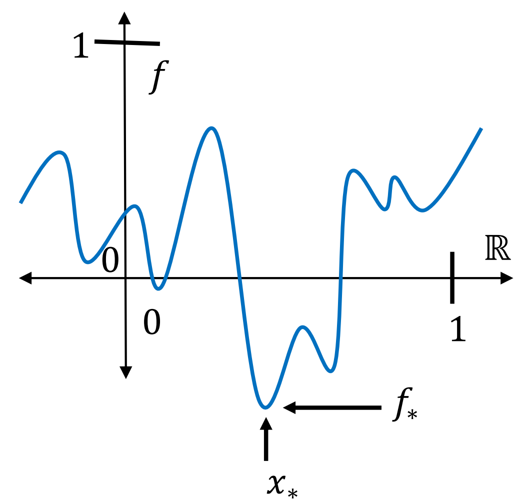
Lecture #15 - 11/5/20

MS&E 213 / CS 2690



Aaron Sidford

sidford@stanford.edu



Plan for Today

Recap

- Online linear optimization
- Follow the regularized leader (FTRL)

Applications

- Learning from experts
- Subgradient descent
- Smooth convex optimization

Mirror Descent

- An alternative algorithm

Stochastic Methods

More algorithms, apply, build

Online Linear Optimization

- Given closed convex $S \subseteq \mathbb{R}^n$
- Iterate for round $t \in [T]$
  - We pick $x_t \in S$
  - Adversary reveals penalty $p_t \in \mathbb{R}^n$

Goal:

- Minimize $\mathrm{regret}(T) \stackrel{\text{def}}{=} \sum_{t \in [T]} p_t^\top x_t - \min_{x \in s} \sum_{t \in [T]} p_t^\top x$
- Want average regret $\rightarrow 0$: $\lim_{T \rightarrow \infty} \frac{\mathrm{regret}(T)}{T} \rightarrow 0$

Methods

Online Linear Optimization

For $t \in [T]$
- We pick $x_t \in S$
- Adversary picks $p_t \in \mathbb{R}^n$

$$\text{regret}(T) \stackrel{\text{def}}{=} \sum_{t\in[T]} p_t^\top x_t - \min_{x\in S} \sum_{t\in[T]} p_t^\top x.$$

Learning from Experts

- $S = \Delta^n = \{x \in \mathbb{R}^n_{\geq 0} \mid \sum_{i\in[n]} x_i = 1\}$
- $p_t(i) \in [0,1]$ for all $i \in [n]$, $t \in [T]$

Some Ideas
- This is a "learning problem"
- Want to focus in on best expert

Problem:
- Past expert performance says nothing about future

Hope
- When past doesn't signal future ⇒ random is good
- Idea: concentrate on good solution slowly

~~*arbitrary*~~

~~*random*~~

~~*greedy*~~
~~*follow the leader (FTL)*~~

Follow the Regularized Leader (FTRL)

careful combination of greedy and random

Online Linear Optimization

- For $t \in [T]$
- We pick $x_t \in S$
- Adversary picks $p_t \in \mathbb{R}^n$
- $\text{regret}(T) \stackrel{\text{def}}{=} \sum_{t\in[T]} p_t^\top x_t - \min_{x\in S} \sum_{t\in[T]} p_t^\top x.$

Some "ideas"

- η is a "learning" rate
- $\sum_{t\in[T]} p_t^\top x$ measure greedy / past performance
- $r(x)$ encourages stability / "uniformity"

Algorithm

- Pick "regularizer" $r: S \to \mathbb{R}$ that is differentiable and μ-strongly convex with respect to $\|\cdot\|$
- Let $\Phi_T(x) \stackrel{\text{def}}{=} \eta \sum_{t\in[T]} p_t^\top x + r(x)$
- Let $\Phi_0(x) \stackrel{\text{def}}{=} r(x)$
- Let $x_{T+1} = \text{argmin}_{x\in S} \Phi_T(x)$

Analysis "Plan"

- Relate penalty to change in $\Phi_T(x_{T+1})$
- Leverage to relate penalty to OPT
- Use strong convexity and stability of Φ_T

Analysis

Online Linear Optimization

For $t \in [T]$
- We pick $x_t \in S$
- Adversary picks $p_t \in \mathbb{R}^n$

$\text{regret}(T) \stackrel{\text{def}}{=} \sum_{t\in[T]} p_t^\top x_t - \min_{x\in S} \sum_{t\in[T]} p_t^\top x.$

FTRL
- $\Phi_T(x) \stackrel{\text{def}}{=} \eta \sum_{t\in[T]} p_t^\top x + r(x)$
- Let $x_{T+1} = \text{argmin}_{x\in S} \Phi_T(x)$

Lemma #2: $\Phi_T(x_{T+1}) - \Phi_{T-1}(x_T) \geq \eta \cdot p_T^\top x_T - \frac{\eta^2}{2\mu} \|p_T\|_*^2$

Theorem (FTRL): For $L = \max_{t\in[T]} \|p_t\|_*^2$:

$$\sum_{t\in[T]} p_t^\top (x_t - z) \leq \frac{\eta}{2\mu} \cdot T \cdot L^2 + \frac{1}{\eta}\left[r(z) - \min_{x\in S} r(x)\right]$$

Corollary: If $D \geq \max_{x\in S} r(x) - \min_{x\in S} r(x)$ *and* $\eta = \sqrt{\frac{D\mu}{TL^2}}$ *then* $\frac{\text{regret}(T)}{T} \leq \sqrt{\frac{DL^2}{2\mu T}}$.

Proof:
- $\text{regret}(T) = \max_{z\in S} \sum_{t\in[T]} p_t^\top (x_t - z) = \sum_{t\in[T]} p_t^\top x_t - \min_{x\in S} \sum_{t\in[T]} p_t^\top x$

Notes
- Theorem says more than corollary!
- Better bounds when $r(z)$ smaller

Plan for Today

Recap ✓

- Online linear optimization
- Follow the regularized leader (FTRL)

Applications

- Learning from experts
- Subgradient descent
- Smooth convex optimization

Mirror Descent

- An alternative algorithm

Stochastic Methods

More algorithms, apply, build

Learning from Experts

Online Linear Optimization

- Given closed convex $S \subseteq \mathbb{R}^n$
- Iterate for round $t \in [T]$
  - We pick $x_t \in S$
  - Adversary reveals penalty $p_t \in \mathbb{R}^n$

Goal:

- $\text{regret}(T) \stackrel{\text{def}}{=} \sum_{t\in[T]} p_t^\top x_t - \min_{x\in S} \sum_{t\in[T]} p_t^\top x$
- Want average regret $\to 0$

$$\lim_{T\to\infty} \frac{\text{regret}(T)}{T} \to 0$$

What is needed to do as well as the experts?

Learning from Experts

- $S = \Delta^n = \left\{x \in \mathbb{R}^n_{\geq 0} \mid \sum_{i\in[n]} x_i = 1\right\}$
- $p_t(i) \in [0,1]$ for all $i \in [n]$, $t \in [T]$
- Note: $\min_{x\in S} a^T x = \min_{i\in[n]} a_i$

Story experts $1, \dots, n$

- Expert $i \in [n]$ makes prediction $[p_t]_i$ on day t
- Larger p_t means worse prediction (penalty)
- Each day we pick distribution on experts x_t
- At end of day p_t are revealed
- $p_t^\top x_t$ = expected penalty
- Regret = our expected penalty – penalty of best expert

Learning from Experts

- $L = \max_{t \in [T]} \|p_t\|_* :$
- $\sum_{t \in [T]} p_t^\top (x_t - z) \leq \frac{\eta}{2\mu} \cdot T \cdot L^2 + \frac{1}{\eta}\left[r(z) - \min_{x \in S} r(x)\right].$
- $D \geq \max_{x \in S} r(x) - \min_{x \in S} r(x)$ and $\eta \stackrel{\text{def}}{=} \sqrt{D\mu/(TL^2)}$
- $\text{regret}(T)/T \leq \sqrt{DL^2/(2\mu T)}$

Setup

- $S = \Delta^n = \left\{x \in \mathbb{R}^n_{\geq 0} | \vec{1}^\top x = 1\right\}$
- $p_t(i) \in [0,1]$

FTRL

- $\Phi_T(x) \stackrel{\text{def}}{=} \eta \sum_{t \in [T]} p_t^\top x + r(x)$
- Let $x_{T+1} = \text{argmin}_{x \in S} \Phi_T(x)$

Questions

- What regularizer?
- What norm?

Idea

- $\|p_t\|_\infty \leq 1$ and $\|x_t\|_1 = 1$
- Use ℓ_1 norm (dual ℓ_∞)

ℓ_1-Regularizer?

Want bounded strongly convex function with respect to ℓ_1 on simplex

Negative Entropy: $e: \mathbb{R}^n \to \mathbb{R}$ with $e(x) = \sum_{i \in [n]} x_i \ln x_i$

Lemma: e is 1-strongly convex with respect to $\|\cdot\|_1$ on Δ^n.

- $[\nabla e(x)]_i = 1 + \ln x_i$
- $[\nabla^2 e(x)] = \text{diag}\left(\frac{1}{x}\right)$
- $\|z\|_1 = \left(\sum_{i \in [n]} \frac{|z_i|}{\sqrt{x_i}} \sqrt{x_i}\ \right)^2$
- $\leq \left[\sum_{i \in [n]} \frac{z_i^2}{x_i}\right] \left[\sum_{i \in [n]} x_i\right] = z^\top \nabla^2 e(x) z$

after ignoring zero coordinates of x

ℓ_1-Regularizer?

Want bounded strongly convex function with respect to ℓ_1 on simplex

Negative Entropy: $e: \mathbb{R}^n \rightarrow \mathbb{R}$ with $e(x) = \sum_{i \in [n]} x_i \ln x_i$

Lemma: e is 1-strongly convex with respect to $\|\cdot\|_1$ on Δ^n.

Lemma: $z = \operatorname*{argmin}_{x \in \Delta^n} f(x) = p^\top x + e(x)$ has $\mathrm{z_i} = \frac{\exp(-p_i)}{\sum_{j \in [n]} \exp(-p_j)}$ for all $i \in [n]$.

- $z \in \Delta^n$
- Minimizer if and only if for all $x \in \Delta^n$, $\nabla f(z)^\top (x - z) \geq 0$
- $\nabla f(z) = \alpha \vec{1}$ for some α
- $\vec{1}^\top x = 1$ for all $x \in \Delta^n$

ℓ_1-Regularizer?

Want bounded strongly convex function with respect to ℓ_1 on simplex

Negative Entropy: $e: \mathbb{R}^n \to \mathbb{R}$ with $e(x) = \sum_{i \in [n]} x_i \ln x_i$

Lemma: e is 1-strongly convex with respect to $\| \cdot \|_1$ on Δ^n.

Lemma: $z = \operatorname{argmin}_{x \in \Delta^n} f(x) = p^\top x + e(x)$ has $z_i = \frac{\exp(-p_i)}{\sum_{j \in [n]} \exp(-p_j)}$ for all $i \in [n]$.

Lemma: $\max_{x \in \Delta^n} e(x) - \min_{x \in \Delta^n} e(x) \leq \ln n$

- $x \ln x$ convex with $1 \ln 1 = 0 \ln 0 = 0$
- $\max_{x \in \Delta^n} e(x) = 0$
- $\min_{x \in \Delta^n} e(x) = e\left(\frac{1}{n}\vec{1}\right) = -\ln n$

Learning from Experts

- $L = \max_{t\in[T]} \|p_t\|_*$:
- $\sum_{t\in[T]} p_t^\top (x_t - z) \le \frac{\eta}{2\mu} \cdot T \cdot L^2 + \frac{1}{\eta}\left[r(z) - \min_{x\in S} r(x)\right]$.
- $D \ge \max_{x\in S} r(x) - \min_{x\in S} r(x)$ and $\eta \overset{\text{def}}{=} \sqrt{D\mu/(TL^2)}$
- $\text{regret}(T)/T \le \sqrt{DL^2/(2\mu T)}$

Setup

- $S = \Delta^n = \left\{x \in \mathbb{R}^n_{\ge 0} \middle| \vec{1}^\top x = 1\right\}$
- $p_t(i) \in [0,1]$

Exponentiated gradient descent

Multiplicative Weights

- Let $w_0 = \vec{1} \in \mathbb{R}^n$
- For $t \in [T]$
  - $x_t(i) = \frac{w_i}{\|w\|_1}$
  - p_t revealed
  - $[w_t]_i = [w_{t-1}]_i \cdot \exp[-\eta \cdot [p_t]_i]$

FTRL

- $\Phi_T(x) \overset{\text{def}}{=} \eta \sum_{t\in[T]} p_t^\top x + r(x)$
- Let $x_{T+1} = \text{argmin}_{x\in S} \Phi_T(x)$

Note

- $x_t(i) = \dfrac{\exp\left(-\eta \sum_{k\in[t]} [p_k]_i\right)}{\sum_{j\in[n]} \exp\left(-\eta \sum_{k\in[t]} [p_k]_j\right)}$

Corollary: for $\eta = \sqrt{\frac{2 \lg n}{T}}$ have

$$\frac{\text{regret}(T)}{T} \le \sqrt{\frac{2 \lg n}{T}}$$

- $L = 1$
- $D = \ln n$
- $\mu = 1$

Plan for Today

Recap

- Online linear optimization
- Follow the regularized leader (FTRL)

Applications

- Learning from experts
- Subgradient descent
- Smooth convex optimization

Mirror Descent

- An alternative algorithm

Stochastic Methods

More algorithms, apply, build

Subgradient Oracle

Subgradient Descent

If $x_{t+1} = x_t - \eta g_t$ where $g_t \in \partial f(x_t)$ and $\|g_t\|_2 \leq L$ then for proper choice of η

$$f\left(\frac{1}{T}\sum_{t\in[T]} x_t\right) - f_* \leq \sqrt{\frac{L\|x_1 - x_*\|_2^2}{T}}.$$

query $x \in \mathbb{R}^n$ → **subgradient oracle** → output $g \in \partial f(x)$

Question
How prove / derive with FTRL?

***Subgradient**: g is subgradient of f at x (i.e. $g \in \partial f(x)$) if and only if $f(y) \geq f(x) + g^\top (y - x)$ for all $y \in \mathbb{R}^n$*

Assumption #1

- f is convex
- $\Leftrightarrow \partial f(x) \neq \emptyset$ for all $x \in \mathbb{R}^n$

Assumption #2

- f is L-Lipschitz with respect to $\|\cdot\|$
- $\Leftrightarrow$ $\|g\|_* \leq L$ for all $g \in \partial f(x)$ and $x \in \mathbb{R}^n$

FTRL Optimization

- $L = \max_{t \in [T]} \|p_t\|_* :$
- $\sum_{t \in [T]} p_t^\top (x_t - z) \le \frac{\eta}{2\mu} \cdot T \cdot L^2 + \frac{1}{\eta}\left[r(z) - \min_{x \in S} r(x)\right].$
- $D \ge \max_{x \in S} r(x) - \min_{x \in S} r(x)$ and $\eta \stackrel{\text{def}}{=} \sqrt{D\mu/(TL^2)}$
- $\text{regret}(T)/T \le \sqrt{DL^2/(2\mu T)}$

Suppose

- $f: \mathbb{R}^n \to \mathbb{R}$ convex
- $p_t \in \partial f(x_t)$ with $\|p_t\|_2 \le L$
- $x_* \in \operatorname{argmin}_{x \in \mathbb{R}^n} f(x)$ and $\|x_0 - x_*\|_2 \le R$

Analysis

- $f(x_*) \ge f(x_t) + p_t^\top (x_* - x_t)$
- $f(x_t) - f_* \le p_t^\top (x_t - x_*)$

Analysis

- $f\left(\frac{1}{T}\sum_{t \in [T]} x_t\right) - f_*$
- $\le \frac{1}{T}\sum_{t \in [T]} f(x_t) - f_*$ *Jensen's inequality*
- $\le \frac{1}{T}\sum_{t \in [T]} p_t^\top (x_t - x_*)$
- $\le \frac{1}{T}\left[\sum_{t \in [T]} p_t^\top x_t - \min_{x \in S} \sum_{t \in [T]} p_t^\top x\right]$
- $= \frac{\text{regret}(T)}{T}$

How apply?

Pick a regularizer?

Problem
If r strongly convex and $S = \mathbb{R}^n$ then $D = \infty$!

FTRL Optimization

- $L = \max_{t\in[T]} \|p_t\|_*$:
- $\sum_{t\in[T]} p_t^\top (x_t - z) \le \frac{\eta}{2\mu} \cdot T \cdot L^2 + \frac{1}{\eta}\left[r(z) - \min_{x\in S} r(x)\right]$.
- $D \ge \max_{x\in S} r(x) - \min_{x\in S} r(x)$ and $\eta \stackrel{\text{def}}{=} \sqrt{D\mu/(TL^2)}$
- $\text{regret}(T)/T \le \sqrt{DL^2/(2\mu T)}$

Suppose

- $f: \mathbb{R}^n \to \mathbb{R}$ convex
- $p_t \in \partial f(x_t)$ with $\|p_t\|_2 \le L$
- $x_* \in \operatorname{argmin}_{x\in\mathbb{R}^n} f(x)$ and $\|x_0 - x_*\|_2 \le R$

FTRL Algorithm

- $r(x) = \frac{1}{2}\|x - x_0\|_2^2$
- $x_{t+1} = \operatorname{argmin}_{x\in\mathbb{R}^n} \eta \sum_{k\in[t]} p_k^t x + r(x)$
- $x_{t+1} = x_0 - \sum_{k\in[t]} \eta p_k = x_t - \eta p_t$

Subgradient Descent

Generalizes to constraints and other norms!

ϵ-optimal with R^2L^2/ϵ^2 queries!

Analysis

- $f\left(\frac{1}{T}\sum_{t\in[T]} x_t\right) - f_*$
- $\le \frac{1}{T}\sum_{t\in[T]} f(x_t) - f_*$
- $\le \frac{1}{T}\sum_{t\in[T]} p_t^\top (x_t - x_*)$
- $\le \sum_{t\in[T]} p_t^\top x_t - \min_{x\in S} \sum_{t\in[T]} p_t^\top x$
- $= \frac{\text{regret}(T)}{T}$
- $\le \frac{\eta}{2\mu} \cdot L^2 + \frac{1}{\eta T}\left[r(x_*) - \min_{x\in S} r(x)\right]$
- $\le \frac{\eta}{2} \cdot L^2 + \frac{1}{2\eta T} \cdot R^2$
- $\eta = \sqrt{R^2/(TL^2)} \Rightarrow \le \sqrt{R^2L^2/T}$

Plan for Today

Recap

- Online linear optmization
- Follow the regularized leader (FTRL)

Applications

- Learning from experts
- Subgradient descent
- Smooth convex optimization

Mirror Descent

- An alternative algorithm

Stochastic Methods

More algorithms, apply, build

Idea: use full Theorem

FTRL for Smooth Convex Functions

Suppose

- $f: \mathbb{R}^n \to \mathbb{R}$ convex and L-smooth
- $x_* \in \operatorname{argmin}_{x \in \mathbb{R}^n} f(x)$

ϵ-optimal point with $2L\|x_0 - x_*\|_2^2/\epsilon^2$ gradient queries

FTRL Algorithm

- $p_t = \nabla f(x_t)$
- $r(x) = \frac{1}{2}\|x - x_0\|_2^2$
- $x_{t+1} = \operatorname{argmin}_{x \in \mathbb{R}^n} \eta \sum_{k \in [t]} p_k^t x + r(x)$
- $x_{t+1} = x_0 - \sum_{k \in [t]} \eta p_k = x_t - \eta p_t$
- $\eta = \frac{1}{2L} \Rightarrow f\left(\frac{1}{T}\sum_{t \in [T]} x_t\right) - f_* \leq \frac{2L}{T} \cdot \|x_0 - x_*\|_2^2$

Analysis

Different "best-case scenario" than gradient descent.

- $f\left(\frac{1}{T}\sum_{t \in [T]} x_t\right) - f_*$
- $\leq \frac{1}{T}\sum_{t \in [T]} f(x_t) - f_*$
- $\leq \frac{1}{T}\sum_{t \in [T]} p_t^\top (x_t - x_*)$
- $\leq \frac{\eta}{2\mu} \cdot L^2 + \frac{1}{\eta T}\left[r(z) - \min_{x \in S} r(x)\right]$
- $\leq \frac{\eta}{2\mu T}\sum_{t \in [T]}\|p_t\|_*^2 + \frac{1}{\eta T}\left[r(z) - \min_{x \in S} r(x)\right]$
- $\leq \frac{\eta L}{T}\sum_{t \in [T]}[f(x_t) - f_*] + \frac{1}{\eta}\left[r(z) - \min_{x \in S} r(x)\right]$
- $\Rightarrow \left(\frac{1 - \eta L}{T}\right)\sum_{t \in [T]}[f(x_t) - f_*] \leq \frac{1}{2\eta T}\|x_0 - x_*\|_2^2$

Plan for Today

Recap ✓
- Online linear optimization
- Follow the regularized leader (FTRL)

Applications ✓
- Learning from experts ✓
- Subgradient descent ✓
- Smooth convex optimization ✓

Mirror Descent
- An alternative algorithm

Stochastic Methods

More algorithms, apply, build

Another Online Learning Algorithm

<u>Idea</u>

- Start at x_0
- For $t \in [T]$
- Take small step in "direction" -p_t

<u>How Design Step?</u>

- Pick strongly convex r
- Center regularizer around x_t
  - Notation: $D(x||x_t)$
- Minimizer $\eta \cdot p_t^\top x + D(x||x_t)$

Mirror Descent

<u>Intuition</u>

- *Similar to before*
- x_t *do not move to much relative to reference point*

Centered Regularizer?

Bregman Divergence Distance

- Let r be differentiable and μ-strongly with respect to some norm.
- $D_r(x||c) = r(x) - [r(c) + \nabla r(c)^\top (x - c)]$

Notes

- $D_r(x||c) \geq \frac{\mu}{2} \|x - c\|^2$
- $D_r(x||c) = 0 \Leftrightarrow x = c$
- $D_r(x||y)$ not necessarily $D_r(y||x)$

Examples

- $r(x) = \frac{1}{2} \|x - d\|_2^2$
  - $D_r(x||c) = \frac{1}{2} \|x - c\|_2^2$
  - Doesn't depend on d
- $r(x) = \sum_{i \in [n]} x_i \log x_i$
  - $D_r(x||c) = \sum_{i \in [n]} x_i \log(x_i / c_i)$
  - KL-divergence

Bregman Projection

Bregman Divergence Distance

- Let r be differentiable and μ-strongly with respect to some norm.
- $D_r(x||c) = r(x) - [r(c) + \nabla r(c)^\top (x - c)]$

Bregman Projection

- For closed convex $S \subseteq \mathbb{R}^n$ let $\pi_S^r(y) = \underset{x \in S}{\operatorname{argmin}}\, D_r(x||y)$

Bregman Pythagorean Theorem (*obtuse angles*)

- $D_r(x||y) + D_r(y||z) \leq D_r(x||z)$ for all $x \in S$
- $\Leftrightarrow y = \pi_S^r(z)$

Proof

- Algebra and optimality of projection

Mirror Descent Analysis

- $D_r(x||c) = r(x) - [r(c) + \nabla r(c)^\top (x - y)]$
- $x_{t+1} = \underset{x \in S}{\operatorname{argmin}}\, \eta \cdot p_t^\top x + D_r(x||x_t)$

Lemma

- $D_r(z||x_{t+1}) - D_r(z||x_t) \leq \frac{\eta^2}{2\mu} \|p_t\|_*^2 - \eta p_t^\top (x_t - z)$

Theorem: Mirror Descent

- $\sum_{t \in [T]} p_t^\top (x_t - z) \leq \sum_{t \in [T]} \frac{\eta}{2\mu} \|p_t\|_*^2 + \frac{1}{\eta} D_r(z||x_0)$

- *Similar bound!*
- *Resulting methods are often the same!*

Theorem: Dual Averaging (FTRL)

- $\sum_{t \in [T]} p_t^\top (x_t - z) \leq \sum_{t \in [T]} \frac{\eta}{2\mu} \|p_t\|_*^2 + \frac{1}{\eta} \left[r(z) - \min_{x \in S} r(x) \right]$

Plan for Today

Recap

- Online linear optimization
- Follow the regularized leader (FTRL)

Applications

- Learning from experts
- Subgradient descent
- Smooth convex optimization

Mirror Descent

- An alternative algorithm

Stochastic Methods

More algorithms, apply, build