Choosing the Metric in Subgradient Methods

John Duchi

Outline

I Mirror descent methods
1. Bregman divergence
2. Motivation via gradient method
3. Convergence proof
4. Example

II Adaptive metric methods
1. Motivation
2. Examples
3. Convergence guarantees (overview)

Motivation

Consider usual problem

$$\text{minimize } f(x) \quad \text{subject to } x \in C \subset \mathbb{R}^n.$$

Assume that n is very large (high-dimensional). Then

- Norm of gradient scales as $\|\nabla f(x)\|_2 = \sqrt{\sum_{i=1}^n [\nabla f(x)]_i^2} \approx \sqrt{n}$
- Can we do better?

Bregman divergences

Let $h : C \to \mathbb{R}$ be a differentiable convex function. The *Bregman divergence* associated with h is

$$D_h(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y \rangle$$

Mirror descent (non-Euclidean gradient descent)

- Compute subgradient $g_k \in \partial f(x_k)$
- Update

$$x_{k+1} = \operatorname*{argmin}_{x \in C} \left\{ \langle g_k, x \rangle + \frac{1}{\alpha_k} D_h(x, x_k) \right\}$$

Convergence analysis

Main assumption (recall homework): $h : C \to \mathbb{R}$ is strongly convex with respect to some norm $\|\cdot\|$ on C,

$$h(y) \geq h(x) + \langle \nabla h(x), y - x \rangle + \frac{1}{2} \|x - y\|^2$$

Not strictly necessary assumption: divergence is upper bounded,

$$D_h(x^\star, x) \leq R^2$$

for all $x \in C$ (or that stepsize α is constant)

Dual norms

Recall *dual norm*

$$\|y\|_* = \sup_{x:\|x\|\le 1} \langle x, y \rangle$$

which satisfies $\|x\| = \sup_{y:\|y\|_*\le 1} \langle x, y \rangle$ (in finite dimensions)

Convergence analysis

Progress of a single update:

$$f(x_k) - f(x^\star) \le \langle g_k, x_k - x^\star \rangle$$

Convergence analysis II

Single update progress:

$$f(x_k) - f(x^\star) \le \frac{1}{\alpha_k}\left[D_h(x^\star, x_k) - D_h(x^\star, x_{k+1}) - D_h(x_{k+1}, x_k)\right] + \langle g_k, x_{k+1} - x_k \rangle$$

Convergence analysis III

Telescope the sum

$$\sum_{k=1}^{K}[f(x_k) - f(x^\star)] \le \sum_{k=1}^{K} \frac{1}{\alpha_k} [D_h(x^\star, x_k) - D_h(x^\star, x_{k+1})] + \sum_{k=1}^{K} \frac{\alpha_k}{2} \|g_k\|_*^2$$

Convergence guarantee

with fixed stepsize $\alpha_k = \alpha$,

$$\frac{1}{K}\sum_{k=1}^{K}[f(x_k) - f(x^\star)] \le \frac{1}{\alpha K} D_h(x^\star, x_1) + \frac{\alpha}{2K} M^2$$

where we assume $\|g_k\|_* \le M$ for all k
In general, convergence if

- $D_h(x^\star, x_1) < \infty$
- $\sum_k \alpha_k = \infty$ but $\alpha_k \to 0$
- subgradients are bounded, i.e. $\|g\|_* \le M$ for $g \in \partial f(x)$ where $x \in C$

Example: entropic mirror descent

Suppose we wish to solve problem over probability simplex,

$$C = \{x \in \mathbb{R}^n_+ : \langle \mathbf{1}, x \rangle = 1\}.$$

Use negative entropy

$$h(x) = \sum_{i=1}^n x_i \log x_i$$

- Strongly convex with respect to ℓ_1-norm over simplex
- $D_h(x, y) = \sum_{i=1}^n x_i \log \frac{x_i}{y_i}$,

$$D_h(x, \mathbf{1}/n) \le \log n$$

- Need only $\|g\|_\infty \le M_\infty$

Entropic mirror descent update

Solve update for $C = \{x \in \mathbb{R}^n_+ : \langle \mathbf{1}, x \rangle = 1\}$

$$\operatorname*{argmin}_{x \in C} \{\langle g, x \rangle + D_h(x, y)\}.$$

Entropic mirror descent versus projected gradient descent

$$\min \; f(x) = \frac{1}{m} \|Ax - b\|_1 \quad \text{s.t.} \quad x \in C = \{x \in \mathbb{R}^n_+ : \langle \mathbf{1}, x \rangle = 1\}$$

where $A = [a_1 \; \cdots \; a_m]^\top \in \mathbb{R}^{m \times n}$.

Projected gradient

- $\|x_1 - x^\star\|_2^2 \le 1$
- $\|g\|_2 \approx \max_i \|a_i\|_2$

Convergence

$$f(x_K) - f(x^\star) \le \frac{\|a\|_2}{\sqrt{K}}$$

Mirror descent

- $D_h(x^\star, x_1) \le \log n$
- $\|g\|_\infty \approx \max_i \|a_i\|_\infty$

Convergence

$$f(x_K) - f(x^\star) \le \frac{\|a\|_\infty \sqrt{\log n}}{\sqrt{K}}.$$

Example

Robust regression problem (an LP):

$$\text{minimize } f(x) = \|Ax - b\|_1 = \sum_{i=1}^{m} |a_i^T x - b_i|$$
$$\text{subject to } x \in C = \{x \in \mathbb{R}^n_+ \mid \mathbf{1}^T x = 1\}$$

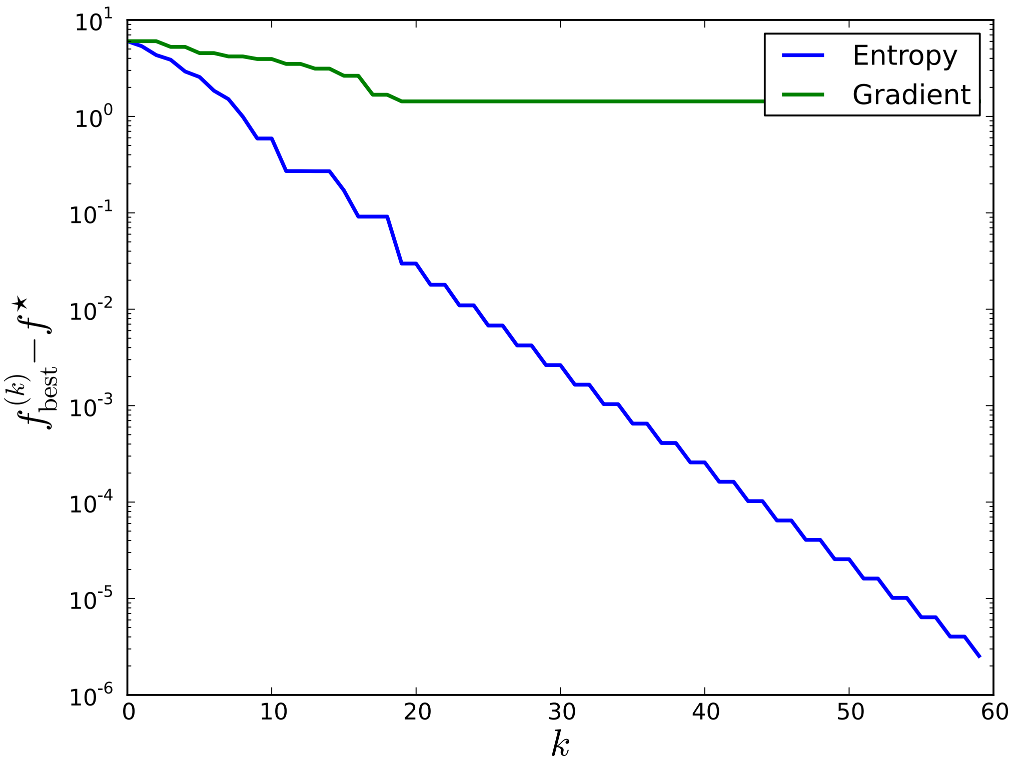
subgradient of objective is $g = \sum_{i=1}^m \operatorname{sign}(a_i^T x - b_i) a_i$

- Projected subgradient update ($h(x) = (1/2)\|x\|_2^2$): annoying
- Mirror descent update ($h(x) = \sum_{i=1}^n x_i \log x_i$):

$$x_i^{(k+1)} = \frac{x_i^{(k)} \exp(-\alpha g_i^{(k)})}{\sum_{j=1}^n x_j^{(k)} \exp(-\alpha g_j^{(k)})}$$

Example

Robust regression problem with $a_i \sim N(0, I_{n \times n})$ and $b_i = (a_{i,1} + a_{i,2})/2 + \varepsilon_i$ where $\varepsilon_i \sim N(0, 10^{-2})$, $m = 20, n = 3000$



stepsizes chosen according to best bounds (but still sensitive to stepsize choice)

Variable metric subgradient methods

Back to Euclidean case, use a metric based on matrix $H_k \succ 0$

(1) Get subgradient $g_k \in \partial f(x_k)$ (or stochastic subgradient with $\mathbb{E}[g_k] \in \partial f(x_k)$)

(2) update (often diagonal) matrix H_k

(3) update

$$x_{k+1} = \operatorname*{argmin}_{x \in C} \left\{ \langle g_k, x \rangle + \frac{1}{2}(x - x_k)^\top H_k (x - x_k) \right\}$$

So H_k generlizes stepsize and metric

Variable metric subgradient methods (projection)

Projected gradient variant (same procedure) with projection in H_k metric

(1) Get subgradient $g_k \in \partial f(x_k)$ (or stochastic subgradient with $\mathbb{E}[g_k] \in \partial f(x_k)$)

(2) update (often diagonal) matrix H_k

(3) update

$$x_{k+1} = \pi_C^{H_k}(x_k - H_k^{-1} g_k)$$

where

$$\pi_C^H(x) = \operatorname*{argmin}_{y \in C}\{\|y - x\|_H^2\}$$

and $\|x\|_H^2 = x^\top H x$

Convergence analysis

$$\frac{1}{2}\left\|x_{k+1} - x^\star\right\|_{H_k}^2$$

Convergence analysis II

$$f(x_k) - f(x^\star) \le \frac{1}{2}\left[\|x_k - x^\star\|_{H_k}^2 - \|x_{k+1} - x^\star\|_{H_k}^2\right] + \frac{1}{2}\|g_k\|_{H_k^{-1}}^2 .$$

Final guarantee (homework)

With choice $\overline{x}_K = \frac{1}{K}\sum_{k=1}^K x_k$,

$$f(\overline{x}_K) - f(x^\star) \le \frac{1}{2K}\left[\|x_1 - x^\star\|_{H_1}^2 + \sum_{k=1}^K \|g_k\|_{H_k^{-1}}^2\right] + \frac{1}{2K}\sum_{k=2}^K \left(\|x_k - x^\star\|_{H_k}^2 - \|x_k - x^\star\|_{H_{k-1}}^2\right).$$

- Convergence if differences $\|\cdot\|_{H_k}^2 - \|\cdot\|_{H_{k-1}}^2$ go to zero and $\sum_{k=1}^K \|g_k\|_{H_k^{-1}}^2$ grows slower than K

AdaGrad

AdaGrad — adaptive subgradient method

(1) get subgradient $g^{(k)} \in \partial f(x^{(k)})$

(2) choose metric H_k:

- set $S_k = \sum_{i=1}^{k} \mathrm{diag}(g_i)^2$
- set $H_k = \frac{1}{\alpha} S_k^{\frac{1}{2}}$

(3) update $x_{k+1} = \pi_C^{H_k}(x_k - H_k^{-1} g_k)$

where $\alpha > 0$ is step-size

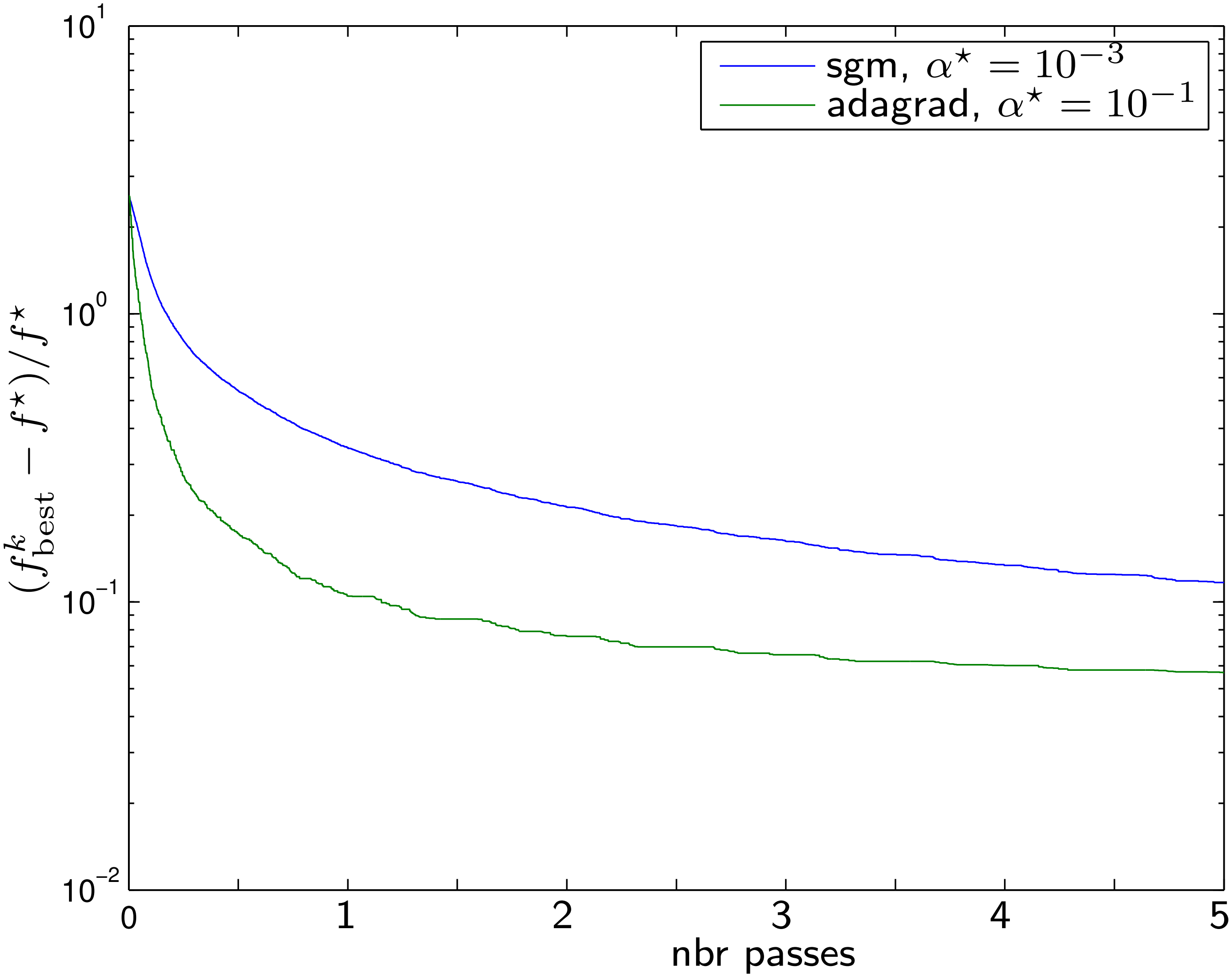
Convergence: homework!

Example

Classification problem:

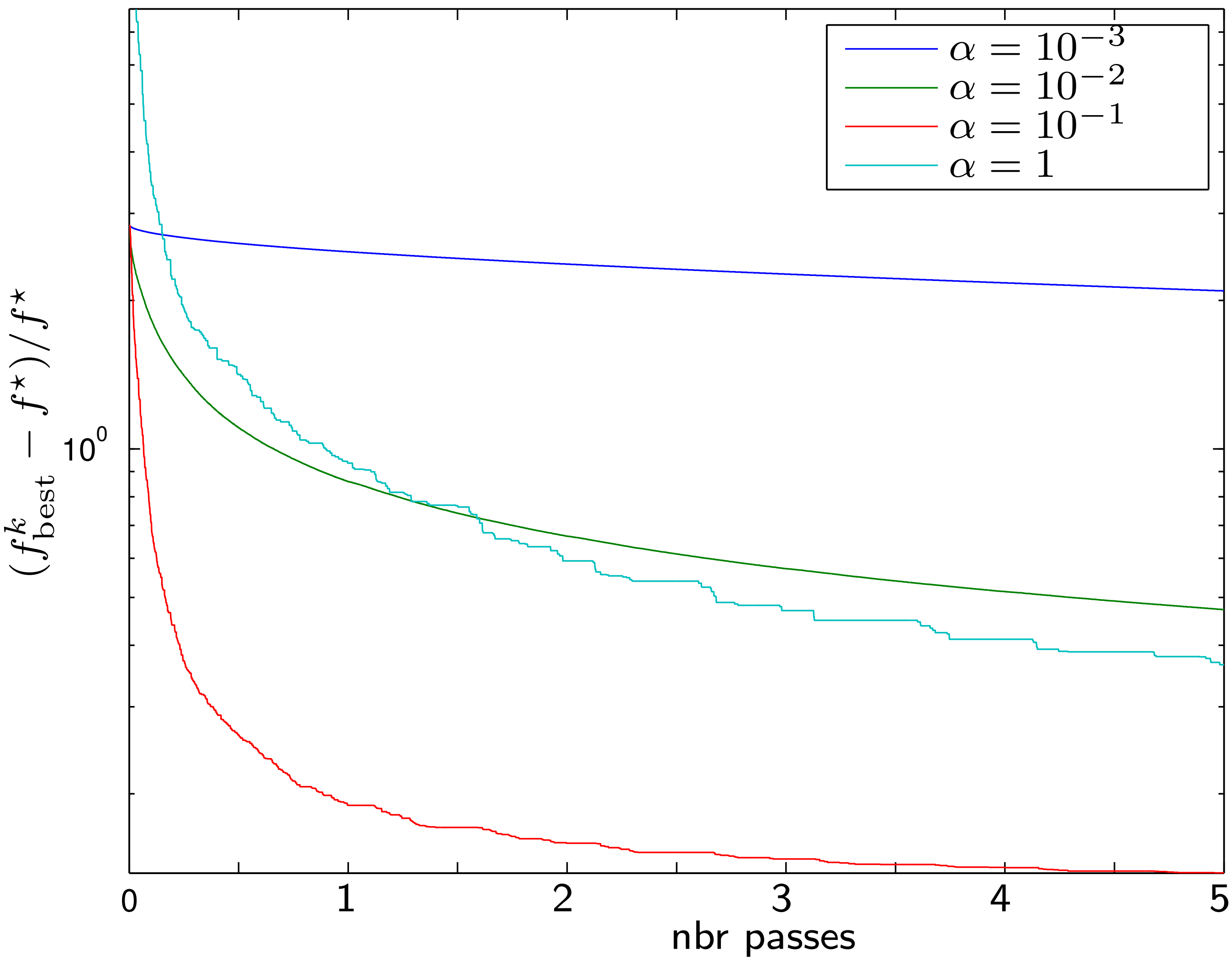
- **Data**: $\{a_i, b_i\}$, $i = 1, \ldots, 50000$
  - $a_i \in \mathbb{R}^{1000}$
  - $b \in \{-1, 1\}$
  - Data created with 5% mis-classifications w.r.t. $w = \mathbf{1}$, $v = 0$
- **Objective**: find classifiers $w \in \mathbb{R}^{1000}$ and $v \in \mathbb{R}$ such that
  - $a_i^\top w + v > 1$ if $b = 1$
  - $a_i^\top w + v < -1$ if $b = -1$
- **Optimization method**:
  - Minimize hinge-loss: $\sum_i \left[1 - b_i \langle a_i, w \rangle + v\right]_+$
  - Choose example uniformly at random, take sub-gradient step w.r.t. that example

Best subgradient method vs best AdaGrad



Often best AdaGrad performs better than best subgradient method

AdaGrad with different step-sizes α:



Sensitive to step-size selection (like standard subgradient method)