The choice of metric in subgradient methods

Stephen Boyd & John Duchi (with help from P. Giselsson)

EE364b, Stanford University

Mirror descent methods

subgradient method without using Euclidean steps

- let h be a differentiable convex function, then associated Bregman divergence is

$$D_h(y, x) = h(y) - h(x) - \nabla h(x)^T (y - x)$$

- mirror (or non-linear) subgradient method

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

  (2) update

$$x^{(k+1)} = \operatorname*{argmin}_{x \in C} \left\{ g^{(k)T} x + \frac{1}{\alpha_k} D_h(x, x^{(k)}) \right\}$$

generalizes projected subgradient decent (take $h(x) = \frac{1}{2} \|x\|_2^2$)

Convergence analysis

properties of h required: *strong convexity* with respect to norm $\|\cdot\|$

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

For any $x^\star \in C$,

$$\begin{aligned} f(x^{(k)}) - f(x^\star) &\leq g^{(k)T}(x^{(k)} - x^\star) \\ &= g^{(k)T}(x^{(k+1)} - x^\star) + g^{(k)T}(x^{(k)} - x^{(k+1)}) \end{aligned}$$

Use optimality conditions for $x^{(k+1)}$:

$$(\alpha_k g^{(k)} + \nabla h(x^{(k+1)}) - \nabla h(x^{(k)}))^T (y - x^{(k+1)}) \geq 0, \quad \text{all } y \in C$$

so (take $y = x^\star$)

$$g^{(k)T}(x^{(k+1)} - x^\star) \leq \frac{1}{\alpha_k} (\nabla h(x^{(k+1)}) - \nabla h(x^{(k)}))^T (x^\star - x^{(k+1)})$$

Convergence analysis continued

identity for divergences

$$(\nabla h(x^{(k+1)}) - \nabla h(x^{(k)}))^T (x^\star - x^{(k+1)}) \\ = D_h(x^\star, x^{(k)}) - D_h(x^\star, x^{(k+1)}) - D_h(x^{(k)}, x^{(k+1)})$$

for any $x^\star \in C$,

$$f(x^{(k)}) - f(x^\star) \leq g^{(k)T}(x^{(k+1)} - x^\star) + g^{(k)T}(x^{(k)} - x^{(k+1)}) \\ \leq \frac{1}{\alpha_k}\left[D_h(x^\star, x^{(k)}) - D_h(x^\star, x^{(k+1)})\right] - \frac{1}{\alpha_k} D_h(x^{(k)}, x^{(k+1)}) \\ + g^{(k)T}(x^{(k)} - x^{(k+1)})$$

apply Fenchel-Young inequality ($x^T y \leq \frac{1}{2\alpha}\|x\|^2 + \frac{\alpha}{2}\|y\|_*^2$)

$$\leq \frac{1}{\alpha_k}\left[D_h(x^\star, x^{(k)}) - D_h(x^\star, x^{(k+1)})\right] - \frac{1}{\alpha_k} D_h(x^{(k)}, x^{(k+1)}) \\ + \frac{\alpha_k \left\|g^{(k)}\right\|_*^2}{2} + \frac{1}{2\alpha_k}\left\|x^{(k)} - x^{(k+1)}\right\|^2 \\ \leq \frac{1}{\alpha_k}\left[D_h(x^\star, x^{(k)}) - D_h(x^\star, x^{(k+1)})\right] + \frac{\alpha_k}{2}\left\|g^{(k)}\right\|_*^2$$

Convergence guarantees

with fixed stepsize $\alpha_k = \alpha$,

$$\frac{1}{k}\sum_{i=1}^{k} f(x^{(i)}) - f(x^\star) \leq \frac{1}{\alpha k} D_h(x^\star, x^{(1)}) + \frac{\alpha}{2k} \max_i \left\| g^{(i)} \right\|_*^2$$

in general, converges if

- $D_h(x^\star, x^{(1)}) < \infty$
- $\sum_k \alpha_k = \infty$ and $\alpha_k \to 0$
- for all $g \in \partial f(x)$ and $x \in C$, $\|g\|_* \leq G$ for some $G < \infty$

Mirror descent examples

- Usual (projected) subgradient descent: $h(x) = \frac{1}{2} \|x\|_2^2$
- With constraints of simplex, $C = \{x \in \mathbf{R}^n_+ \mid \mathbf{1}^T x = 1\}$, use negative entropy

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

(1) Strongly convex with respect to ℓ_1-norm

(2) With $x^{(1)} = \mathbf{1}/n$, have $D_h(x^\star, x^{(1)}) \le \log n$ for $x^\star \in C$

(3) If $G_\infty \ge \|g\|_\infty$ for $g \in \partial f(x)$ for $x \in C$,

$$f_{\mathrm{best}}^{(k)} - f^\star \le \frac{\log n}{\alpha k} + \frac{\alpha}{2k} G_\infty$$

(4) Can be much better than regular subgradient decent...

Example

Robust regression problem (an LP):

$$\begin{array}{ll} \text{minimize} & f(x) = \|Ax - b\|_1 = \sum_{i=1}^{m} |a_i^T x - b_i| \\ \text{subject to} & x \in C = \{x \in \mathbf{R}_+^n \mid \mathbf{1}^T x = 1\} \end{array}$$

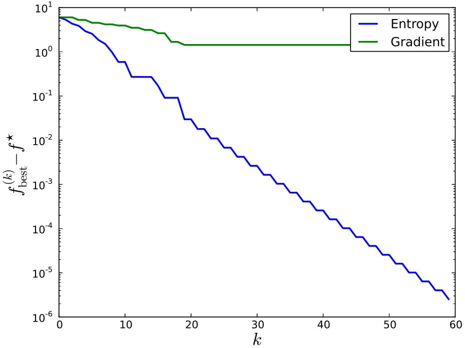
subgradient of objective is $g = \sum_{i=1}^{m} \operatorname{sign}(a_i^T x - b_i) a_i$

- Projected subgradient update ($h(x) = (1/2)\|x\|_2^2$): homework
- 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

subgradient method with variable metric $H_k \succ 0$:

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

(2) update (diagonal) metric H_k

(3) update $x^{(k+1)} = x^{(k)} - H_k^{-1} g^{(k)}$

- matrix H_k generalizes step-length α_k

there are many such methods (Ellipsoid method, AdaGrad, ...)

Variable metric projected subgradient method

same, with projection carried out in the H_k metric:

(1) get subgradient $g^{(k)} \in \partial f(x^{(k)})$
(2) update (diagonal) metric H_k
(3) update $x^{(k+1)} = P_{\mathcal{X}}^{H_k}\left(x^{(k)} - H_k^{-1} g^{(k)}\right)$

where

$$\Pi_{\mathcal{X}}^{H}(y) = \underset{x \in \mathcal{X}}{\operatorname{argmin}} \|x - y\|_H^2$$

and $\|x\|_H = \sqrt{x^T H x}$.

Convergence analysis

since $\Pi_{\mathcal{X}}^{H_k}$ is non-expansive in the $\|\cdot\|_{H_k}$ norm, we get

$$\begin{aligned}
\|x^{(k+1)} - x^\star\|_{H_k}^2 &= \left\| P_{\mathcal{X}}^{H_k}\left(x^{(k)} - H_k^{-1}g^{(k)}\right) - P_{\mathcal{X}}^{H_k}(x^\star)\right\|_{H_k}^2 \\
&\le \|x^{(k)} - H_k^{-1}g^{(k)} - x^\star\|_{H_k}^2 \\
&= \|x^{(k)} - x^\star\|_{H_k}^2 - 2(g^{(k)})^T(x^{(k)} - x^\star) + \|g^{(k)}\|_{H_k^{-1}}^2 \\
&\le \|x^{(k)} - x^\star\|_{H_k}^2 - 2(f(x^{(k)}) - f^\star) + \|g^{(k)}\|_{H_k^{-1}}^2.
\end{aligned}$$

using $f^\star = f(x^\star) \ge f(x^{(k)}) + g^{(k)T}(x^\star - x^{(k)})$

apply recursively, use

$$\sum_{i=1}^{k} \left(f(x^{(i)}) - f^\star \right) \geq k \left(f_{\text{best}}^{(k)} - f^\star \right)$$

and rearrange to get

$$\begin{aligned} f_{\text{best}}^{(k)} - f^\star \leq & \frac{\|x^{(1)} - x^\star\|_{H_1}^2 + \sum_{i=1}^{k} \|g^{(i)}\|_{H_i^{-1}}^2}{2k} \\ & + \frac{\sum_{i=2}^{k} \left(\|x^{(i)} - x^\star\|_{H_i}^2 - \|x^{(i)} - x^\star\|_{H_{i-1}}^2 \right)}{2k} \end{aligned}$$

numerator of additional term can be bounded to get estimates

- for general $H_k = \mathbf{diag}(h_k)$

$$f_{\rm best}^k - f^\star \le \frac{R_\infty^2 \|H_1\|_1 + \sum_{i=1}^k \|g^{(i)}\|_{H_i^{-1}}^2}{2k} + \frac{R_\infty^2 \sum_{i=2}^k \|H_i - H_{i-1}\|_1}{2k}$$

- for $H_k = \mathbf{diag}(h_k)$ with $h_i \ge h_{i-1}$ for all i

$$f_{\rm best}^k - f^\star \le \frac{\sum_{i=1}^k \|g^{(i)}\|_{H_i^{-1}}^2}{2k} + \frac{R_\infty^2 \|h_k\|_1}{2k}$$

where $\max_{1 \le i \le k} \|x^{(i)} - x^\star\|_\infty \le R_\infty$

converges if

- $R_\infty < \infty$ (e.g. if $\mathcal{X}$ is compact)
- $\sum_{i=1}^{k} \|g^{(i)}\|^2_{H_i^{-1}}$ grows slower than k
- $\sum_{i=2}^{k} \|H_i - H_{i-1}\|_1$ grows slower than k **or**
  $h_i \geq h_{i-1}$ for all i and $\|h_k\|_1$ 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 $S_k = \sum_{i=1}^{k} \mathbf{diag}(g^{(i)})^2$
- set $H_k = \frac{1}{\alpha} S_k^{\frac{1}{2}}$

(3) update $x^{(k+1)} = P_{\mathcal{X}}^{H_k} \left(x^{(k)} - H_k^{-1} g^{(k)} \right)$

where $\alpha > 0$ is step-size

AdaGrad – motivation

- for fixed $H_k = H$ we have estimate:

$$f_{\text{best}}^{(k)} - f^\star \le \frac{1}{2k}(x^{(1)} - x^\star)^T H (x^{(1)} - x^\star) + \frac{1}{2k}\sum_{i=1}^{k} \|g^{(i)}\|_{H^{-1}}^2$$

- **idea**: Choose *diagonal* $H_k \succ 0$ that minimizes this estimate in hindsight:

$$H_k = \operatorname*{argmin}_{h} \max_{x,y \in C} (x-y)^T \,\mathbf{diag}(h)(x-y) + \sum_{i=1}^{k} \|g^{(i)}\|_{\mathbf{diag}(h)^{-1}}^2$$

- optimal $H_k = \frac{1}{R_\infty}\,\mathbf{diag}\left(\sqrt{\sum_{i=1}^k (g_1^{(i)})^2}, \ldots, \sqrt{\sum_{i=1}^k (g_n^{(i)})^2}\right)$
- **intuition**: adapt step-length based on historical step lengths

AdaGrad – convergence

by construction, $H_i = \frac{1}{\alpha}\,\mathbf{diag}(h_i)$ and $h_i \geq h_{i-1}$, so

$$\begin{aligned} f_{\mathrm{best}}^{(k)} - f^\star &\leq \frac{1}{2k}\sum_{i=1}^{k} \|g^{(i)}\|_{H_i^{-1}}^2 + \frac{1}{2k\alpha} R_\infty^2 \left\|h_k\right\|_1 \\ &\leq \frac{\alpha}{k}\left\|h_k\right\|_1 + \frac{1}{2k\alpha} R_\infty^2 \left\|h_k\right\|_1 \end{aligned}$$

(second line is a theorem)

also have (with $\alpha = R_\infty^2$) and for compact sets C

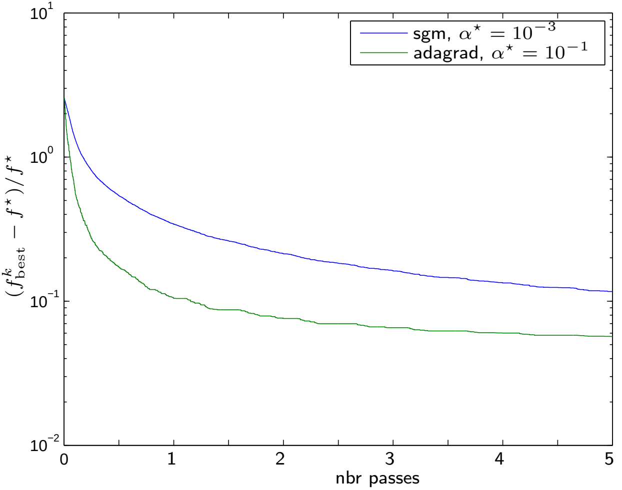
$$f_{\mathrm{best}}^{(k)} - f^\star \leq \frac{2}{k}\inf_{h \geq 0}\left\{ \sup_{x,y \in C} (x-y)^T\,\mathbf{diag}(h)(x-y) + \sum_{i=1}^{k} \|g^{(i)}\|_{\mathbf{diag}(h)^{-1}}^2 \right\}$$

Example

Classification problem:

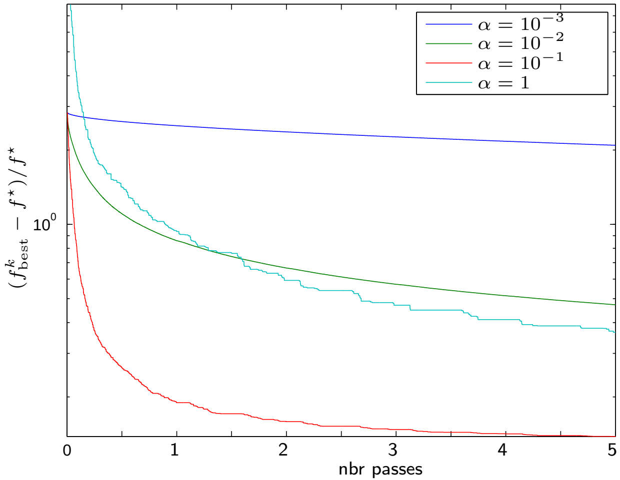
- **Data**: $\{a_i, b_i\}$, $i = 1, \ldots, 50000$
  - $a_i \in \mathbf{R}^{1000}$
  - $b \in \{-1, 1\}$
  - Data created with 5% mis-classifications w.r.t. $w = \mathbf{1}$, $v = 0$
- **Objective**: find classifiers $w \in \mathbf{R}^{1000}$ and $v \in \mathbf{R}$ such that
  - $a_i^T w + v > 1$ if $b = 1$
  - $a_i^T w + v < 1$ if $b = -1$
- **Optimization method**:
  - Minimize hinge-loss: $\sum_i \max(0, 1 - b_i(a_i^T w + v))$
  - 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)