Mirror Descent and Variable Metric Methods

Stephen Boyd & John Duchi & Mert Pilanci

EE364b, Stanford University

April 24, 2019

Mirror descent

- due to Nemirovski and Yudin (1983)
- recall the *projected subgradient* method:

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

  (2) update

$$x^{(k+1)} = \operatorname*{argmin}_{x \in C} \left\{ g^{(k)T} x + \frac{1}{2\alpha_k} \left\| x - x^{(k)} \right\|_2^2 \right\}$$

replace $\|\cdot\|_2^2$ with an alternate distance^2-like function

Convergence rate of projected subgradient method

Consider $\min_{x \in C} f(x)$

- Bounded subgradients: $||g||_2 \leq G$ for all $g \in \partial f$
- Initialization radius: $||x^{(1)} - x^*||_2 \leq R$

Projected sub-gradient method iterates will satisfy

$$f_{\text{best}}^{(k)} - f^* \leq \frac{R^2 + G^2 \sum_{i=1}^k \alpha_i^2}{2 \sum_{i=1}^k \alpha_i}$$

setting $\alpha_i = (R/G)/\sqrt{k}$ gives

$$f_{\text{best}}^{(k)} - f^* \leq \frac{RG}{\sqrt{k}}$$

$G = \max_{x \in C} ||\partial f(x)||_2$ and $R = \max_{x,y \in C} ||x - y||_2$ The analysis and the convergence results depend on Euclidean (ℓ_2) norm

Bregman Divergence

h convex differentiable over an open convex set C.

- The Bregman divergence associated to h is defined by

$$D_h(x,y) = h(x) - \left[h(y) + \nabla h(y)^T (x-y)\right]$$

can be interpreted as
distance between x and y as measured by the function h
Example: $h(x) = ||x||_2^2$

Strong convexity

$h(x)$ is λ-strongly convex with respect to the norm $||\cdot||$ if

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

Properties of Bregman divergence

For a λ-strongly convex function h, Bregman divergence

$$D_h(x, y) = h(x) - \left[h(y) + \nabla h(y)^T (x - y)\right]$$

satisfies

$$D_h(x, y) \geq \frac{\lambda}{2} ||x - y||^2 \geq 0$$

Pythagorean theorem

Bregman projection

$$P_C^h(y) = \arg\min_{x \in C} D_h(x, y)$$

$$D_h(x, y) \geq D_h(x, P_C^h(y)) + D_h(P_C^h(y), y)$$

Projected Gradient Descent

$$x^{(k+1)} = P_C \arg\min_x \left\{ f(x^k) + g^{(k)T}(x - x^{(k)}) + \frac{1}{2\alpha_k}||x - x^{(k)}||_2^2) \right\}$$

where P_C is the Euclidean projection onto C.

Mirror Descent

$$x^{(k+1)} = P_C^h \operatorname*{argmin}_x \left\{ f(x^k) + g^{(k)T}(x - x^{(k)}) + \frac{1}{\alpha_k} D_h(x, x^{(k)}) \right\}$$

where $D_h(x, y)$ is the Bregman divergence

$$D_h(x, y) = h(x) - \left[h(y) + \nabla h(y)^T (x - y) \right]$$

and $h(x)$ is strongly convex with respect to $|| \cdot ||$
P_C^h is the Bregman projection:

$$P_C^h(y) = \arg \min_{x \in C} D_h(x, y)$$

Mirror Descent update rule

$$\begin{aligned}
x^{(k+1)} &= P_C^h \operatorname*{argmin}_x \left\{ f(x^k) + g^{(k)T}(x - x^{(k)}) + \frac{1}{\alpha_k} D_h(x, x^{(k)}) \right\} \\
&= P_C^h \operatorname*{argmin}_x \left\{ g^{(k)T} x + \frac{1}{\alpha_k} D_h(x, x^{(k)}) \right\} \\
&= P_C^h \operatorname*{argmin}_x \left\{ g^{(k)T} x + \frac{1}{\alpha_k} h(x) - \frac{1}{\alpha_k} \nabla h(x^{(k)})^T x \right\}
\end{aligned}$$

optimality condition for $y =$ argmin:$g^k + \frac{1}{\alpha_k}\nabla h(y) - \frac{1}{\alpha_k}\nabla h(x^{(k)}) = 0$

$D_h(x, y)$ is the Bregman divergence

$$D_h(x, y) = h(x) - \left[h(y) + \nabla h(y)^T (x - y) \right]$$

P_C^h is the Bregman projection:

$$\begin{aligned}
P_C^h(y) = \arg\min_{x \in C} D_h(x, y) &= \arg\min_{x \in C} h(x) - \nabla h(y)^T x \\
&= \arg\min_{x \in C} h(x) - (\nabla h(x^{(k)}) - \alpha_k g^{(k)})^T x \\
&= \arg\min_{x \in C} D_h(x, x^{(k)}) + \alpha_k g^{(k)T} x
\end{aligned}$$

Mirror Descent update rule (simplified)

$$x^{(k+1)} = P_C^h \operatorname*{argmin}_x \left\{ f(x^k) + g^{(k)T}(x - x^{(k)}) + \frac{1}{\alpha_k} D_h(x, x^{(k)}) \right\}$$

$$= \arg\min_{x \in C} \ D_h(x, x^{(k)}) + \alpha_k {g^{(k)}}^T x$$

where $D_h(x, y)$ is the Bregman divergence

$$D_h(x, y) = h(x) - \left[h(y) + \nabla h(y)^T (x - y) \right]$$

Convergence guarantees

Let $||g||_* \leq G_{||\cdot||}$ for all $g \in \partial f$, or equivalently

$$f(x) - f(y) \leq G_{||\cdot||}||x - y||$$

Let $x^h = \arg\min_{x \in C} h(x)$ and $R^h_{||\cdot||} = \left(2 \max_y D_h(x^h, y)/\lambda\right)^{1/2}$, then

$$||x - x^h|| \leq R^h_{||\cdot||}$$

General guarantee:

$$\sum_{i=1}^{k} \alpha_i [f(x^{(i)}) - f(x^\star)] \leq D(x^\star, x^{(1)}) + \frac{1}{2} \sum_{i=1}^{k} \alpha_i^2 \|g^{(i)}\|_*^2$$

Choose step size $\alpha_k = \frac{\lambda R^h_{||\cdot||}}{G_{||\cdot||}\sqrt{k}}$. Mirror Descent iterates satisfy

$$f^{(k)}_{best} - f^* \leq \frac{R^h_{||\cdot||} G_{||\cdot||}}{\sqrt{k}}$$

Standard setups for Mirror Descent

$$
\begin{aligned}
x^{(k+1)} &= P_C^h \operatorname*{argmin}_x \left\{ f(x^k) + g^{(k)T}(x - x^{(k)}) + \frac{1}{\alpha_k} D_h(x, x^{(k)}) \right\} \\
&= \arg\min_{x \in C} \; D_h(x, x^{(k)}) + \alpha_k {g^{(k)}}^T x
\end{aligned}
$$

where $D_h(x, y)$ is the Bregman divergence

$$
D_h(x, y) = h(x) - \left[h(y) + \nabla h(y)^T (x - y) \right]
$$

- simplest version is when $h(x) = \frac{1}{2}||x||_2^2$, which is strongly convex w.r.t. $||\cdot||_2$. Mirror Descent= Projected Subgradient Descent $D_h(x, y) = ||x - y||_2^2$
- negative entropy $h(x) = \sum_{i=1}^n x_i \log x_i$, which is 1-strongly convex wrt w.r.t. $||x||_1$ (Pinsker's inequality). $D_h(x, y) = \sum_{i=1}^n x_i \log \frac{x_i}{y_i} - (x_i - y_i)$ is the generalized Kullback-Leibler divergence

Negative Entropy

- negative entropy $h(x) = \sum_{i=1}^n x_i \log x_i$
  $D_h(x, y) = \sum_{i=1}^n x_i \log \frac{x_i}{y_i} - (x_i - y_i)$ is the generalized Kullback-Leibler divergence
- unit simplex $C = \Delta_n = \{x \in \mathbf{R}_+^n \ : \ \sum_i x_i = 1\}$.
- Bregman projection onto the simplex is a simple renormalization

$$P^h(Y) = \frac{y}{||y||_1}$$

- Mirror Descent:

$$x^{(k+1)} = P_C^h \operatorname*{argmin}_x \left\{ f(x^k) + g^{(k)T}(x - x^{(k)}) + \frac{1}{\alpha_k} D_h(x, x^{(k)}) \right\}$$

- $y \in \arg\min \implies \nabla h(y) = \log(y) + 1 = \nabla h(x^{(k)}) - \alpha_k g^k$
  $\implies y_i = x_i^{(k)} \exp\left(-\alpha_k g_i^k\right)$
- Mirror Descent update:

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

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 $D_h(x^\star, x^{(1)}) \leq \log n$ for $x^\star \in C$

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

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

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

Example

Robust regression problem (an LP):

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

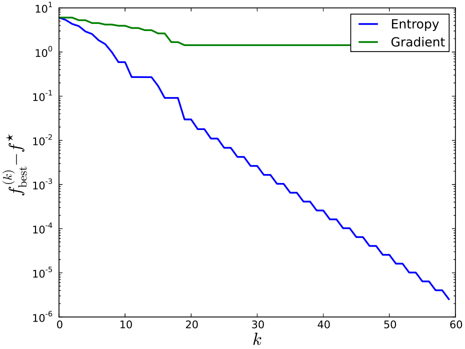
subgradient of objective is $g = \sum_{i=1}^m \mathrm{sign}(a_i^T x - b_i) a_i$

- 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)

Example: Spectrahedron

minimizing a function on the spectrahedron S_n defined as

$$S_n = \{X \in \mathbb{S}^n_+ \ : \ \mathbf{tr}(X) = 1\}$$

Example: Spectrahedron

$$S_n = \{X \in \mathbb{S}^n_+ \ : \ \mathbf{tr}(X) = 1\}$$

- von Neumann entropy:

$$h(X) = \sum_{i=1}^{n} \lambda_i(X) \log \lambda_i(X)$$

  where $\lambda_1(X), ..., \lambda_n(X)$ are the eigenvalues of X.
- $\frac{1}{2}$ strongly convex with respect to the norm

$$||X||_{tr} = \sum_{i=1}^{n} \lambda_i(X)$$

- Mirror Descent update:

$$Y_{t+1} = \exp(\log X_t - \alpha_t \nabla f(X_t))$$
$$X_{t+1} = P_C^h(Y_t + 1) = Y_{t+1}/||Y_{t+1}||_{tr}$$

Mirror Descent Analysis

distance generating function h, 1-strongly-convex w.r.t. $\|\cdot\|$:

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

Fenchel conjugate

$$h^*(\theta) = \sup_{x \in C} \left\{ \theta^T x - h(x) \right\} \;,\; \nabla h^*(\theta) = \operatorname*{argmax}_{x \in C} \left\{ \theta^T x - h(x) \right\}$$

∇h, ∇h^* take us "through the mirror" and back

$$x \underset{\nabla h^*}{\overset{\nabla h}{\rightleftarrows}} \theta$$

miror descent iterations for $C = \mathbf{R}^n$

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

$h(x) = \frac{1}{2} \|x\|_2^2$ recovers standard case

Convergence analysis

$$g^{(k)} \in \partial f(x^{(k)}) \ , \ \theta^{(k+1)} = \theta^{(k)} - \alpha_k g^{(k)} \ , \ x^{(k+1)} = \nabla h^*(\theta^{(k+1)})$$

Bregman divergence

$$D_{h^*}(\theta', \theta) = h^*(\theta') - h^*(\theta) - \nabla h^*(\theta)^T(\theta' - \theta)$$

Let $\theta^\star = \nabla h(x^\star)$,

$$\begin{aligned} D_{h^*}(\theta^{(k+1)}, \theta^\star) = \ & D_{h^*}(\theta^{(k)}, \theta^\star) \\ & + (\theta^{(k+1)} - \theta^{(k)})^T(\nabla h^*(\theta^{(k)}) - \nabla h^*(\theta^\star)) \\ & + D_{h^*}(\theta^{(k+1)}, \theta^{(k)}) \end{aligned}$$

and

$$(\theta^{(k+1)} - \theta^{(k)})^T(\nabla h^*(\theta^{(k)}) - \nabla h^*(\theta^\star)) = -\alpha_k {g^{(k)}}^T(x^{(k)} - x^\star)$$

Convergence analysis continued

From convexity and $g^{(k)} \in \partial f(x^{(k)})$,

$$f(x^{(k)}) - f(x^\star) \le {g^{(k)}}^T (x^{(k)} - x^\star)$$

Therefore

$$\begin{aligned}\alpha_k[f(x^{(k)}) - f(x^\star)] \le D_{h^*}(\theta^{(k)}, \theta^\star) - D_{h^*}(\theta^{(k+1)}, \theta^\star) \\ + D_{h^*}(\theta^{(k+1)}, \theta^{(k)})\end{aligned}$$

Fact: h is 1-strongly-convex w.r.t. $\|\cdot\| \Leftrightarrow D_h(x', x) \ge \frac{1}{2}\|x' - x\|^2 \Leftrightarrow$
h^* is 1-smooth w.r.t. $\|\cdot\|_*$ $\quad\Leftrightarrow D_{h^*}(\theta', \theta) \le \frac{1}{2}\|\theta' - \theta\|_*^2$

Bounding the $D_{h^*}(\theta^{(k+1)}, \theta^{(k)})$ terms and telescoping gives

$$\sum_{i=1}^{k} \alpha_i [f(x^{(i)}) - f(x^\star)] \le D_{h^*}(\theta^{(1)}, \theta^\star) + \frac{1}{2}\sum_{i=1}^{k} \alpha_i^2 \|g^{(i)}\|_*^2$$

Convergence guarantees

Note: $D_{h^*}(\theta^{(1)}, \theta^\star) = D_h(x^\star, x^{(1)})$

Most general guarantee,

$$\sum_{i=1}^{k} \alpha_i [f(x^{(i)}) - f(x^\star)] \le D_h(x^\star, x^{(1)}) + \frac{1}{2} \sum_{i=1}^{k} \alpha_i^2 \|g^{(i)}\|_*^2$$

Fixed step size $\alpha_k = \alpha$

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

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 $x \in C$, $\|g\|_* \le G$ for some $G < \infty$

Stochastic gradients are fine!

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

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

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

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

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

numerator of additional term can be bounded to get estimates

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

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

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

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

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

converges if

- $R_\infty < \infty$ (e.g. if $\mathcal{X}$ is compact)
- $\sum_{i=1}^{k} \|g^{(i)}\|_{H_i^{-1}}^2$ 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

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

(second line is a theorem)

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

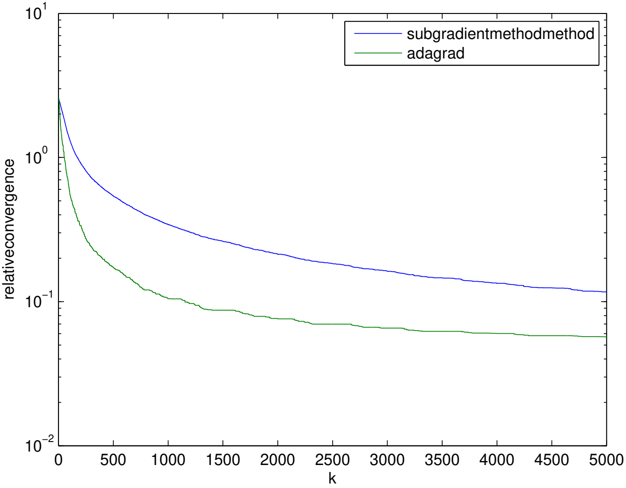
$$f_{\rm best}^{(k)} - f^\star \leq \frac{2}{k}\inf_{h\geq 0}\left\{\sup_{x,y\in C}(x-y)^T\,\mathbf{diag}(h)(x-y) + \sum_{i=1}^k \|g^{(i)}\|_{\mathbf{diag}(h)^{-1}}^2\right\}$$

Example

Classification problem:

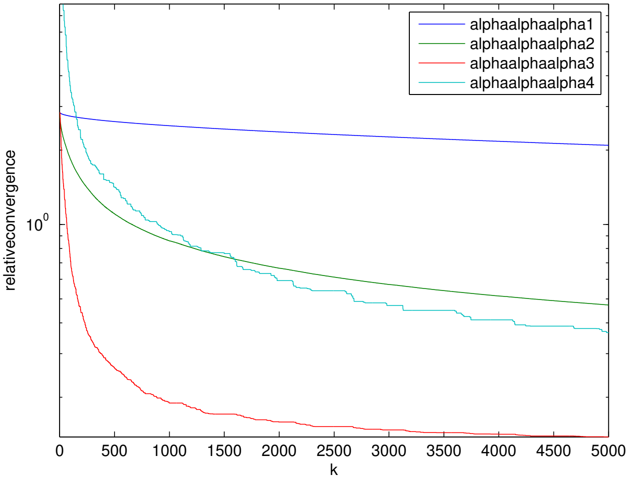
- **Data**: $\{a_i, b_i\}$, $i = 1, \ldots, 50000$
  - $a_i \in \mathbf{R}^{1000}$
  - $b \in \{-1, 1\}$
  - Data created with 5% mis-classifications w.r.t. $w = \mathbf{1}$, $v = 0$
- **Objective**: find classifiers $w \in \mathbf{R}^{1000}$ and $v \in \mathbf{R}$ such that
  - $a_i^T w + v > 1$ if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