Subgradient Methods

John Duchi

Outline

The problem

Problem for now:

$$\underset{x}{\text{minimize}}\ \ f(x)$$

where f convex, not necessarily differentiable

Gradient method

Consider

$$\underset{x}{\text{minimize}} \;\; f(x)$$

where f convex and continuously differentiable

Gradient method: For some stepsize sequence α_k, iterate

$$\begin{aligned} x_{k+1} &= x_k - \alpha_k \nabla f(x_k) \\ &= \operatorname*{argmin}_x \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\alpha_k} \left\| x - x_k \right\|_2^2 \right\} \end{aligned}$$

Subgradient method

Iterate

$$\text{Choose } \textit{any } g_k \in \partial f(x_k)$$
$$\text{Update } x_{k+1} = x_k - \alpha_k g_k$$

- Not a descent method
- $\alpha_k > 0$ is kth step size

Convergence proof start

A few assumptions to make our lives easier:

- Optimal point: $f^\star = \inf_x f(x) > -\infty$ and there is $x^\star \in \mathbb{R}^n$ with $f(x^\star) = f^\star$
- Lipschitz condition: $\|g\|_2 \leq M$ for all $g \in \partial f(x)$ and all x
- $\|x_1 - x^\star\|_2 \leq R$

(Stronger than needed but whatever)

Convergence proof

Key quantity: distance to optimal point $x^\star$

Convergence proof II

Key step: recursion

Convergence guarantee

Have guarantees

$$\sum_{k=1}^{K} \alpha_k [f(x_k) - f(x^\star)] \le \frac{1}{2} \|x_1 - x^\star\|_2^2 + \sum_{k=1}^{K} \frac{\alpha_k^2}{2} \|g_k\|_2^2$$

or, if $\overline{x}_K = \sum_{k=1}^{K} \alpha_k x_k / \sum_{k=1}^{K} \alpha_K$,

$$f(\overline{x}_K) - f(x^\star) \le \frac{R^2 + \frac{1}{2}\sum_{k=1}^{K} \alpha_k^2 M^2}{\sum_{k=1}^{K} \alpha_k}$$

Convergence guarantee

For fixed stepsize α and $\overline{x}_K = \frac{1}{K}\sum_{k=1}^{K} x_k$, have

$$f(\overline{x}_K) - f(x^\star) \le \frac{R^2}{\alpha K} + \frac{\alpha}{2} M^2.$$

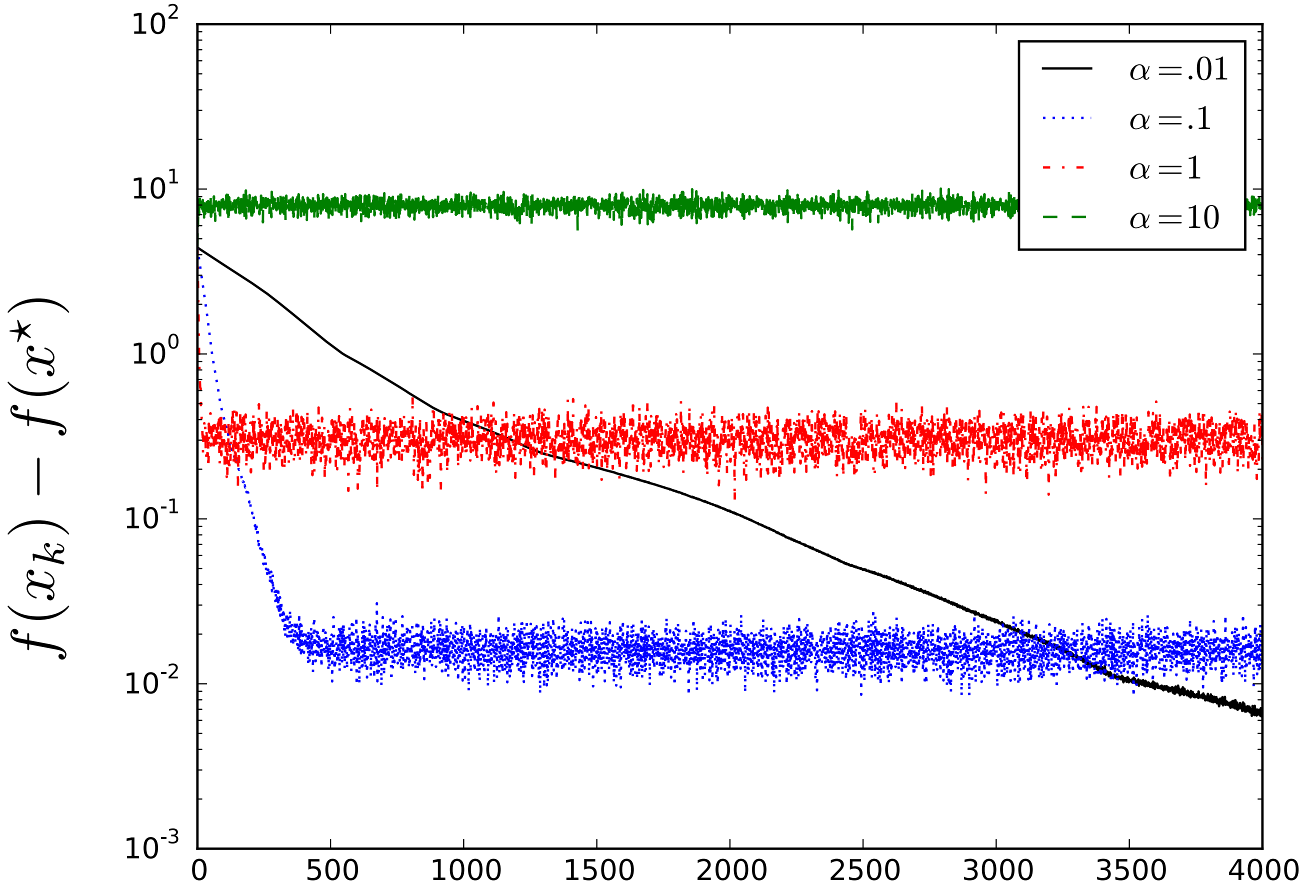
Example: robust regression

$$\text{minimize } f(x) = \frac{1}{m} \|Ax - b\|_1 = \frac{1}{m} \sum_{i=1}^{m} |a_i^T x - b_i|.$$

(Recall: $\partial \|x\|_1 = \text{sign}(x)$, so $\partial f(x) = A^T \text{sign}(Ax - b)$)

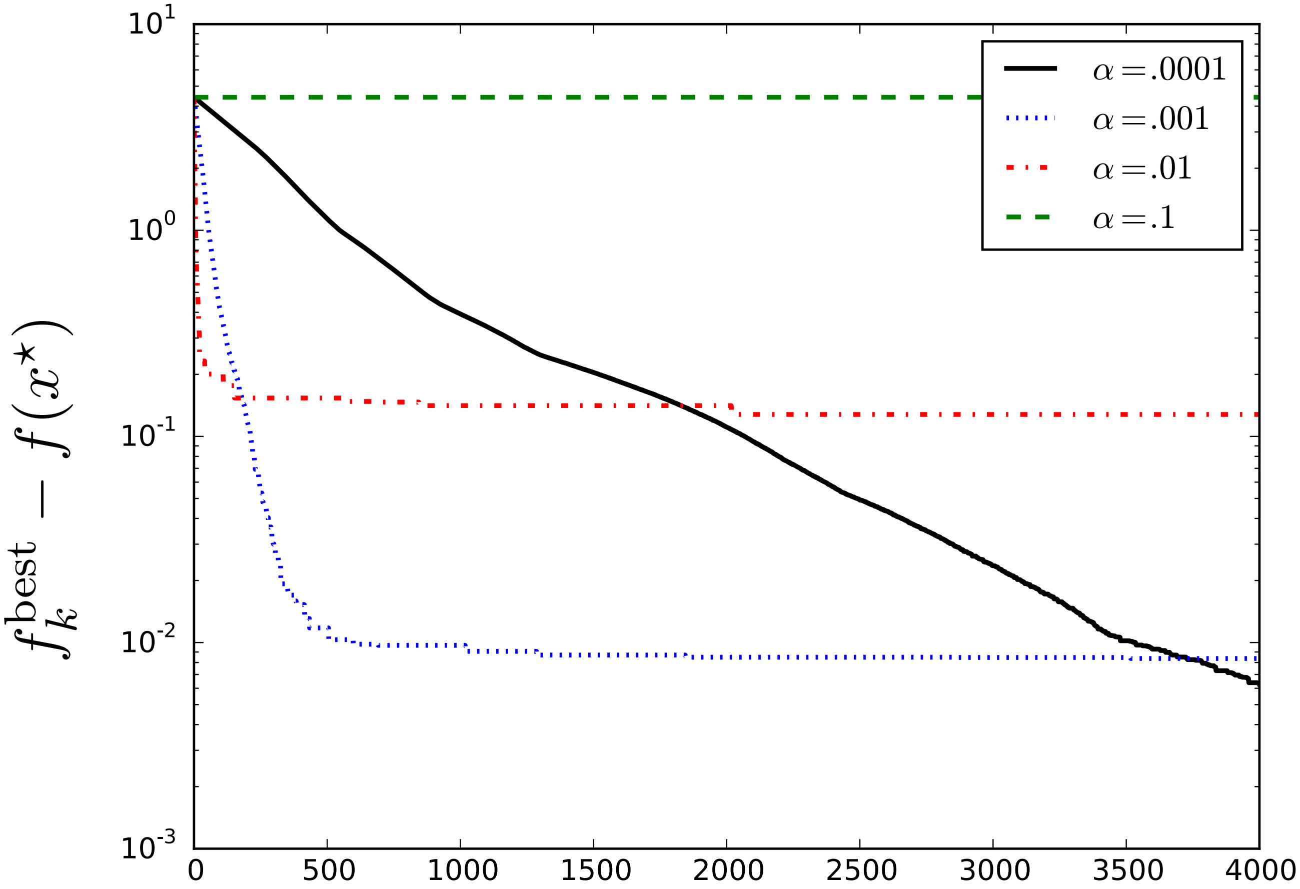
- Perform subgradient descent with fixed stepsize $\alpha \in \{10^{-2}, 10^{-1}, 1, 10\}$.
- Plot $f(x_k) - f^\star$
- Use $f_k^{\text{best}} = \min_{i \le k} f(x_i)$ and plot $f_k^{\text{best}} - f^\star$

Robust regression example



Fixed stepsizes, showing $f(x_k) - f(x^\star)$ for $f(x) = \|Ax - b\|_1$. Here $A \in \mathbb{R}^{100 \times 50}$

Robust regression example



Fixed stepsizes, showing $f_k^{\text{best}} - f(x^\star)$ for $f(x) = \|Ax - b\|_1$. Here $A \in \mathbb{R}^{100 \times 50}$

Projected subgradient method

Solve problem

$$\underset{x}{\text{minimize}} \;\; f(x) \quad \text{subject to} \;\; x \in C$$

where C is a closed convex set

Projected gradient method Iterate:

- Pick $g_k \in \partial f(x_k)$
- Update

$$\begin{aligned} x_{k+1} &= \pi_C(x_k - \alpha_k g_k) \\ &= \underset{x \in C}{\operatorname{argmin}} \left\{ \langle g_k, x \rangle + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 \right\} \end{aligned}$$

where

$$\pi_C(x) := \underset{y \in C}{\operatorname{argmin}} \|x - y\|_2^2 .$$

Projected subgradient method

- Pick $g_k \in \partial f(x_k)$
- Update

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

where

$$\pi_C(x) := \operatorname*{argmin}_{y \in C} \|x - y\|_2^2 .$$

Projected subgradient method: Convergence

Assume: $\|x - x^{\star}\|_2^2 \le R^2$ for all $x \in C$

One inequality to rule them all

$$\|\pi_C(x) - y\|_2^2 \le \|x - y\|_2^2$$

for $y \in C$

Projected subgradient method: Convergence II

Variant on recursion:

$$f(x_k) - f(x^\star) \le \frac{1}{2\alpha_k}\left[\|x_k - x^\star\|_2^2 - \|x_{k+1} - x^\star\|_2^2\right] + \frac{\alpha_k}{2}\|g_k\|_2^2.$$

Projected subgradient method: Convergence III

Variant on recursion:

$$\sum_{k=1}^{K} [f(x_k) - f(x^\star)] \le \frac{1}{2\alpha_K} R^2 + \sum_{k=1}^{K} \frac{\alpha_k}{2} \|g_k\|_2^2 .$$

Example

ℓ_2-constraint:

Let $C = \{x \in \mathbb{R}^n : \|x\|_2 \le R\}$. Then $\|x - x^\star\|_2 \le 2R$ for all $x, x^\star$ and

$$\pi_C(x) = \begin{cases} x & \text{if } \|x\|_2 \le R \\ R\frac{x}{\|x\|_2} & \text{otherwise.} \end{cases}$$

Stochastic subgradient methods

Stochastic subgradient: Given function f, a *stochastic* subgradient for a point x is a random vector with

$$\mathbb{E}[g \mid x] \in \partial f(x).$$

Standard example: Expectations. Let S be random variable,

$$f(x) = \mathbb{E}[F(x; S)] = \int F(x; s) dP(s)$$

where $F(\cdot; s)$ is convex. Given x, draw $S \sim P$ and set

$$g = g(x; S) \in \partial F(x; S).$$

(Projected) stochastic subgradient method

Problem:

$$\text{minimize} \;\; f(x) \quad \text{subject to} \, x \in C$$

given access to *stochastic gradients* of f

Method: Iterate with stepsizes $\alpha_k > 0$

- Get stochastic gradient g_k for f at x_k, i.e. $\mathbb{E}[g_k \mid x_k] \in \partial f(x_k)$
- Update

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

Motivation and example

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} F(x; S_i)$$

for very large sample $\{S_1, \ldots, S_N\}$.

- True subgradient: take $g_i \in \partial F(x; S_i)$ and

$$g = \frac{1}{N} \sum_{i=1}^{N} g_i$$

- Stochastic subgradient: choose $i \in \{1, \ldots, N\}$ uniformly at random, take $g \in \partial F(x; S_i)$.

Motivation and example

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} F(x; S_i)$$

for very large sample $\{S_1, \ldots, S_N\}$.

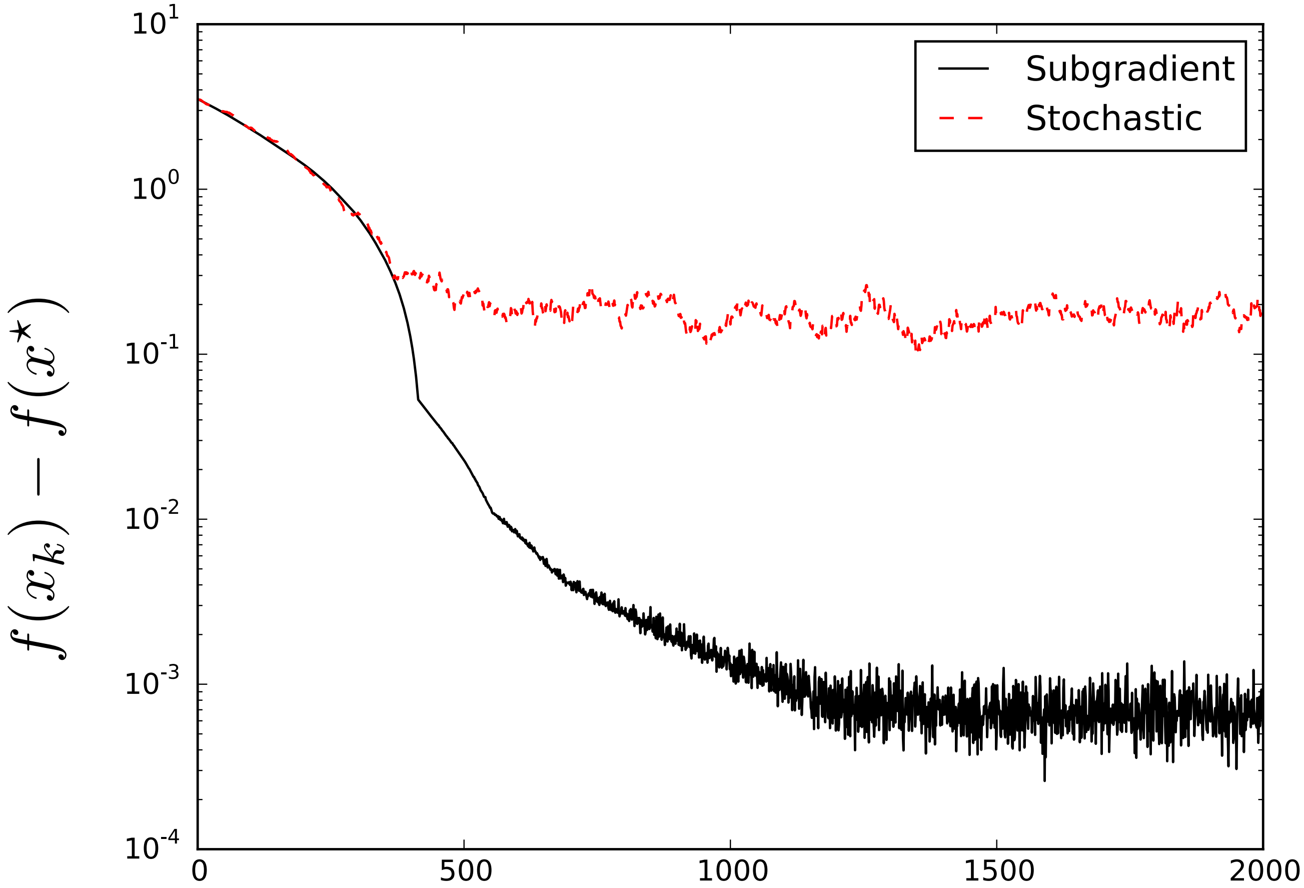
- True subgradient: take $g_i \in \partial F(x; S_i)$ and

$$g = \frac{1}{N} \sum_{i=1}^{N} g_i$$

- Stochastic subgradient: choose $i \in \{1, \ldots, N\}$ uniformly at random, take $g \in \partial F(x; S_i)$.

Example: robust regression

$$f(x) = \frac{1}{m}\left\|Ax - b\right\|_1 = \frac{1}{m}\sum_{i=1}^{m} |\left\langle a_i, x\right\rangle - b_i|.$$



Convergence proof

- Compact set C, so $\|x - y\|_2 \le R$ for all $x, y \in C$
- $\mathbb{E}[\|g\|_2^2] \le M^2$ for stochastic subgradients
- Define error $\xi_k = g_k - f'(x_k)$, where $\mathbb{E}[g_k \mid x_k] = f'(x_k) \in \partial f(x_k)$

Starting point:

$$\|x_{k+1} - x^\star\|_2^2 = \|\pi_C(x_k - \alpha_k g_k) - x^\star\|_2^2 \le \|x_k - \alpha_k g_k - x^\star\|_2^2$$

Convergence proof II

$$\|x_{k+1} - x^\star\|_2^2 \le \|x_k - x^\star\|_2^2 - 2\alpha_k \left\langle f'(x_k), x_k - x^\star \right\rangle + \alpha_k^2 \|g_k\|_2 \\ - 2\alpha_k \langle \xi_k, x_k - x^\star \rangle$$

Convergence of Stochastic Gradient Descent

Final convergence guarantee if C compact and $\|x - y\|_2 \le R$ for $x, y \in C$:

$$\sum_{k=1}^{K} [f(x_k) - f(x^\star)] \le \frac{1}{2\alpha_K} R^2 + \frac{1}{2} \sum_{k=1}^{K} \alpha_k \|g_k\|_2^2 - \sum_{k=1}^{K} \langle \xi_k, x_k - x^\star \rangle .$$

Take Expectations:

Convergence of Stochastic Gradient Descent II

Expected convergence guarantee: If $\alpha_k = R/M\sqrt{k}$ and $\overline{x}_K = \frac{1}{K}\sum_{k=1}^{K} x_k$,

$$\mathbb{E}[f(\overline{x}_K) - f(x^\star)] \leq \frac{3}{2}\frac{RM}{\sqrt{K}}.$$

High Probability Convergence

Question: Can we get convergence with high probability?

Theorem: (Azuma-Hoeffding inequality). Let $Z_1, Z_2, \ldots, Z_K$ be a sequence of conditionally mean-zero random variables with $|Z_k| \le B$ for all k, i.e.

$$\mathbb{E}[Z_k \mid Z_1, \ldots, Z_{k-1}] = 0 \quad \text{and} \quad \max_k |Z_k| \le B < \infty.$$

Then

$$\mathbb{P}\left(\frac{1}{K}\sum_{k=1}^{K} Z_k \ge t\right) \le \exp\left(-\frac{Kt^2}{2B^2}\right)$$

for all $t \ge 0$.

High Probability Convergence

Assume that $\|g\|_2 \le M$ for any stochastic subgradient g. Have guarantee (always)

$$f(\overline{x}_K) - f(x^\star) \le \frac{1}{2K\alpha_K} R^2 + \frac{1}{K}\sum_{k=1}^{K} \frac{\alpha_k}{2} M^2 - \frac{1}{K}\sum_{k=1}^{K} \langle \xi_k, x_k - x^\star \rangle \, .$$

High Probability Convergence

Theorem: If $\alpha_k > 0$ is non-increasing, $\|x - y\|_2 \le R$ for all $x, y \in C$, and $\|g\|_2 \le M$ for all stochastic gradients, then

$$f(\overline{x}_K) - f(x^\star) \le \frac{1}{2K\alpha_K} R^2 + \frac{1}{K} \sum_{k=1}^{K} \frac{\alpha_k}{2} M^2 + \frac{2MR}{\sqrt{K}} \epsilon$$

with probability at least $1 - \exp(-\epsilon^2)$.