Optimization methods

Optimization-Based Data Analysis

http://www.cims.nyu.edu/~cfgranda/pages/OBDA_spring16

Carlos Fernandez-Granda

2/8/2016

Introduction

Aim: Overview of optimization methods that

- Tend to scale well with the problem dimension
- Are widely used in machine learning and signal processing
- Are (reasonably) well understood theoretically

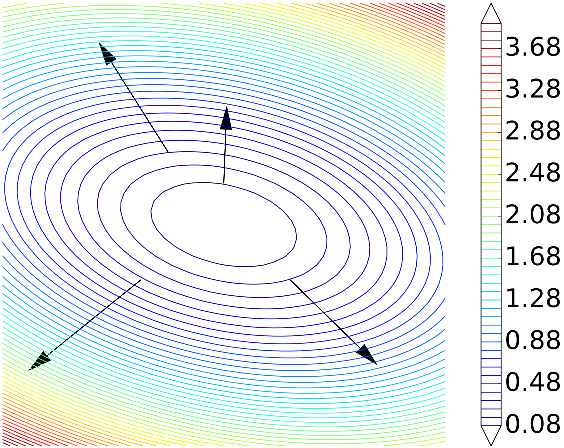
Differentiable functions

- Gradient descent
- Convergence analysis of gradient descent
- Accelerated gradient descent
- Projected gradient descent

Nondifferentiable functions

- Subgradient method
- Proximal gradient method
- Coordinate descent

Gradient



Direction of maximum variation

Gradient descent (aka steepest descent)

Method to solve the optimization problem

$$\text{minimize} \quad f(x),$$

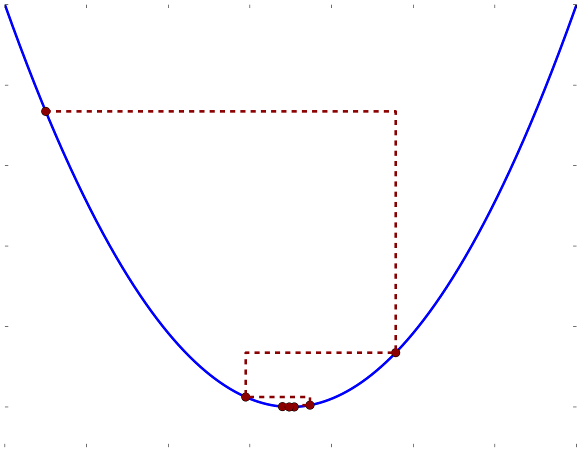
where f is differentiable

Gradient-descent iteration:

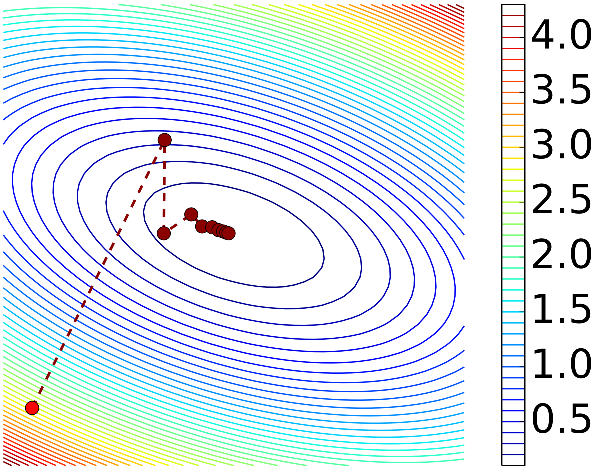
$$x^{(0)} = \text{arbitrary initialization}$$
$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f\left(x^{(k)}\right)$$

where α_k is the step size

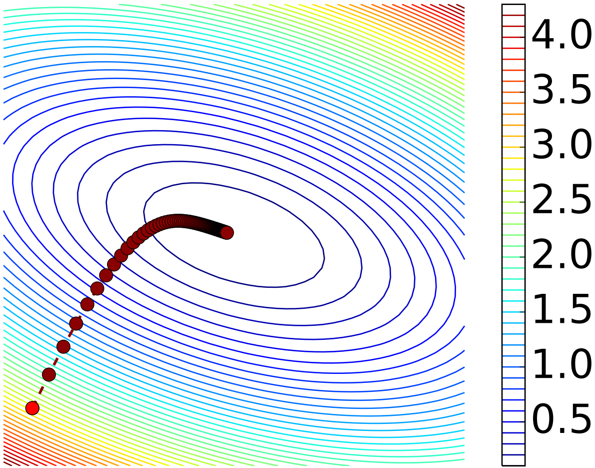
Gradient descent (1D)



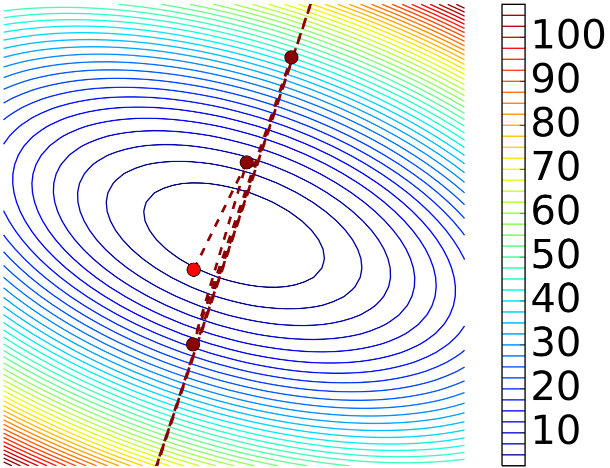
Gradient descent (2D)



Small step size



Large step size



Line search

- Exact

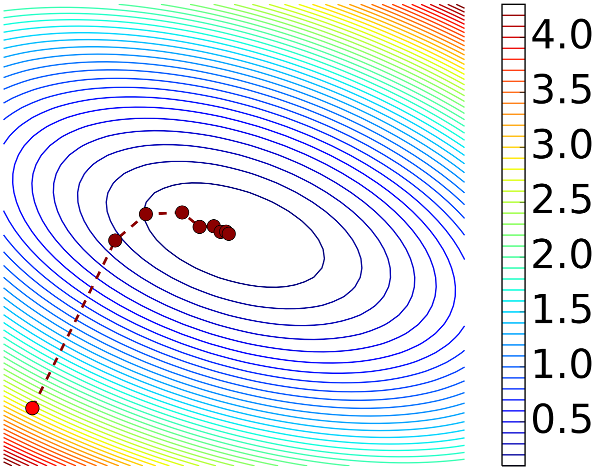
$$\alpha_k := \arg\min_{\beta \geq 0} f\left(x^{(k)} - \beta \nabla f\left(x^{(k)}\right)\right)$$

- Backtracking (Armijo rule)

  Given $\alpha^0 \geq 0$ and $\beta \in (0, 1)$, set $\alpha_k := \alpha^0 \, \beta^i$ for the smallest i such that

$$f\left(x^{(k+1)}\right) \leq f\left(x^{(k)}\right) - \frac{1}{2} \alpha_k \left|\left|\nabla f\left(x^{(k)}\right)\right|\right|_2^2$$

Backtracking line search



Lipschitz continuity

A function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is Lipschitz continuous with Lipschitz constant L if for any $x, y \in \mathbb{R}^n$

$$||f(y) - f(x)||_2 \leq L\,||y - x||_2$$

Example:

$f(x) := Ax$ is Lipschitz continuous with $L = \sigma_{\max}(A)$

Quadratic upper bound

If the gradient of $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz continuous with constant L

$$||\nabla f(y) - \nabla f(x)||_2 \leq L\,||y - x||_2$$

then for any $x, y \in \mathbb{R}^n$

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}\,||y - x||_2^2$$

Consequence of quadratic bound

Since $x^{(k+1)} = x^{(k)} - \alpha_k \nabla f\left(x^{(k)}\right)$

$$f\left(x^{(k+1)}\right) \leq f\left(x^{(k)}\right) - \alpha_k \left(1 - \frac{\alpha_k L}{2}\right) \left\|\nabla f\left(x^{(k)}\right)\right\|_2^2$$

If $\alpha_k \leq \frac{1}{L}$ the value of the function always decreases!

$$f\left(x^{(k+1)}\right) \leq f\left(x^{(k)}\right) - \frac{\alpha_k}{2} \left\|\nabla f\left(x^{(k)}\right)\right\|_2^2$$

Gradient descent with constant step size

Conditions:

- f is convex
- ∇f is L-Lipschitz continuous
- There exists a solution x^* such that $f(x^*)$ is finite

If $\alpha_k = \alpha \leq \frac{1}{L}$

$$f\left(x^{(k)}\right) - f(x^*) \leq \frac{\left|\left|x^{(k)} - x^{(0)}\right|\right|_2^2}{2\,\alpha\,k}$$

We need $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ iterations to get an ϵ-optimal solution

Proof

Recall that if $\alpha \leq \frac{1}{L}$

$$f\left(x^{(i)}\right) \leq f\left(x^{(i-1)}\right) - \frac{\alpha}{2}\left|\left|\nabla f\left(x^{(i-1)}\right)\right|\right|_2^2$$

By the first-order characterization of convexity

$$f\left(x^{(i-1)}\right) - f\left(x^*\right) \leq \nabla f\left(x^{(i-1)}\right)^T\left(x^{(i-1)} - x^*\right)$$

This implies

$$\begin{aligned}
f\left(x^{(i)}\right) - f\left(x^*\right) &\leq \nabla f\left(x^{(i-1)}\right)^T\left(x^{(i-1)} - x^*\right) - \frac{\alpha}{2}\left|\left|\nabla f\left(x^{(i-1)}\right)\right|\right|_2^2 \\
&= \frac{1}{2\,\alpha}\left(\left|\left|x^{(i-1)} - x^*\right|\right|_2^2 - \left|\left|x^{(i-1)} - x^* - \alpha\nabla f\left(x^{(i-1)}\right)\right|\right|_2^2\right) \\
&= \frac{1}{2\,\alpha}\left(\left|\left|x^{(i-1)} - x^*\right|\right|_2^2 - \left|\left|x^{(i)} - x^*\right|\right|_2\right)
\end{aligned}$$

Proof

Because the value of f never increases,

$$\begin{aligned}
f\left(x^{(k)}\right) - f\left(x^{*}\right) &\leq \frac{1}{k}\sum_{i=1}^{k} f\left(x^{(i)}\right) - f\left(x^{*}\right) \\
&= \frac{1}{2\,\alpha\,k}\left(\left|\left|x^{(0)} - x^{*}\right|\right|_2^2 - \left|\left|x^{(k)} - x^{*}\right|\right|_2^2\right) \\
&\leq \frac{\left|\left|x^{(0)} - x^{*}\right|\right|_2^2}{2\,\alpha\,k}
\end{aligned}$$

Backtracking line search

If the gradient of $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz continuous with constant L the step size in the backtracking line search satisfies

$$\alpha_k \geq \alpha_{\min} := \min \left\{ \alpha^0, \frac{\beta}{L} \right\}$$

Proof

Line search ends when

$$f\left(x^{(k+1)}\right) \leq f\left(x^{(k)}\right) - \frac{\alpha_k}{2}\left|\left|\nabla f\left(x^{(k)}\right)\right|\right|_2^2$$

but we know that if $\alpha_k \leq \frac{1}{L}$

$$f\left(x^{(k+1)}\right) \leq f\left(x^{(k)}\right) - \frac{\alpha_k}{2}\left|\left|\nabla f\left(x^{(k)}\right)\right|\right|_2^2$$

This happens as soon as $\beta/L \leq \alpha^0\beta^i \leq 1/L$

Gradient descent with backtracking

Under the same conditions as before gradient descent with backtracking line search achieves

$$f\left(x^{(k)}\right) - f\left(x^*\right) \leq \frac{\left|\left|x^{(0)} - x^*\right|\right|_2^2}{2\,\alpha_{\min}\,k}$$

$\mathcal{O}\left(\frac{1}{\epsilon}\right)$ iterations to get an ϵ-optimal solution

Strong convexity

A function $f : \mathbb{R}^n$ is strongly convex if for any $x, y \in \mathbb{R}^n$

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + S \|y - x\|^2 .$$

Example:

$$f(x) := \|Ax - y\|_2^2$$

where $A \in \mathbb{R}^{m \times n}$ is strongly convex with $S = \sigma_{\min}(A)$ if $m > n$

Gradient descent for strongly convex functions

If f is S-strongly convex and ∇f is L-Lipschitz continuous

$$f\left(x^{(k)}\right) - f\left(x^*\right) \leq \frac{c^k L \left|\left|x^{(k)} - x^{(0)}\right|\right|_2^2}{2}$$

$$c := \frac{\frac{L}{S} - 1}{\frac{L}{S} + 1}$$

We need $\mathcal{O}\left(\log \frac{1}{\epsilon}\right)$ iterations to get an ϵ-optimal solution

Lower bounds for convergence rate

There exist convex functions with L-Lipschitz-continuous gradients such that for any algorithm that selects $x^{(k)}$ from

$$x^{(0)} + \text{span}\left\{\nabla f\left(x^{(0)}\right), \nabla f\left(x^{(1)}\right), \ldots, \nabla f\left(x^{(k-1)}\right)\right\}$$

we have

$$f\left(x^{(k)}\right) - f\left(x^{*}\right) \geq \frac{3L\left|\left|x^{(0)} - x^{*}\right|\right|_2^2}{32\left(k+1\right)^2}$$

Nesterov's accelerated gradient method

Achieves lower bound, i.e. $\mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\right)$ convergence

Uses momentum variable

$$
\begin{aligned}
y^{(k+1)} &= x^{(k)} - \alpha_k \nabla f\left(x^{(k)}\right) \\
x^{(k+1)} &= \beta_k y^{(k+1)} + \gamma_k y^{(k)}
\end{aligned}
$$

Despite guarantees, *why* this works is not completely understood

Differentiable functions

- Gradient descent
- Convergence analysis of gradient descent
- Accelerated gradient descent
- **Projected gradient descent**

Nondifferentiable functions

- Subgradient method
- Proximal gradient method
- Coordinate descent

Projected gradient descent

Optimization problem

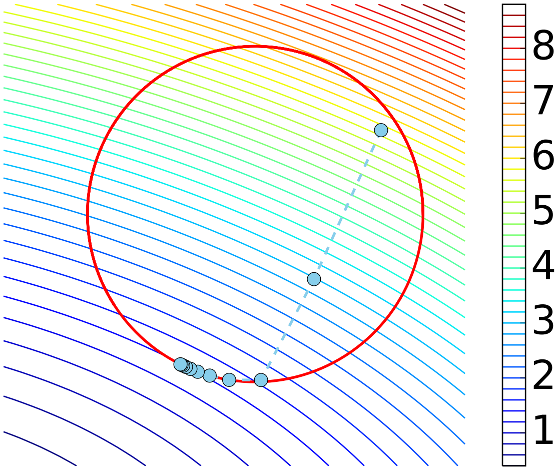
$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & x \in \mathcal{S}, \end{aligned}$$

where f is differentiable and $\mathcal{S}$ is convex

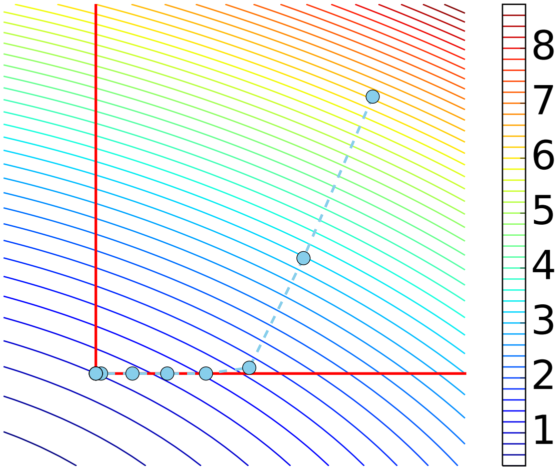
Projected-gradient-descent iteration:

$$\begin{aligned} x^{(0)} &= \text{arbitrary initialization} \\ x^{(k+1)} &= \mathcal{P}_{\mathcal{S}}\left(x^{(k)} - \alpha_k \nabla f\left(x^{(k)}\right)\right) \end{aligned}$$

Projected gradient descent



Projected gradient descent



Differentiable functions

- Gradient descent
- Convergence analysis of gradient descent
- Accelerated gradient descent
- Projected gradient descent

Nondifferentiable functions

- Subgradient method
- Proximal gradient method
- Coordinate descent

Subgradient method

Optimization problem

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

where f is convex but nondifferentiable

Subgradient-method iteration:

$$x^{(0)} = \text{arbitrary initialization}$$
$$x^{(k+1)} = x^{(k)} - \alpha_k \, q^{(k)}$$

where $q^{(k)}$ is a subgradient of f at $x^{(k)}$

Least-squares regression with ℓ_1-norm regularization

$$\text{minimize} \quad \frac{1}{2} \left|\left|Ax - y\right|\right|_2^2 + \lambda \left|\left|x\right|\right|_1$$

Sum of subgradients is a subgradient of the sum

$$q^{(k)} = A^T \left(Ax^{(k)} - y \right) + \lambda \, \text{sign} \left(x^{(k)} \right)$$

Subgradient-method iteration:

$$x^{(0)} = \text{arbitrary initialization}$$
$$x^{(k+1)} = x^{(k)} - \alpha_k \left(A^T \left(Ax^{(k)} - y \right) + \lambda \, \text{sign} \left(x^{(k)} \right) \right)$$

Convergence of subgradient method

It is not a descent method

Convergence rate can be shown to be $\mathcal{O}\left(1/\epsilon^2\right)$

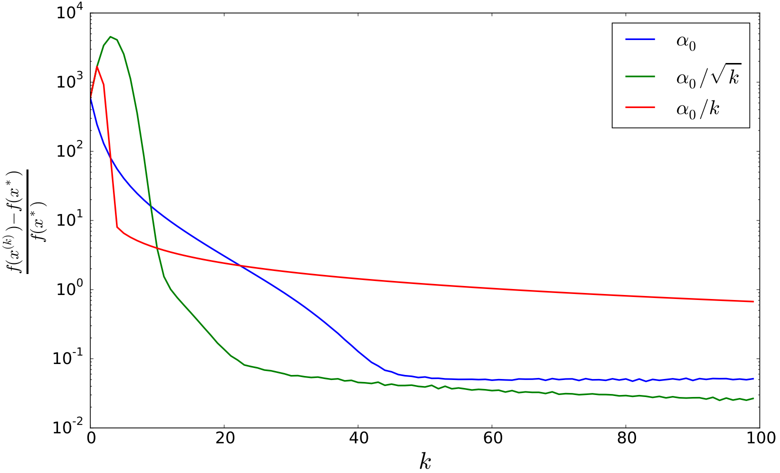
Diminishing step sizes are necessary for convergence

Experiment:

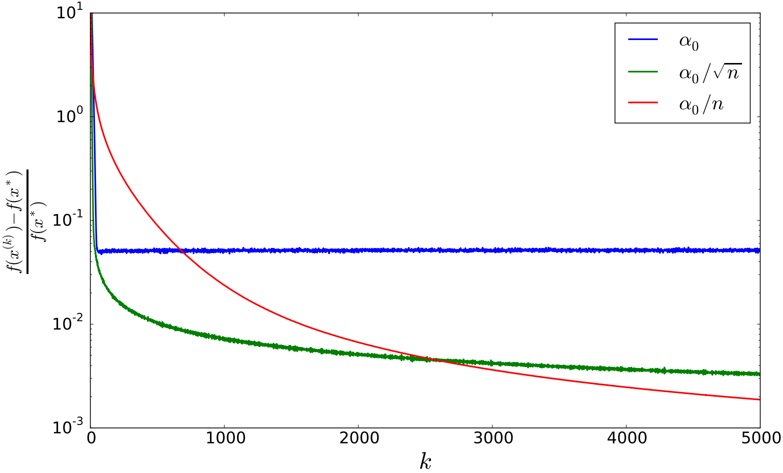
$$\text{minimize} \quad \frac{1}{2}\left|\left|Ax - y\right|\right|_2^2 + \lambda \left|\left|x\right|\right|_1$$

$A \in \mathbb{R}^{2000\times 1000}$, $y = Ax_0 + z$ where x_0 is 100-sparse and z is iid Gaussian

Convergence of subgradient method



Convergence of subgradient method



Differentiable functions

- Gradient descent
- Convergence analysis of gradient descent
- Accelerated gradient descent
- Projected gradient descent

Nondifferentiable functions

- Subgradient method
- Proximal gradient method
- Coordinate descent

Composite functions

Interesting class of functions for data analysis

$$f(x) + g(x)$$

f convex and differentiable, g convex but not differentiable

Example:

Least-squares regression (f) + ℓ_1-norm regularization (g)

$$\frac{1}{2}\left|\left|Ax - y\right|\right|_2^2 + \lambda\left|\left|x\right|\right|_1$$

Interpretation of gradient descent

Solution of local first-order approximation

$$
\begin{aligned}
x^{(k+1)} &:= x^{(k)} - \alpha_k \nabla f\left(x^{(k)}\right) \\
&= \arg\min_x \left|\left|x - \left(x^{(k)} - \alpha_k \nabla f\left(x^{(k)}\right)\right)\right|\right|_2^2 \\
&= \arg\min_x f\left(x^{(k)}\right) + \nabla f\left(x^{(k)}\right)^T \left(x - x^{(k)}\right) + \frac{1}{2\,\alpha_k} \left|\left|x - x^{(k)}\right|\right|_2^2
\end{aligned}
$$

Proximal gradient method

Idea: Minimize local first-order approximation $+ g$

$$
\begin{aligned}
x^{(k+1)} &= \arg\min_x f\left(x^{(k)}\right) + \nabla f\left(x^{(k)}\right)^T\left(x - x^{(k)}\right) + \frac{1}{2\,\alpha_k}\left|\left|x - x^{(k)}\right|\right|_2^2 \\
&\qquad + g(x) \\
&= \arg\min_x \frac{1}{2}\left|\left|x - \left(x^{(k)} - \alpha_k \nabla f\left(x^{(k)}\right)\right)\right|\right|_2^2 + \alpha_k\, g(x) \\
&= \operatorname{prox}_{\alpha_k g}\left(x^{(k)} - \alpha_k \nabla f\left(x^{(k)}\right)\right)
\end{aligned}
$$

Proximal operator:

$$
\operatorname{prox}_g(y) := \arg\min_x g(x) + \frac{1}{2}\,||y - x||_2^2
$$

Proximal gradient method

Method to solve the optimization problem

$$\text{minimize} \quad f(x) + g(x),$$

where f is differentiable and prox_g is tractable

Proximal-gradient iteration:

$$x^{(0)} = \text{arbitrary initialization}$$
$$x^{(k+1)} = \text{prox}_{\alpha_k\, g}\left(x^{(k)} - \alpha_k \nabla f\left(x^{(k)}\right)\right)$$

Interpretation as a fixed-point method

A vector $\hat{x}$ is a solution to

$$\text{minimize} \quad f(x) + g(x),$$

if and only if it is a fixed point of the proximal-gradient iteration for any $\alpha > 0$

$$\hat{x} = \text{prox}_{\alpha_k g}\left(\hat{x} - \alpha_k \nabla f(\hat{x})\right)$$

Projected gradient descent as a proximal method

The proximal operator of the indicator function

$$\mathcal{I}_{\mathcal{S}}(x) := \begin{cases} 0 & \text{if } x \in \mathcal{S}, \\ \infty & \text{if } x \notin \mathcal{S}. \end{cases}$$

of a convex set $\mathcal{S} \subseteq \mathbb{R}^n$ is projection onto $\mathcal{S}$

Proximal-gradient iteration:

$$\begin{aligned} x^{(k+1)} &= \operatorname{prox}_{\alpha_k \mathcal{I}_{\mathcal{S}}}\left(x^{(k)} - \alpha_k \nabla f\left(x^{(k)}\right)\right) \\ &= \mathcal{P}_{\mathcal{S}}\left(x^{(k)} - \alpha_k \nabla f\left(x^{(k)}\right)\right) \end{aligned}$$

Proximal operator of ℓ_1 norm

The proximal operator of the ℓ_1 norm is the soft-thresholding operator

$$\text{prox}_{\beta\,||\cdot||_1}(y) = \mathcal{S}_\beta(y)$$

where $\beta > 0$ and

$$\mathcal{S}_\beta(y)_i := \begin{cases} y_i - \text{sign}(y_i)\,\beta & \text{if } |y_i| \geq \beta \\ 0 & \text{otherwise} \end{cases}$$

Iterative Shrinkage-Thresholding Algorithm (ISTA)

The proximal gradient method for the problem

$$\text{minimize} \quad \frac{1}{2} \left|\left|Ax - y\right|\right|_2^2 + \lambda \left|\left|x\right|\right|_1$$

is called ISTA

ISTA iteration:

$$\begin{aligned} x^{(0)} &= \text{arbitrary initialization} \\ x^{(k+1)} &= \mathcal{S}_{\alpha_k \lambda} \left(x^{(k)} - \alpha_k A^T \left(Ax^{(k)} - y \right) \right) \end{aligned}$$

Fast Iterative Shrinkage-Thresholding Algorithm (FISTA)

ISTA can be accelerated using Nesterov's accelerated gradient method

FISTA iteration:

$$
\begin{aligned}
x^{(0)} &= \text{arbitrary initialization} \\
z^{(0)} &= x^{(0)} \\
x^{(k+1)} &= \mathcal{S}_{\alpha_k \lambda}\left(z^{(k)} - \alpha_k A^T \left(Az^{(k)} - y\right)\right) \\
z^{(k+1)} &= x^{(k+1)} + \frac{k}{k+3}\left(x^{(k+1)} - x^{(k)}\right)
\end{aligned}
$$

Convergence of proximal gradient method

Without acceleration:

- Descent method
- Convergence rate can be shown to be $\mathcal{O}(1/\epsilon)$ with constant step or backtracking line search

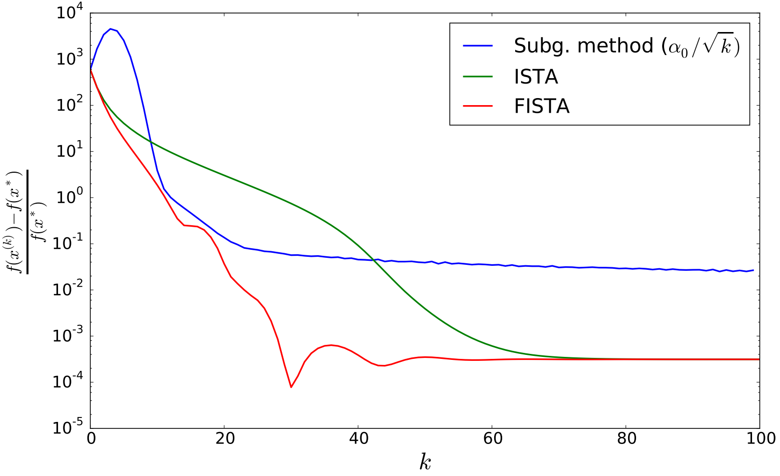
With acceleration:

- Not a descent method
- Convergence rate can be shown to be $\mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\right)$ with constant step or backtracking line search

Experiment: minimize $\quad \frac{1}{2}\left|\left|Ax - y\right|\right|_2^2 + \lambda \left|\left|x\right|\right|_1$

$A \in \mathbb{R}^{2000\times 1000}$, $y = Ax_0 + z$, x_0 100-sparse and z iid Gaussian

Convergence of proximal gradient method



Differentiable functions

- Gradient descent
- Convergence analysis of gradient descent
- Accelerated gradient descent
- Projected gradient descent

Nondifferentiable functions

- Subgradient method
- Proximal gradient method
- Coordinate descent

Coordinate descent

Idea: Solve the n-dimensional problem

$$\text{minimize} \quad h(x_1, x_2, \ldots, x_n)$$

by solving a sequence of 1D problems

Coordinate-descent iteration:

$$x^{(0)} = \text{arbitrary initialization}$$
$$x_i^{(k+1)} = \arg\min_{\alpha} h\left(x_1^{(k)}, \ldots, \alpha, \ldots, x_n^{(k)}\right) \quad \text{for some } 1 \leq i \leq n$$

Coordinate descent

Convergence is guaranteed for functions of the form

$$f(x) + \sum_{i=1}^{n} g_i(x_i)$$

where f is convex and differentiable and $g_1, \ldots, g_n$ are convex

Least-squares regression with ℓ_1-norm regularization

$$h(x) := \frac{1}{2} \|Ax - y\|_2^2 + \lambda \|x\|_1$$

The solution to the subproblem $\min_{x_i} h(x_1, \ldots, x_i, \ldots, x_n)$ is

$$\hat{x}_i = \frac{\mathcal{S}_\lambda(\gamma_i)}{\|A_i\|_2^2}$$

where A_i is the ith column of A and

$$\gamma_i := \sum_{l=1}^{m} A_{li} \left(y_l - \sum_{j \neq i} A_{lj} x_j \right)$$

Computational experiments

Table 5.1 *Lasso for linear regression: Average (standard error) of CPU times over ten realizations, for coordinate descent, generalized gradient, and Nesterov's momentum methods. In each case, time shown is the total time over a path of 20 λ values.*

	$N = 10000,\ p = 100$		$N = 200,\ p = 10000$	
Correlation	0	0.5	0	0.5
Coordinate descent	0.110 (0.001)	0.127 (0.002)	0.298 (0.003)	0.513 (0.014)
Proximal gradient	0.218 (0.008)	0.671 (0.007)	1.207 (0.026)	2.912 (0.167)
Nesterov	0.251 (0.007)	0.604 (0.011)	1.555 (0.049)	2.914 (0.119)

From *Statistical Learning with Sparsity The Lasso and Generalizations*
by Hastie, Tibshirani and Wainwright