Subgradient Descent

David Rosenberg

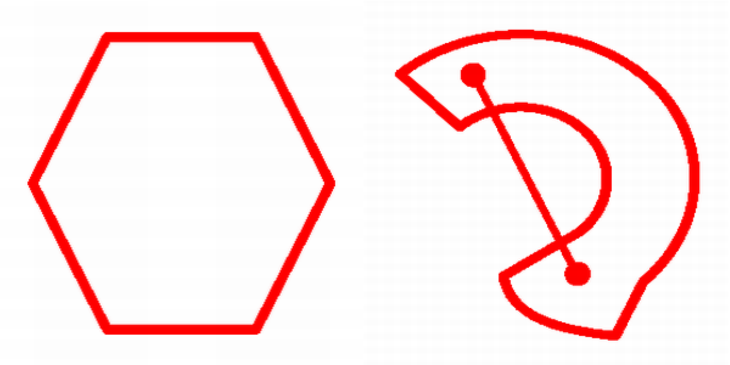
New York University

October 29, 2016

Convex Sets

Definition

A set C is **convex** if the line segment between any two points in C lies in C.

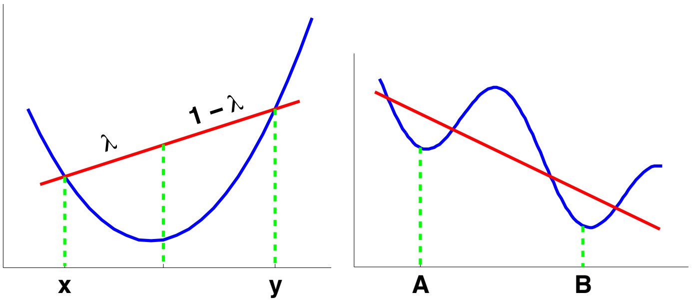


KPM Fig. 7.4

Convex and Concave Functions

Definition

A function $f : \mathbf{R}^n \to \mathbf{R}$ is **convex** if the line segment connecting any two points on the graph of f lies above the graph. f is **concave** if $-f$ is convex.

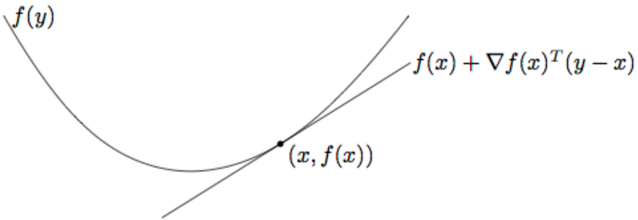


KPM Fig. 7.5

First-Order Approximation

- Suppose $f : \mathbf{R}^n \to \mathbf{R}$ is **differentiable**
- Suppose we know $f(x)$ and $\nabla f(x)$.
- What can we say about $f(y)$, when y is near x?
- We have the following linear approximation:

$$f(y) \approx f(x) + \nabla f(x)^T (y - x)$$



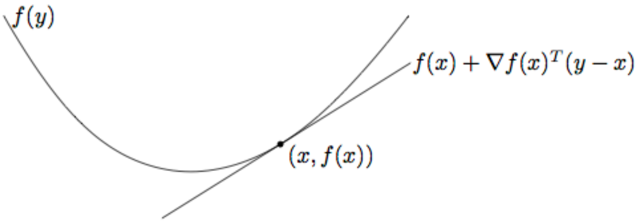
Boyd & Vandenberghe Fig. 3.2

First-Order Condition for Convex, Differentiable Function

- Suppose $f : \mathbf{R}^n \to \mathbf{R}$ is **convex** and **differentiable**
- Then for any $x, y \in \mathbf{R}^n$

$$f(y) \geqslant f(x) + \nabla f(x)^T (y - x)$$

- The linear approximation to f at x is a **global underestimator** of f:



Boyd & Vandenberghe Fig. 3.2

First-Order Condition for Convex, Differentiable Function

- Suppose $f : \mathbf{R}^n \to \mathbf{R}$ is **convex** and **differentiable**
- Then for any $x, y \in \mathbf{R}^n$

$$f(y) \geqslant f(x) + \nabla f(x)^T (y - x)$$

Corollary

If $\nabla f(x) = 0$ then x is a global minimizer of f.

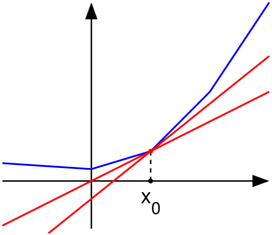
Subgradients

Definition

A vector $g \in \mathbf{R}^n$ is a **subgradient** of $f : \mathbf{R}^n \to \mathbf{R}$ at x if for all z,

$$f(z) \geqslant f(x) + g^T(z - x).$$

- g is a subgradient iff $f(x) + g^T(z - x)$ is a global underestimator of f



Subdifferential

Definitions

- f is **subdifferentiable** at x if $\exists$ at least one subgradient at x.
- The set of all subgradients at x is called the **subdifferential:** $\partial f(x)$

Basic Facts

- If f is convex and differentiable, then $\nabla f(x)$ is the unique subgradient of f at x.
- Any point x, there can be 0, 1, or infinitely many subgradients.
  - Can only be 0 for non-convex f.

Globla Optimality Condition

Definition

A vector $g \in \mathbf{R}^n$ is a **subgradient** of $f : \mathbf{R}^n \to \mathbf{R}$ at x if for all z,

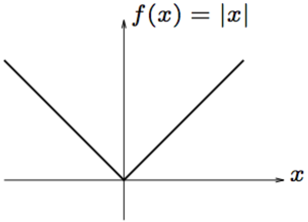
$$f(z) \geqslant f(x) + g^T(z - x).$$

Corollary

If $0 \in \partial f(x)$, then x is a ***global minimizer*** *of f.*

Subdifferential of Absolute Value

- Consider $f(x) = |x|$



- Plot on right shows $\cup\{(x,g) \mid x \in \mathbf{R}, g \in \partial f(x)\}$
- See B&V's notes for more: `http://web.stanford.edu/class/ee364b/lectures/subgradients_notes.pdf`

Boyd EE364b: Subgradients Slides

Subgradient Descent

Subgradient Descent

- Initialize $x = 0$
  - repeat
    - $x \leftarrow x - \eta g$ for $g \in \partial f(x)$ and η chosen according to **step size rule**
  - until stopping criterion satisfied

- Note: Not necessarily a "**descent method**"
  - in a descent method, every step is an improvement
- Always keep track of the best x we've seen as we go

Step Size

- Because not a descent method, can't adaptive step size
  - i.e. we don't use backtracking line search.
- Need to determine step sizes in advance
- Two main choices:
  1. Fixed step size
  2. Step sizes decrease according to Robbins-Monro Conditions:

$$\sum_{t=1}^{\infty} \eta_t^2 < \infty \qquad \sum_{t=1}^{\infty} \eta_t = \infty$$

  - e.g. $\eta_t = 1/t$.

Convergence Theorem for Fixed Step Size

Assume $f : \mathbf{R}^n \to \mathbf{R}$ is convex and

- f is Lipschitz continuous with constant $G > 0$:

$$|f(x) - f(y)| \leqslant G\|x - y\| \text{ for all } x, y$$

Theorem

For fixed step size η, *subgradient method satisfies:*

$$\lim_{k\to\infty} f(x_{best}^{(k)}) \leqslant f(x^*) + G^2 t/2$$

Based on https://www.cs.cmu.edu/~ggordon/10725-F12/slides/06-sg-method.pdf

Convergence Theorems for Decreasing Step Sizes

Assume $f : \mathbf{R}^n \to \mathbf{R}$ is convex and

- f is Lipschitz continuous with constant $G > 0$:

$$|f(x) - f(y)| \leqslant G\|x - y\| \text{ for all } x, y$$

Theorem

For step size respecting Robbins-Monro conditions,

$$\lim_{k \to \infty} f(x_{best}^{(k)}) \leqslant f(x^*)$$

Based on `https://www.cs.cmu.edu/~ggordon/10725-F12/slides/06-sg-method.pdf`

Coordinate Subdifferential of Lasso Objective

- Lasso objective:

$$\min_{w \in \mathbf{R}^d} \sum_{i=1}^{n} \left(w^T x_i - y_i\right)^2 + \lambda |w|_1$$

- Partial derivative of empirical risk (homework):

$$\frac{\partial}{\partial w_k} \sum_{i=1}^{n} \left(w^T x_i - y_i\right)^2 = a_k w_k - c_k$$

where

$$a_j = 2 \sum_{i=1}^{n} x_{ij}^2 \qquad\qquad c_j = 2 \sum_{i=1}^{n} x_{ij} \left(y_i - w_{-j}^T x_{i,-j}\right)$$

Coordinate Subdifferential of Lasso Objective

- Subdifferential of $|w|_1$:

$$\partial_{w_k} \lambda |w| = \begin{cases} -\lambda & w_k < 0 \\ \lambda & w_k > 0 \\ [-\lambda, \lambda] & w_k = 0 \end{cases}$$

- So subdifferential of objective is:

$$\partial_{w_k} (\text{Lasso Objective}) = \begin{cases} a_k w_k - c_k - \lambda & w_k < 0 \\ a_k w_k - c_k + \lambda & w_k > 0 \\ [-c_k - \lambda, -c_k + \lambda] & w_k = 0 \end{cases}$$

Coordinate Subdifferential of Lasso Objective

- Solving for $0 \in \partial_{w_k}(\text{Lasso Objective})$:
  - **Case 1**: $w_k < 0$:

    $$a_k w_k - c_k - \lambda = 0 \implies w_k = (c_k + \lambda)/a_k$$

    So if $c_k < -\lambda$, then $w_k = (c_k + \lambda)/a_k$ is a critical point
  - **Case 2**: $w_k > 0$: If $c_k > \lambda$ then $w_k = (c_k - \lambda)/a_k$ is a critical point
  - **Case 3**: $w_k = 0$: $w_k = 0$ and $c_k \in [-\lambda, \lambda] \implies 0 \in [-c_k - \lambda, -c_k + \lambda]$ so $w_k = 0$ is a critical point
- So $0 \in \partial_{w_k}(\text{Lasso Objective})$ iff

$$w_j(c_j) = \begin{cases} (c_j + \lambda)/a_j & \text{if } c_j < -\lambda \\ 0 & \text{if } c_j \in [-\lambda, \lambda] \\ (c_j - \lambda)/a_j & \text{if } c_j > \lambda \end{cases}$$