# Subgradient Descent (Continued)

David S. Rosenberg

New York University

February 13, 2018

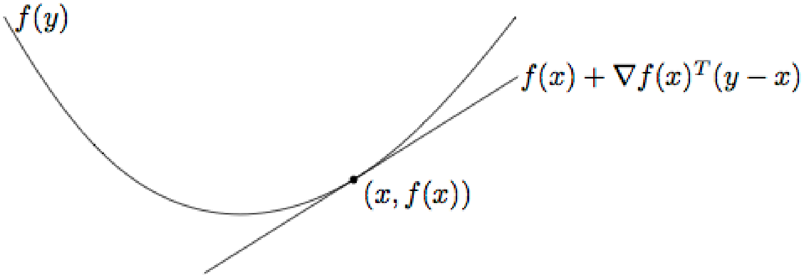
# Contents

# Subgradients: Recap

# First-Order Approximation

- Suppose $f : \mathbf{R}^d \to \mathbf{R}$ is **differentiable.**
- Predict $f(y)$ given $f(x)$ and $\nabla f(x)$?
- Linear (i.e. "**first order**") 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 $f : \mathbf{R}^d \to \mathbf{R}$ is **convex** and **differentiable.**
- Then for any $x, y \in \mathbf{R}^d$

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

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

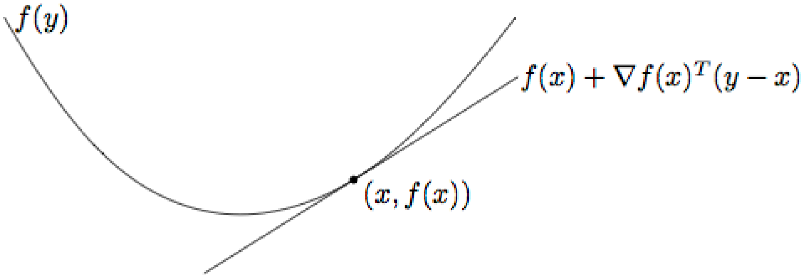


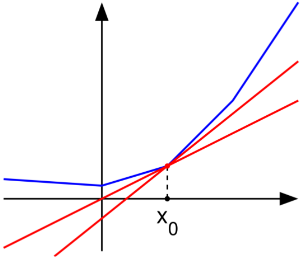

Figure from Boyd & Vandenberghe Fig. 3.2; Proof in Section 3.1.3

# Subgradients

## Definition

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

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




Blue is a graph of $f(x)$.
Each red line $x \mapsto f(x_0) + g^T(x - x_0)$ is a global lower bound on $f(x)$.

# 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

- $f$ is convex and differentiable at $x \implies \partial f(x) = \{\nabla f(x)\}$.
- At any point $x$, there can be 0, 1, or infinitely many subgradients.
- $\partial f(x) = \emptyset \implies f$ is not convex.

# Subgradients give Ascent Directions

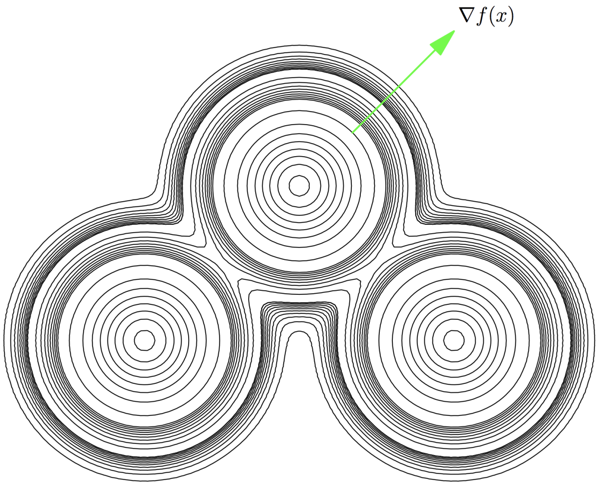
# Contour Lines and Gradients

- For function $f : \mathbf{R}^d \to \mathbf{R}$,
  - **graph** of function lives in $\mathbf{R}^{d+1}$,
  - **gradient** and **subgradient** of $f$ live in $\mathbf{R}^d$, and
  - **contours**, **level sets,** and **sublevel sets** are in $\mathbf{R}^d$.
- $f : \mathbf{R}^d \to \mathbf{R}$ continuously differentiable, $\nabla f(x_0) \neq 0$, then $\nabla f(x_0)$ normal to level set

$$S = \left\{ x \in \mathbf{R}^d \mid f(x) = f(x_0) \right\}.$$

- Proof sketch in notes.

# Gradient orthogonal to sublevel sets




Plot courtesy of Brett Bernstein.

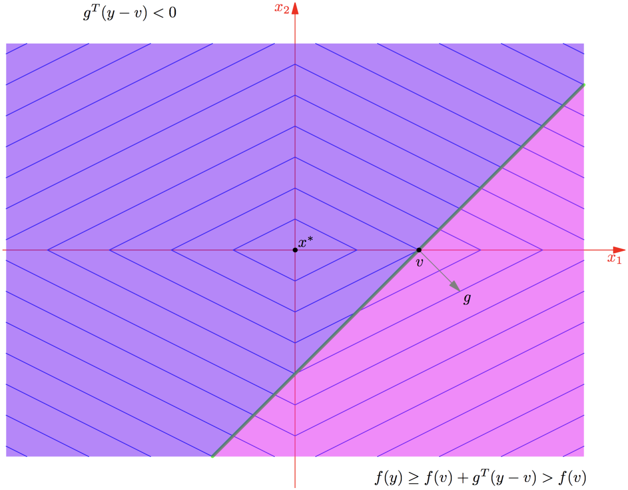
# Contour Lines and Subgradients

- A hyperplane $H$ **supports** a set $S$ if $H$ intersects $S$ and all of $S$ lies one one side of $H$.
- If $f : \mathbf{R}^d \to \mathbf{R}$ has subgradient $g$ at $x_0$, then the hyperplane $H$ orthogonal to $g$ at $x_0$ must **support** the level set $S = \left\{x \in \mathbf{R}^d \mid f(x) = f(x_0)\right\}$.

Proof:

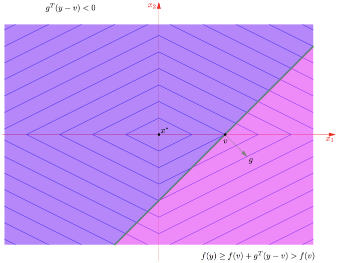
- For any $y$, we have $f(y) \geqslant f(x_0) + g^T(y - x_0)$. (def of subgradient)
- If $y$ is strictly on side of $H$ that $g$ points in,
  - then $g^T(y - x_0) > 0$.
  - So $f(y) > f(x_0)$.
  - So $y$ is not in the level set $S$.
- $\therefore$ All elements of $S$ must be on $H$ or on the $-g$ side of $H$.

# Subgradient of $f(x_1, x_2) = |x_1| + 2|x_2|$




Plot courtesy of Brett Bernstein.

# Subgradient of $f(x_1, x_2) = |x_1| + 2|x_2|$




- Points on $g$ side of $H$ have larger $f$-values than $f(x_0)$. (from proof)
- But points on $-g$ side may **not** have smaller $f$-values.
- So $-g$ may **not** be a descent direction. (shown in figure)

Plot courtesy of Brett Bernstein.

# Subgradient Descent

# Subgradient Descent

- Suppose $f$ is convex, and we start optimizing at $x_0$.
- Repeat
  - Step in a negative subgradient direction:

    $$x = x_0 - tg,$$

    where $t > 0$ is the step size and $g \in \partial f(x_0)$.
- $-g$ not a descent direction – can this work?

# Subgradient Gets Us Closer To Minimizer

**Theorem**

*Suppose $f$ is convex.*

- *Let $x = x_0 - tg$, for $g \in \partial f(x_0)$.*
- *Let $z$ be any point for which $f(z) < f(x_0)$.*
- *Then for small enough $t > 0$,*
$$\|x - z\|_2 < \|x_0 - z\|_2.$$

- Apply this with $z = x^* \in \arg\min_x f(x)$.

$\Longrightarrow$ **Negative subgradient step gets us closer to minimizer.**

# Subgradient Gets Us Closer To Minimizer (Proof)

- Let $x = x_0 - tg$, for $g \in \partial f(x_0)$ and $t > 0$.
- Let $z$ be any point for which $f(z) < f(x_0)$.
- Then

$$\begin{aligned}
\|x - z\|_2^2 &= \|x_0 - tg - z\|_2^2 \\
&= \|x_0 - z\|_2^2 - 2tg^T(x_0 - z) + t^2\|g\|_2^2 \\
&\leqslant \|x_0 - z\|_2^2 - 2t\left[f(x_0) - f(z)\right] + t^2\|g\|_2^2
\end{aligned}$$

- Consider $-2t\left[f(x_0) - f(z)\right] + t^2\|g\|_2^2$.
  - It's a convex quadratic (facing upwards).
  - Has zeros at $t = 0$ and $t = 2\left(f(x_0) - f(z)\right)/\|g\|_2^2 > 0$.
  - Therefore, it's negative for any

$$t \in \left(0, \frac{2\left(f(x_0) - f(z)\right)}{\|g\|_2^2}\right).$$

Based on Boyd EE364b: Subgradients Slides

# 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 $t$, 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,*

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

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