

# Subgradient Descent

Xintian Han & David S. Rosenberg

NYU CDS

Feb 13, 2019

# Contents

- 1 Motivation and Review: Lasso
- 2 Convexity and Sublevel Sets
- 3 Convex and Differentiable Functions
- 4 Subgradients
- 5 Subgradient Descent

## Motivation and Review: Lasso

# The Lasso Problem

- Lasso problem can be parametrized as

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

- We could solve Lasso by Shooting Method and Projected SGD.
- How about using SGD?
- $\|w\|_1 = |w_1| + |w_2|$  is not differentiable!

## Gradient Descent on Lasso Objective?

- The partial gradient of the Lasso objective is

$$\nabla_w J(w) = \frac{1}{n} \sum_{j=1}^n 2\{w^T x_j - y_j\} x_j + \lambda \cdot \text{sign}(w)$$

when  $w_i \neq 0$  for all  $i$ , and **otherwise is undefined**.

Potential arguments for why we shouldn't care about the points of nondifferentiability:

- If we start with a random  $w$ , will we ever hit exactly  $w_i = 0$ ?
- If we did, could we perturb the step size by  $\epsilon$  to miss such a point?
- Does it even make sense to check  $w_i = 0$  with floating point numbers?

## Gradient Descent on Lasso Objective?

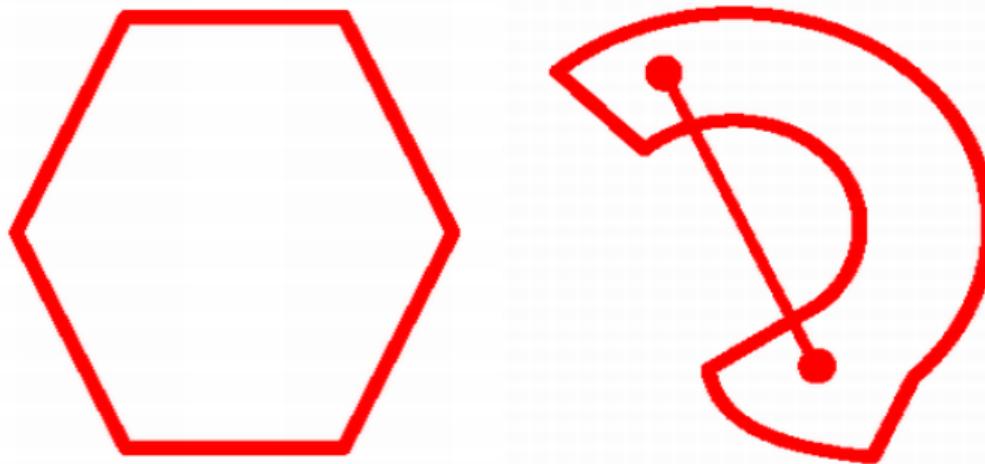
- If we blindly apply gradient descent from a random starting point
  - seems unlikely that we'll hit a point where the gradient is undefined.
  - So it's clear how we could apply gradient descent or SGD to a function that's not differentiable everywhere
- Still, doesn't mean that gradient descent will work if objective not differentiable!
- Theory of subgradients and subgradient descent will clear up any uncertainty.

## Convexity and Sublevel Sets

# Convex Sets

## Definition

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

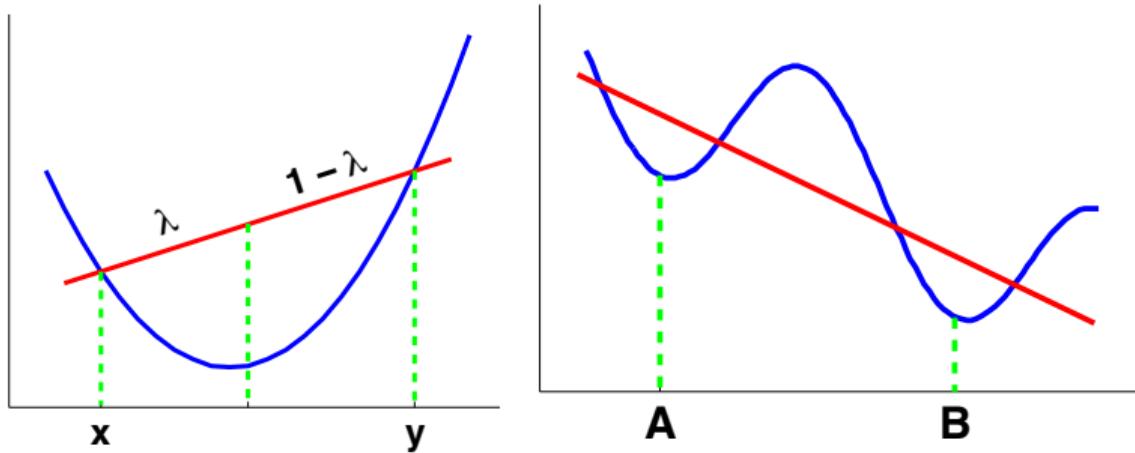


KPM Fig. 7.4

# Convex and Concave Functions

## Definition

A function  $f : \mathbf{R}^d \rightarrow \mathbf{R}$  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

# Examples of Convex Functions on $\mathbf{R}$

## Examples

- $x \mapsto ax + b$  is both convex and concave on  $\mathbf{R}$  for all  $a, b \in \mathbf{R}$ .
- $x \mapsto |x|^p$  for  $p \geq 1$  is convex on  $\mathbf{R}$
- $x \mapsto e^{ax}$  is convex on  $\mathbf{R}$  for all  $a \in \mathbf{R}$
- Every norm on  $\mathbf{R}^n$  is convex (e.g.  $\|x\|_1$  and  $\|x\|_2$ )
- Max:  $(x_1, \dots, x_n) \mapsto \max\{x_1, \dots, x_n\}$  is convex on  $\mathbf{R}^n$

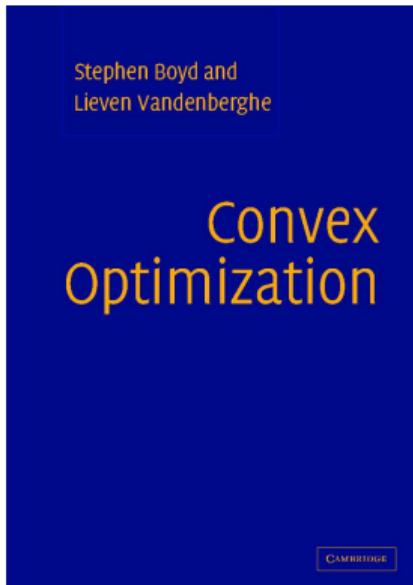
# Simple Composition Rules

## Examples

- If  $g$  is convex, and  $Ax + b$  is an affine mapping, then  $g(Ax + b)$  is convex.
- If  $g$  is convex then  $\exp g(x)$  is convex.
- If  $g$  is convex and nonnegative and  $p \geq 1$  then  $g(x)^p$  is convex.
- If  $g$  is concave and positive then  $\log g(x)$  is concave
- If  $g$  is concave and positive then  $1/g(x)$  is convex.

# Main Reference for Convex Optimization

- Boyd and Vandenberghe (2004)
  - Very clearly written, but has a ton of detail for a first pass.
  - See the [Extreme Abridgement of Boyd and Vandenberghe](#).



# Level Sets and Sublevel Sets

Let  $f : \mathbf{R}^d \rightarrow \mathbf{R}$  be a function. Then we have the following definitions:

## Definition

A **level set** or **contour line** for the value  $c$  is the set of points  $x \in \mathbf{R}^d$  for which  $f(x) = c$ .

## Definition

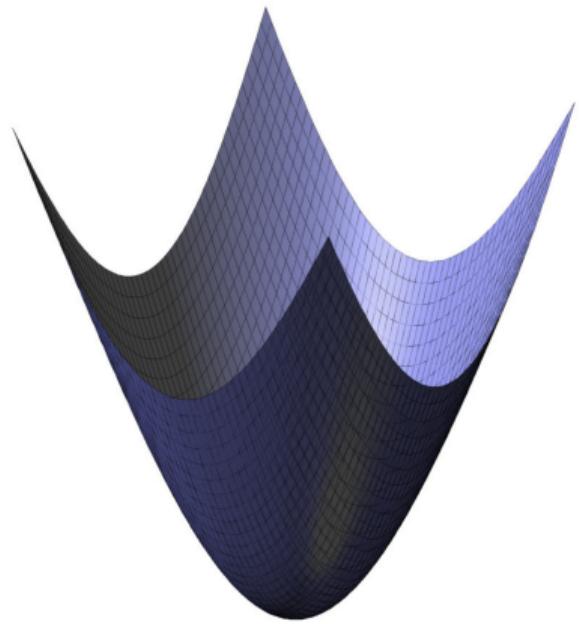
A **sublevel** set for the value  $c$  is the set of points  $x \in \mathbf{R}^d$  for which  $f(x) \leq c$ .

## Theorem

*If  $f : \mathbf{R}^d \rightarrow \mathbf{R}$  is convex, then the sublevel sets are convex.*

(Proof straight from definitions.)

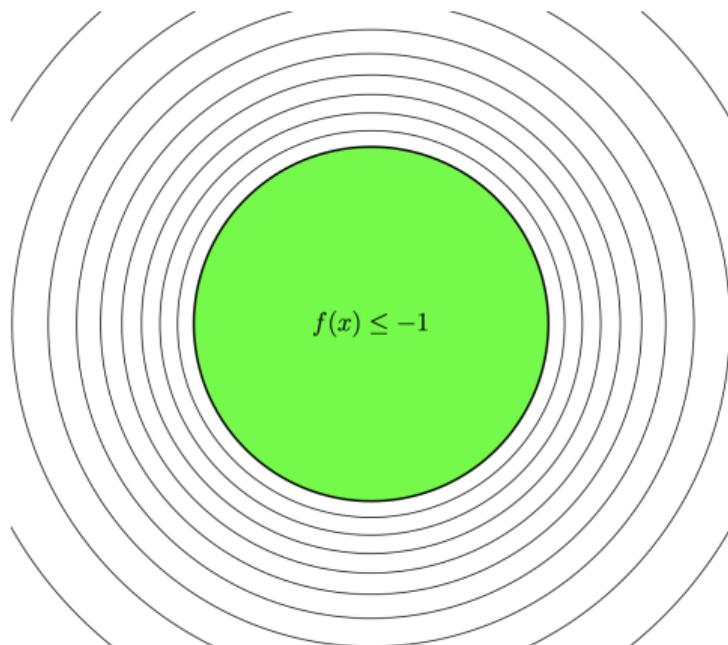
# Convex Function



---

Plot courtesy of Brett Bernstein.

# Contour Plot Convex Function: Sublevel Set

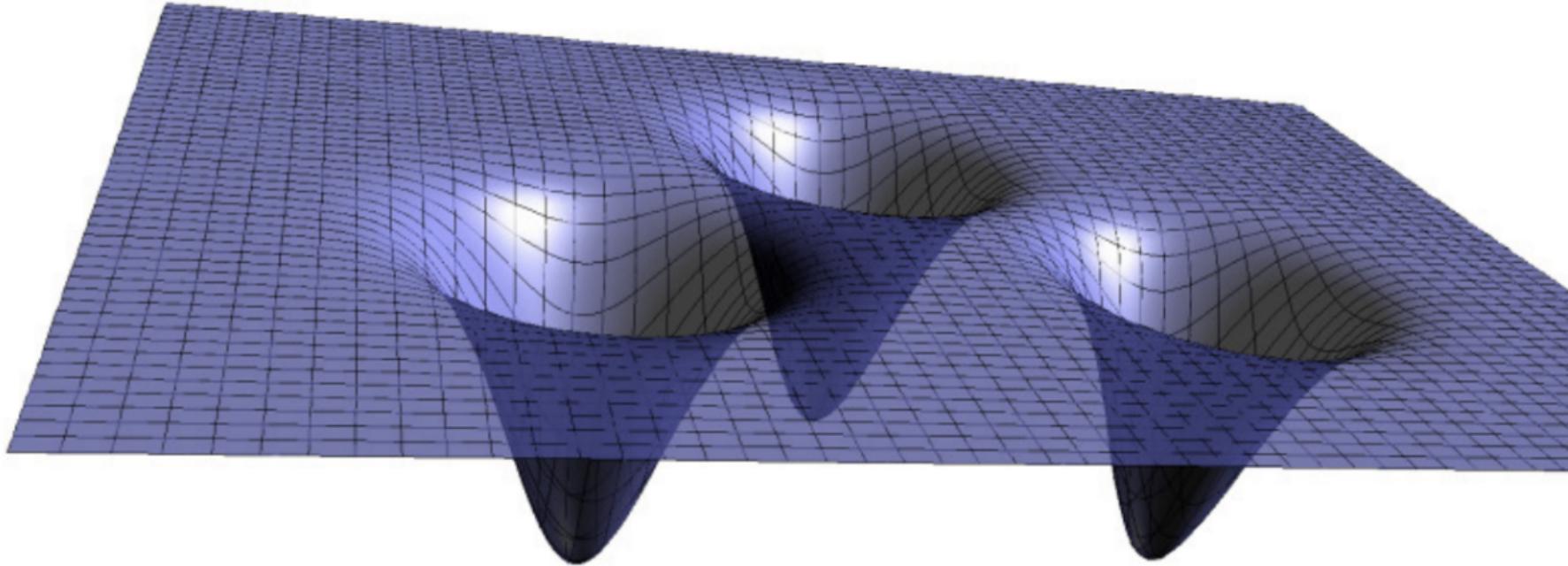


Is the sublevel set  $\{x \mid f(x) \leq 1\}$  convex?

---

Plot courtesy of Brett Bernstein.

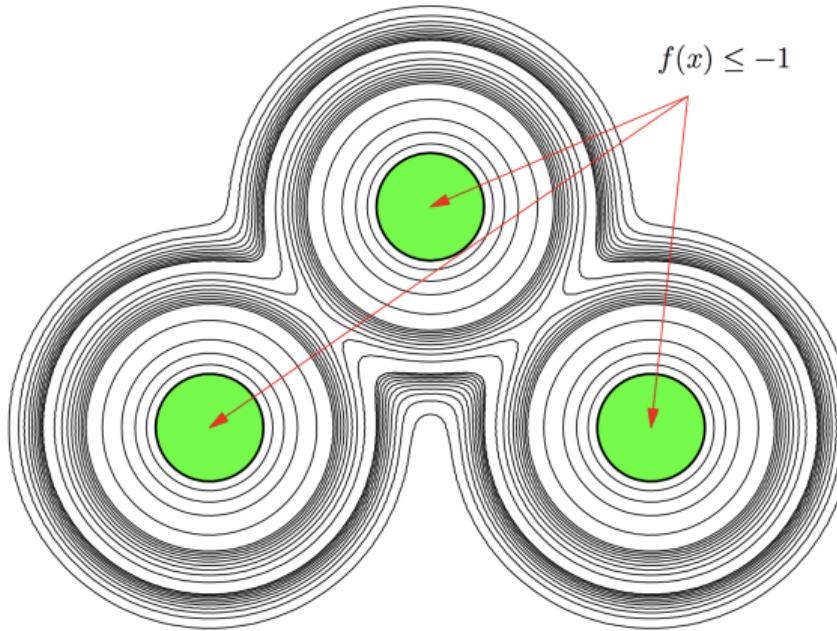
# Nonconvex Function



---

Plot courtesy of Brett Bernstein.

# Contour Plot Nonconvex Function: Sublevel Set

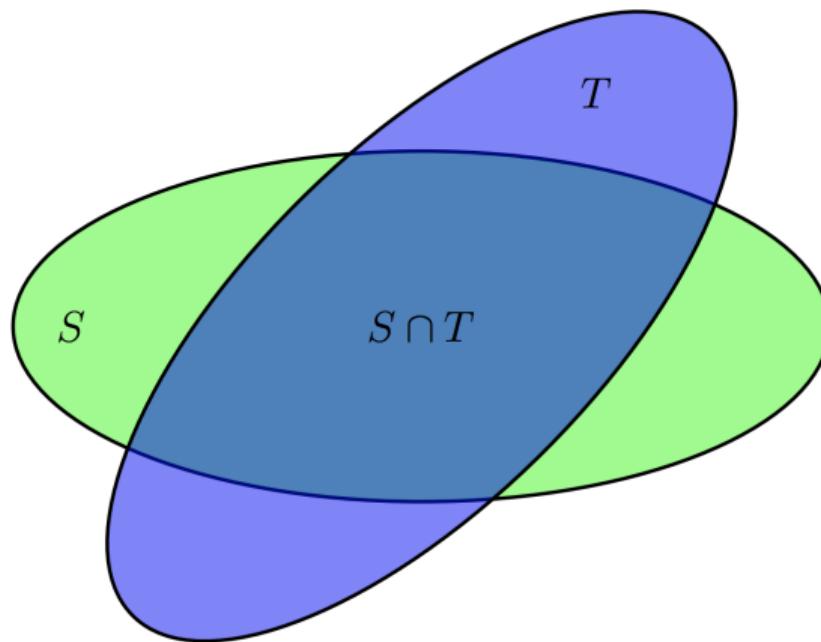


Is the sublevel set  $\{x \mid f(x) \leq 1\}$  convex?

---

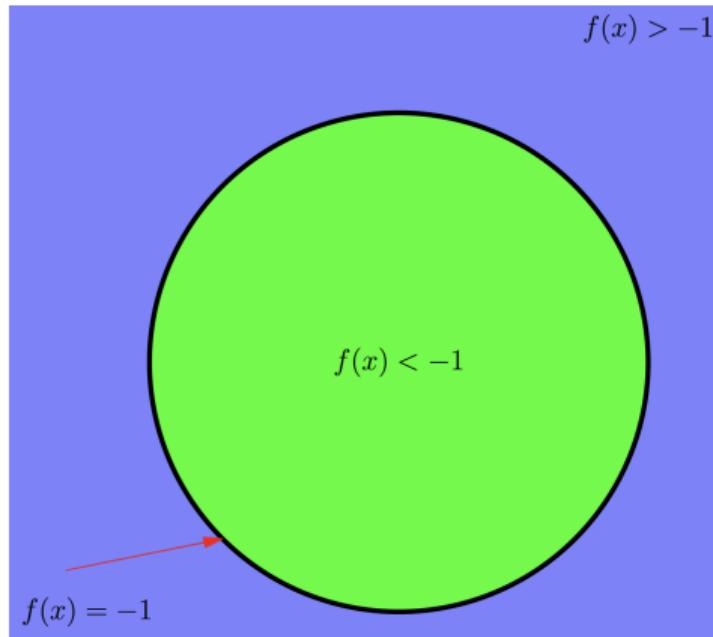
Plot courtesy of Brett Bernstein.

## Fact: Intersection of Convex Sets is Convex



Plot courtesy of Brett Bernstein.

## Level and Superlevel Sets



Level sets and superlevel sets of convex functions are **not** generally convex.

---

Plot courtesy of Brett Bernstein.

## Convex Optimization Problem: Standard Form

### Convex Optimization Problem: Standard Form

$$\begin{aligned} & \text{minimize} && f_0(x) \\ & \text{subject to} && f_i(x) \leq 0, \quad i = 1, \dots, m \end{aligned}$$

where  $f_0, \dots, f_m$  are convex functions.

- What can we say about each constraint set  $\{x \mid f_i(x) \leq 0\}$ ? (convex)
- What can we say about the feasible set  $\{x \mid f_i(x) \leq 0, i = 1, \dots, m\}$ ? (convex)

## Convex Optimization Problem: Implicit Form

### Convex Optimization Problem: Implicit Form

$$\begin{aligned} & \text{minimize} && f(x) \\ & \text{subject to} && x \in C \end{aligned}$$

where  $f$  is a convex function and  $C$  is a convex set.

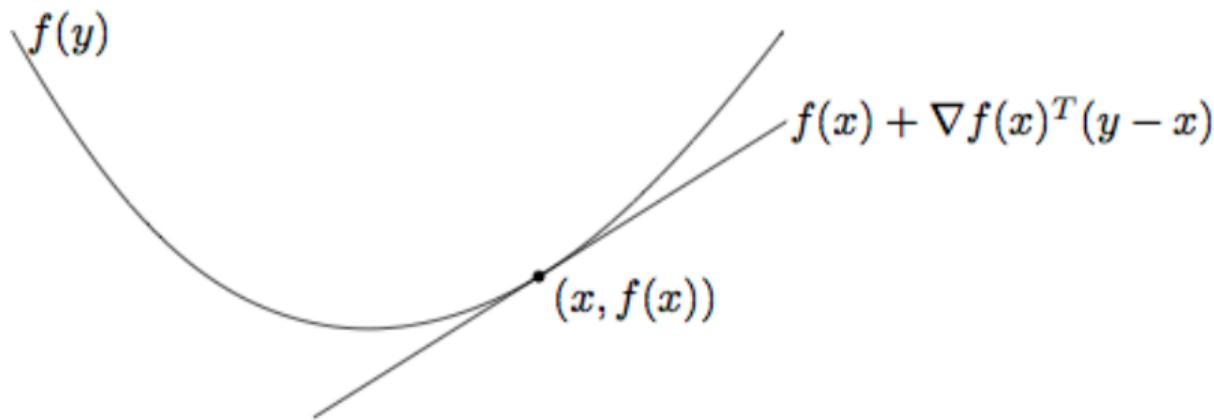
An alternative “generic” convex optimization problem.

## Convex and Differentiable Functions

# First-Order Approximation

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

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

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

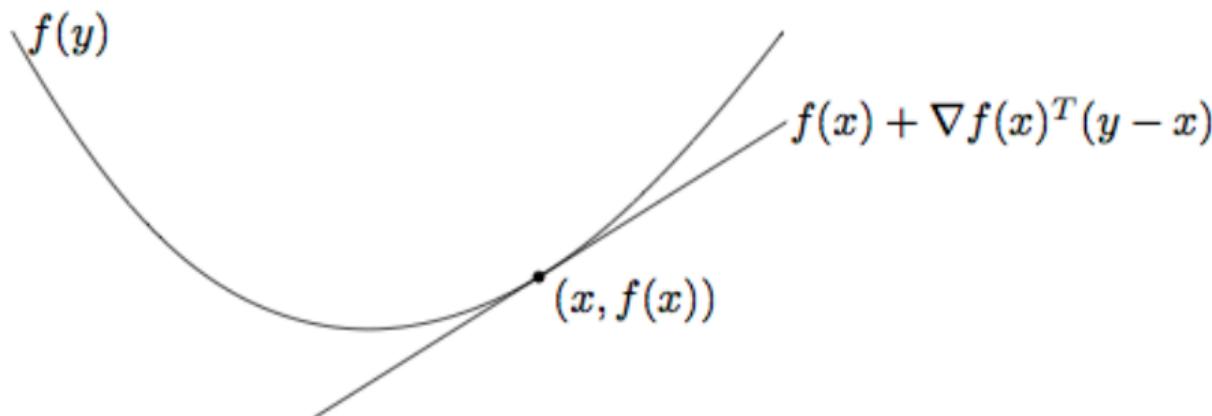


Figure from Boyd & Vandenberghe Fig. 3.2; Proof in Section 3.1.3

# First-Order Condition for Convex, Differentiable Function

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

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

## Corollary

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

For convex functions, local information gives global information.

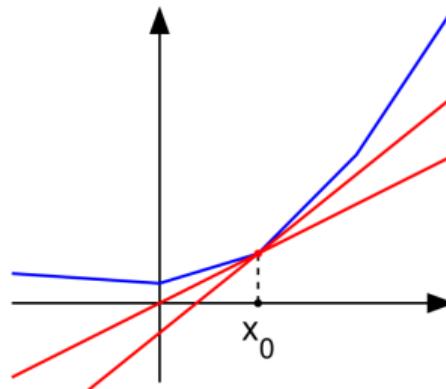
# Subgradients

# Subgradients

## Definition

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

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



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

# Globla Optimality Condition

## Definition

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

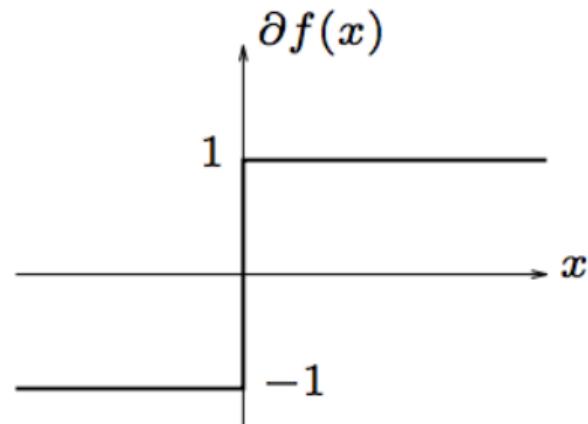
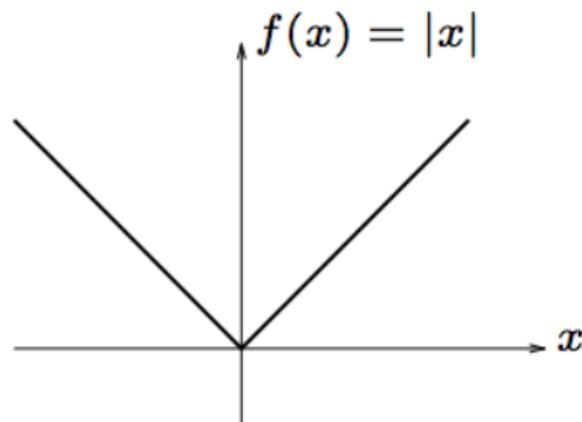
$$f(z) \geq f(x) + g^T(z - x).$$

## Corollary

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

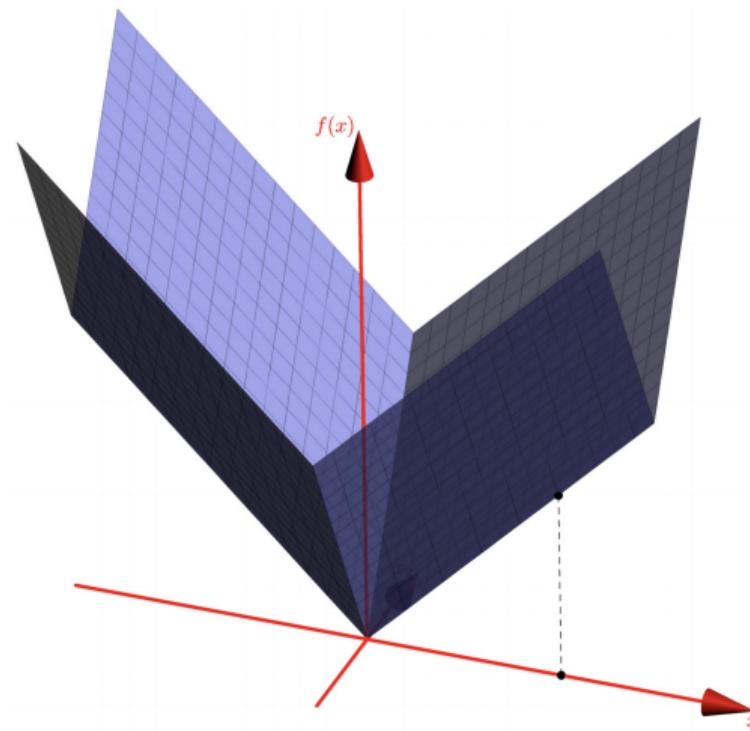
# Subdifferential of Absolute Value

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



- Plot on right shows  $\{(x, g) \mid x \in \mathbb{R}, g \in \partial f(x)\}$

$$f(x_1, x_2) = |x_1| + 2|x_2|$$



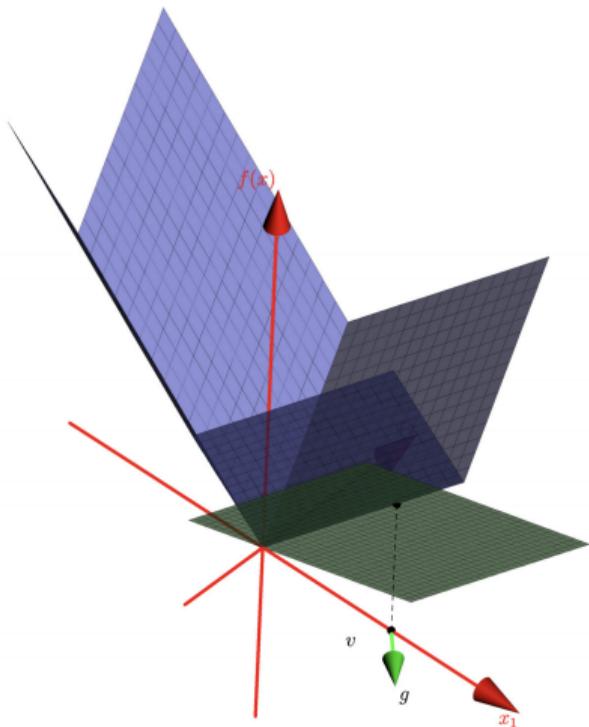
---

Plot courtesy of Brett Bernstein.

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

- Let's find the subdifferential of  $f(x_1, x_2) = |x_1| + 2|x_2|$  at  $(3, 0)$ .
- First coordinate of subgradient must be 1, from  $|x_1|$  part (at  $x_1 = 3$ ).
- Second coordinate of subgradient can be anything in  $[-2, 2]$ .
- So graph of  $h(x_1, x_2) = f(3, 0) + g^T(x_1 - 3, x_2 - 0)$  is a global underestimate of  $f(x_1, x_2)$ , for any  $g = (g_1, g_2)$ , where  $g_1 = 1$  and  $g_2 \in [-2, 2]$ .

# Underestimating Hyperplane to $f(x_1, x_2) = |x_1| + 2|x_2|$



---

Plot courtesy of Brett Bernstein.

# Important Properties of Subdifferential

- If  $f_1, \dots, f_m : \mathbf{R}^d \rightarrow \mathbf{R}$  are convex functions and  $f = f_1 + \dots + f_m$ , then  $\partial f(x) = \partial f_1(x) + \dots + \partial f_m(x)$ .
- For  $\alpha \geq 0$ ,  $\partial(\alpha f)(x) = \alpha \partial f(x)$ .

## Subgradients of $f(x) = \|x\|_1$

- Let's find the subdifferential of  $f(x) = \|x\|_1 = \sum_{i=1}^d |x_i|$  at any given point  $x^0 = (x_1^0, x_2^0, \dots, x_d^0)$ .
- By an important property of subdifferential: If  $f = f_1 + \dots + f_m$ , then  $\partial f(x) = \partial f_1(x) + \dots + \partial f_m(x)$ .
- We could calculate the subgradient of  $f^i(x) = |x_i|$  and sum them up.
- The subgradient  $g^i = (g_1^i, \dots, g_d^i)$  of  $f^i(x) = |x_i|$  at  $x^0 = (x_1^0, x_2^0, \dots, x_d^0)$  is:

$$g_j^i = 0, \quad j \neq i; \quad g_j^i = s(x_j^0), \quad j = i,$$

where  $s(x) = \text{sign}(x)$  if  $x \neq 0$  and  $s(x) \in [-1, 1]$  if  $x = 0$

- We sum all the  $g^i$  up to get the subgradient  $g = (g_1, \dots, g_d)$  of  $f(x)$  at  $x^0$ :

$$g_i = s(x_i^0) \quad \text{for all } i$$

# 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?

# Convergence Theorem for Fixed Step Size

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

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

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

## Theorem

For fixed step size  $t$ , subgradient method satisfies:

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

# Convergence Theorems for Decreasing Step Sizes

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

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

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

## Theorem

For step size respecting Robbins-Monro conditions,

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

# Subgradient Descent for Lasso Problem

- Lasso problem can be parametrized as

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

- Subgradients of  $J(w)$  are

$$\frac{1}{n} \sum_{i=1}^n 2\{w^T x_i - y_i\}x_i + \lambda s,$$

where  $s_i = \text{sign}(w_i)$  if  $w_i \neq 0$  and  $s_i \in [-1, 1]$  if  $w_i = 0$ .

## Subgradient Descent for Lasso Problem: Potential Issues

- Subgradient descent will work for all convex and Lipschitz continuous objective functions.
- BUT, convergence can be very **slow** for non-differentiable functions
- One can often find better approaches by closer examination of the objective function. For example, shooting method or projected SGD.
- Taking small steps in the direction of the (sub)gradient usually may **not** lead to zero coordinates.
- BUT, in practice, we can threshold small values.

## Appendix: 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  $z = x^* \in \arg \min_x f(x)$ .

⇒ Appendix: 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

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

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

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