

# Subgradient Descent

David S. Rosenberg

Bloomberg ML EDU

October 18, 2017

## Motivation and Review: Support Vector Machines

# The Classification Problem

- Output space  $\mathcal{Y} = \{-1, 1\}$       Action space  $\mathcal{A} = \mathbb{R}$
- **Real-valued prediction function**  $f : \mathcal{X} \rightarrow \mathbb{R}$
- The value  $f(x)$  is called the **score** for the input  $x$ .
- Intuitively, magnitude of the score represents the **confidence of our prediction**.
- Typical convention:

$$f(x) > 0 \implies \text{Predict } 1$$

$$f(x) < 0 \implies \text{Predict } -1$$

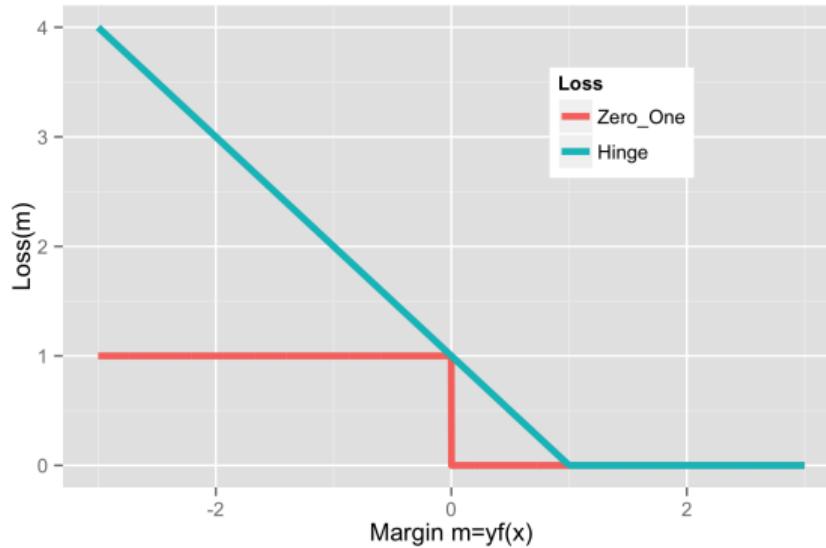
(But we can choose other thresholds...)

# The Margin

- The **margin** (or **functional margin**) for predicted score  $\hat{y}$  and true class  $y \in \{-1, 1\}$  is  $y\hat{y}$ .
- The margin often looks like  $yf(x)$ , where  $f(x)$  is our score function.
- The margin is a measure of how **correct** we are.
- We want to **maximize the margin**.

## [Margin-Based] Classification Losses

SVM/Hinge loss:  $\ell_{\text{Hinge}} = \max\{1 - m, 0\} = (1 - m)_+$



Not differentiable at  $m = 1$ . We have a “margin error” when  $m < 1$ .

## [Soft Margin] Linear Support Vector Machine (No Intercept)

- Hypothesis space  $\mathcal{F} = \{f(x) = w^T x \mid w \in \mathbf{R}^d\}$ .
- Loss  $\ell(m) = \max(1, m)$
- $\ell_2$  regularization

$$\min_{w \in \mathbf{R}^d} \sum_{i=1}^n \max(0, 1 - y_i w^T x_i) + \lambda \|w\|_2^2$$

## SVM Optimization Problem (no intercept)

- SVM objective function:

$$J(w) = \frac{1}{n} \sum_{i=1}^n \max(0, 1 - y_i [w^T x_i]) + \lambda \|w\|^2.$$

- Not differentiable... but let's think about gradient descent anyway.
- Derivative of hinge loss  $\ell(m) = \max(0, 1 - m)$ :

$$\ell'(m) = \begin{cases} 0 & m > 1 \\ -1 & m < 1 \\ \text{undefined} & m = 1 \end{cases}$$

## “Gradient” of SVM Objective

- We need gradient with respect to parameter vector  $w \in \mathbf{R}^d$ :

$$\begin{aligned}\nabla_w \ell(y_i w^T x_i) &= \ell'(y_i w^T x_i) y_i x_i \text{ (chain rule)} \\ &= \begin{cases} 0 & y_i w^T x_i > 1 \\ -1 & y_i w^T x_i < 1 \\ \text{undefined} & y_i w^T x_i = 1 \end{cases} y_i x_i \text{ (expanded } m \text{ in } \ell'(m)) \\ &= \begin{cases} 0 & y_i w^T x_i > 1 \\ -y_i x_i & y_i w^T x_i < 1 \\ \text{undefined} & y_i w^T x_i = 1 \end{cases}\end{aligned}$$

## “Gradient” of SVM Objective

$$\nabla_w \ell(y_i w^T x_i) = \begin{cases} 0 & y_i w^T x_i > 1 \\ -y_i x_i & y_i w^T x_i < 1 \\ \text{undefined} & y_i w^T x_i = 1 \end{cases}$$

So

$$\begin{aligned}\nabla_w J(w) &= \nabla_w \left( \frac{1}{n} \sum_{i=1}^n \ell(y_i w^T x_i) + \lambda \|w\|^2 \right) \\ &= \frac{1}{n} \sum_{i=1}^n \nabla_w \ell(y_i w^T x_i) + 2\lambda w \\ &= \begin{cases} \frac{1}{n} \sum_{i:y_i w^T x_i < 1} (-y_i x_i) + 2\lambda w & \text{all } y_i w^T x_i \neq 1 \\ \text{undefined} & \text{otherwise} \end{cases}\end{aligned}$$

## Gradient Descent on SVM Objective?

- The gradient of the SVM objective is

$$\nabla_w J(w) = \frac{1}{n} \sum_{i:y_i w^T x_i < 1} (-y_i x_i) + 2\lambda w$$

when  $y_i w^T x_i \neq 1$  for all  $i$ , and **otherwise is undefined**.

Suppose we tried gradient descent on  $J(w)$ :

- If we start with a random  $w$ , will we ever hit  $y_i w^T x_i = 1$ ?
- If we did, could we perturb the step size by  $\epsilon$  to miss such a point?
- Does it even make sense to check  $y_i w^T x_i = 1$  with floating point numbers?

## Gradient Descent on SVM Objective?

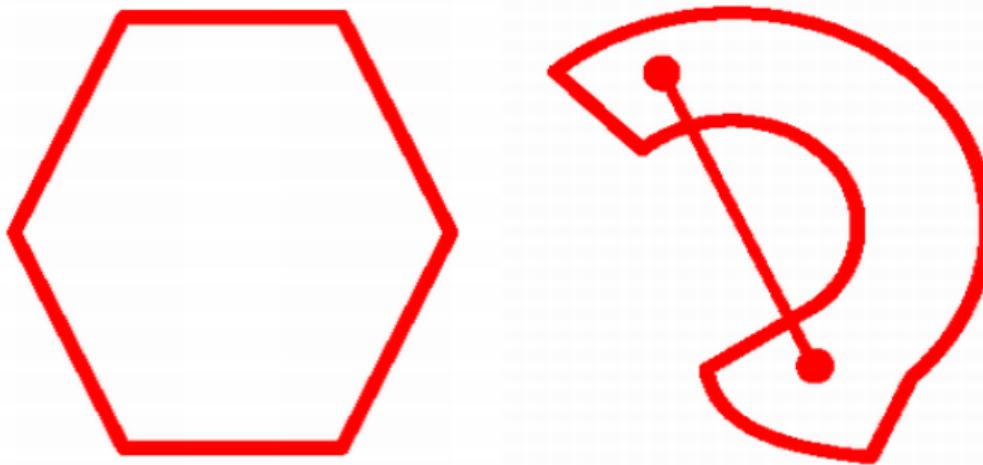
- If we blindly apply gradient descent from a random starting point
  - seems unlikely that we'll hit a point where the gradient is undefined.
- Still, doesn't mean that gradient descent will work if objective not differentiable!
- Theory of subgradients and subgradient descent will clear up any uncertainty.

## Convexity and Sublevel Sets

# Convex Sets

## Definition

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

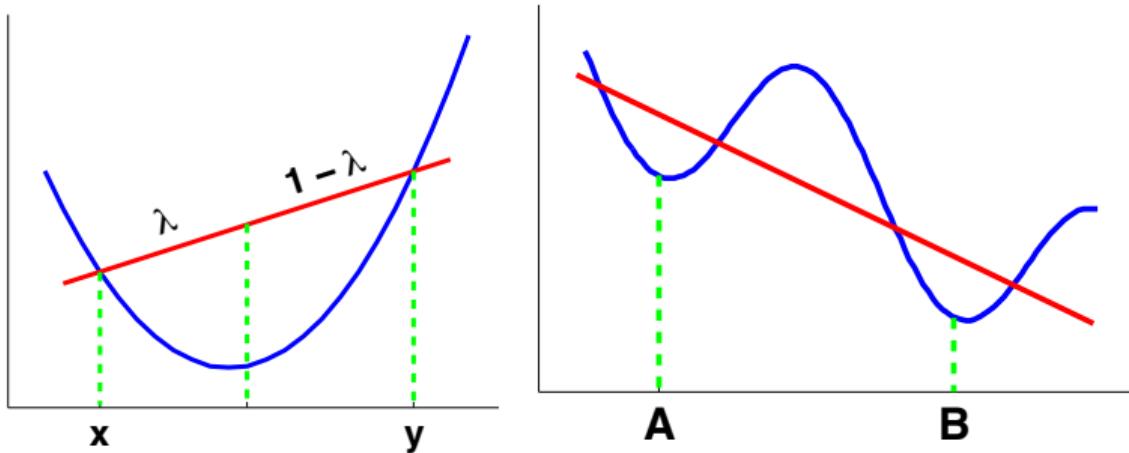


KPM Fig. 7.4

# Convex and Concave Functions

## Definition

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

## 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.

Question: Is the  $\leq$  in the constraint just a convention? Could we also have used  $\geq$  or  $=$ ?

# 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

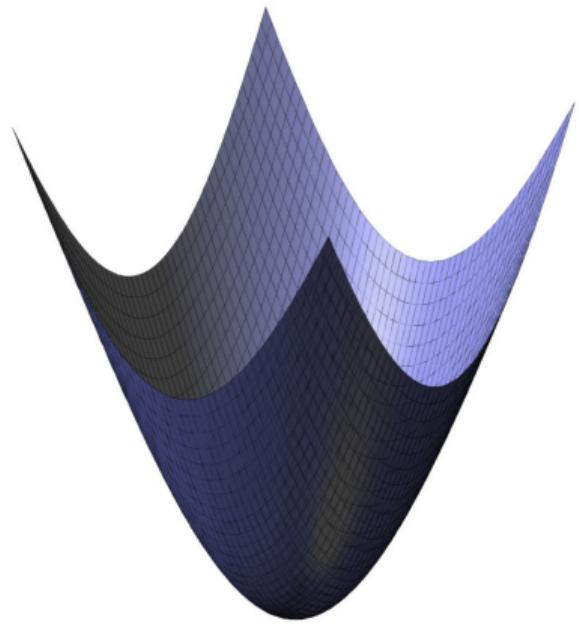
A **sublevel** set for the value  $c$  is the set of points  $x \in \mathbf{R}^d$  for which  $f(x) \leq c$ .

## Theorem

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

(Proof straight from definitions.)

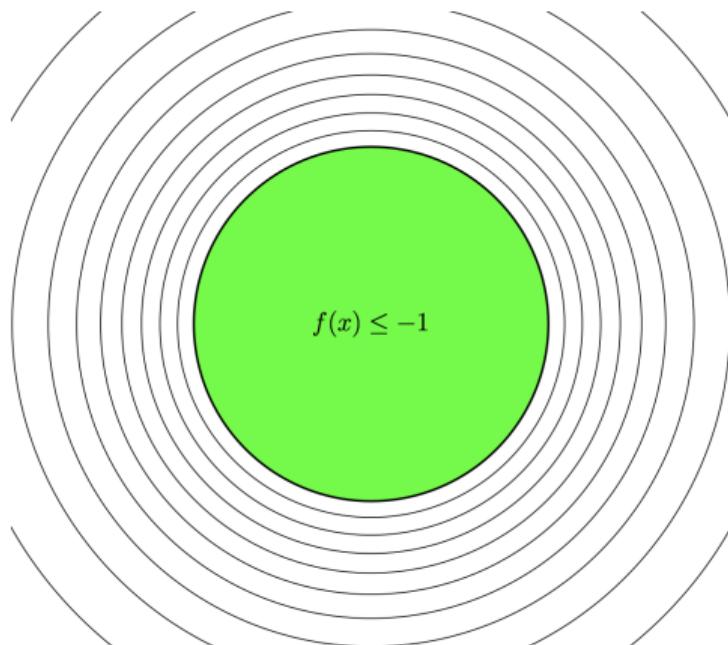
# Convex Function



---

Plot courtesy of Brett Bernstein.

## Contour Plot Convex Function: Sublevel Set

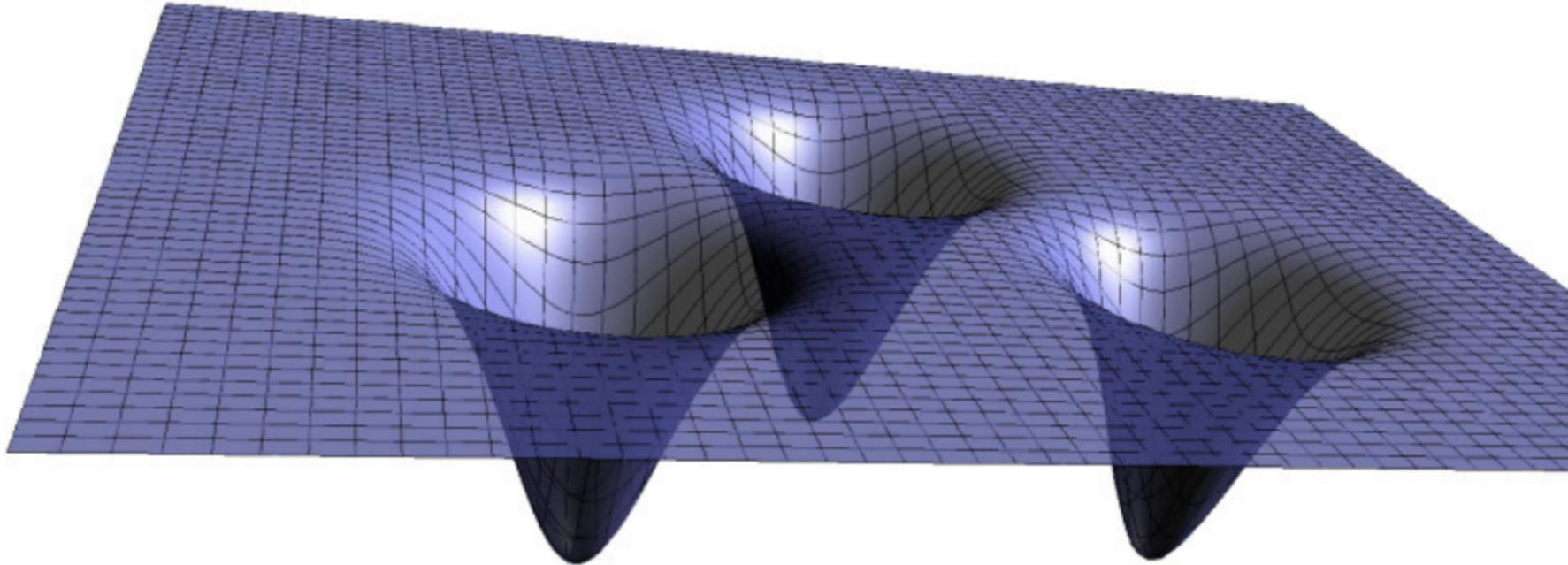


Is the sublevel set  $\{x \mid f(x) \leq 1\}$  convex?

---

Plot courtesy of Brett Bernstein.

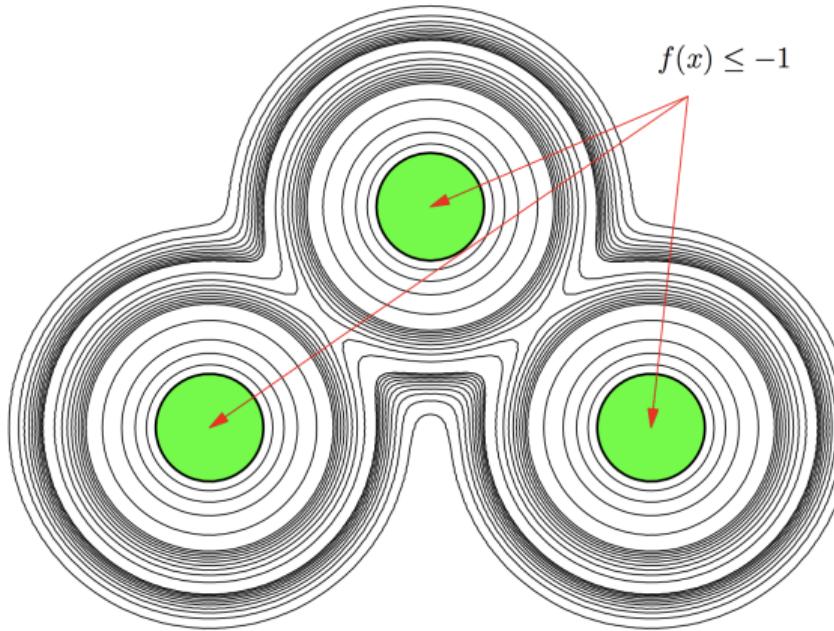
## Nonconvex Function



---

Plot courtesy of Brett Bernstein.

## Contour Plot Nonconvex Function: Sublevel Set

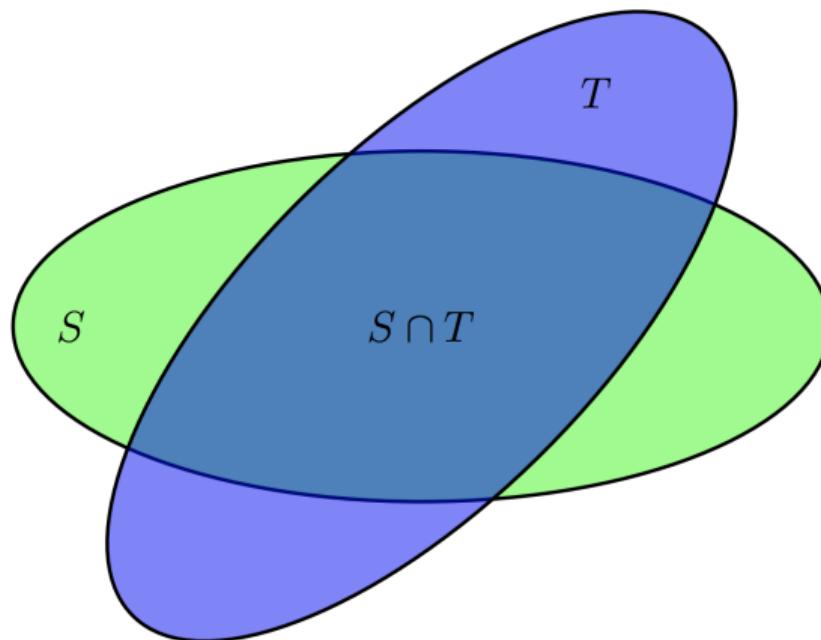


Is the sublevel set  $\{x \mid f(x) \leq 1\}$  convex?

---

Plot courtesy of Brett Bernstein.

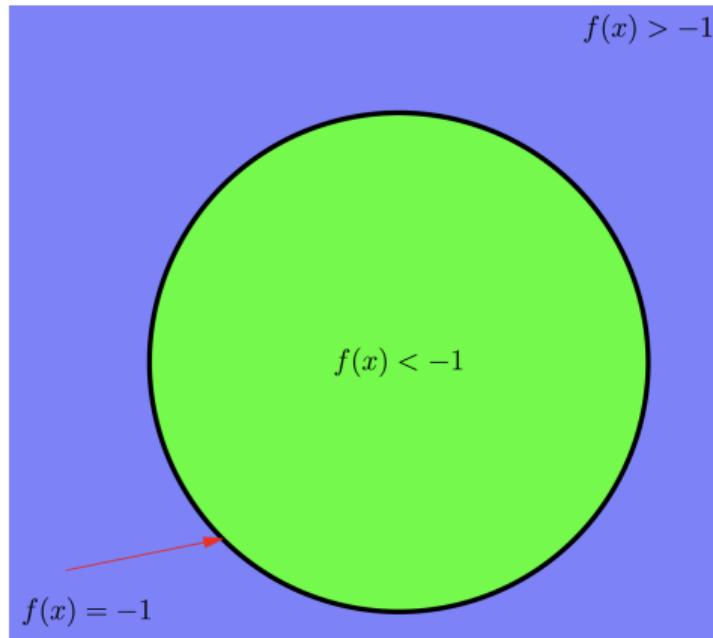
## Fact: Intersection of Convex Sets is Convex



---

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.

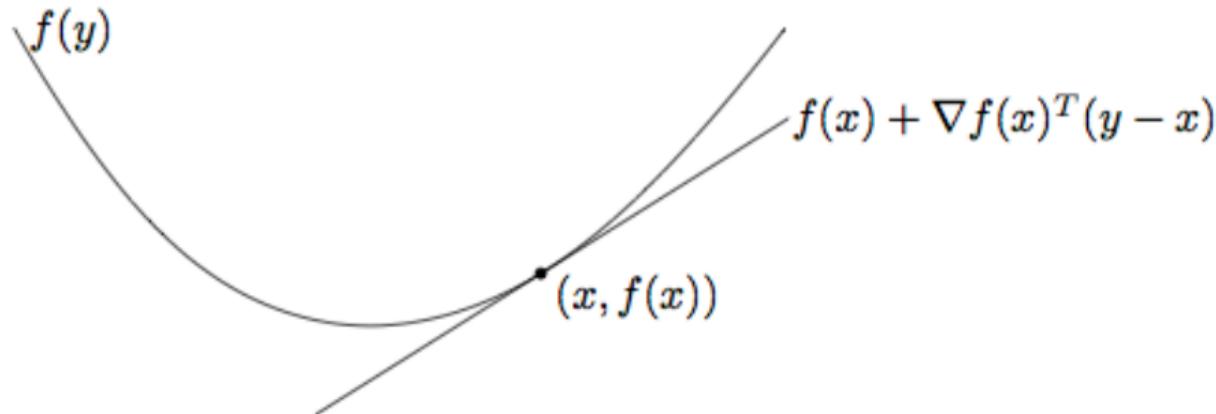
An alternative “generic” convex optimization problem.

## Convex and Differentiable Functions

## First-Order Approximation

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

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

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

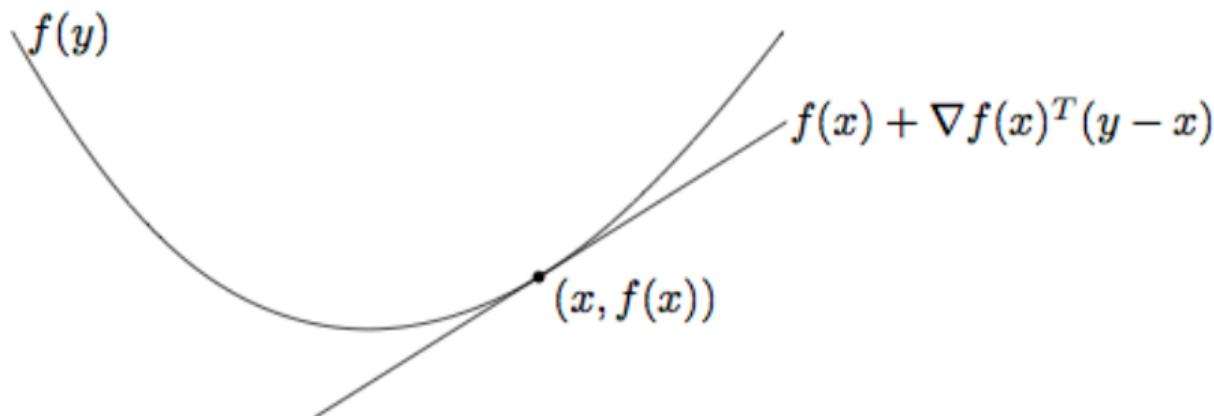


Figure from Boyd & Vandenberghe Fig. 3.2; Proof in Section 3.1.3

# First-Order Condition for Convex, Differentiable Function

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

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

## Corollary

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

For convex functions, local information gives global information.

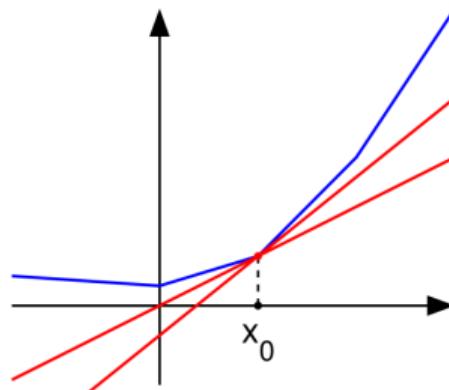
# Subgradients

# Subgradients

## Definition

A vector  $g \in \mathbf{R}^n$  is a **subgradient** of  $f : \mathbf{R}^n \rightarrow \mathbf{R}$  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.

# Global Optimality Condition

## Definition

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

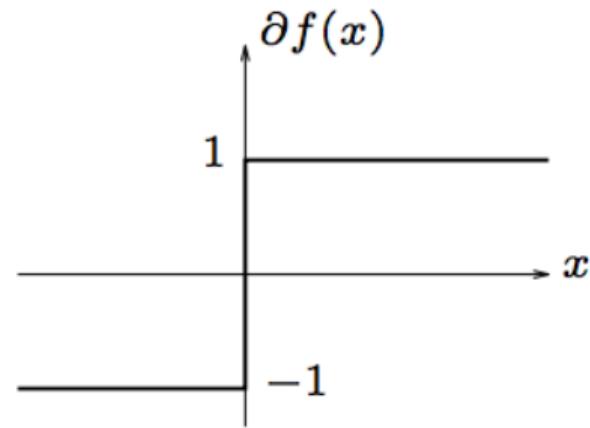
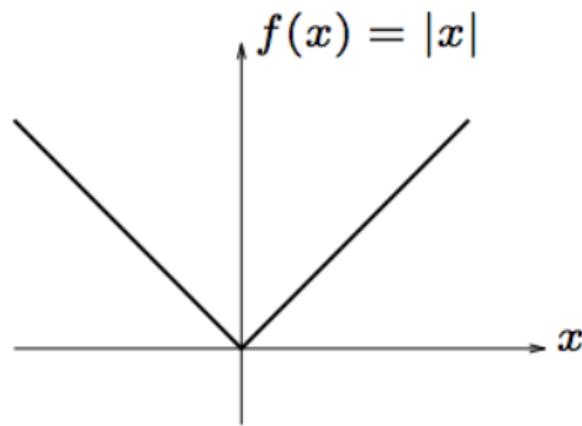
$$f(z) \geq f(x) + g^T(z - x).$$

## Corollary

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

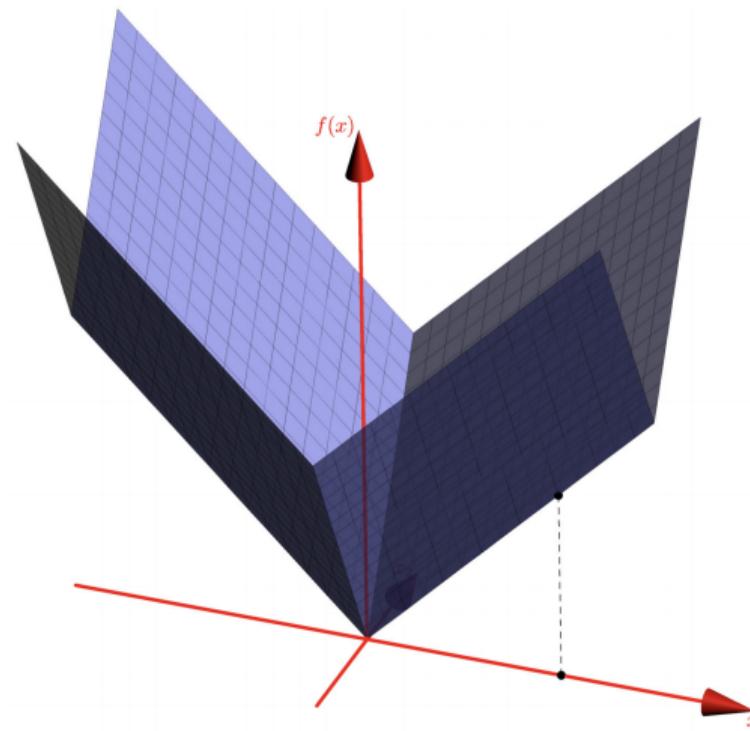
## Subdifferential of Absolute Value

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



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

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



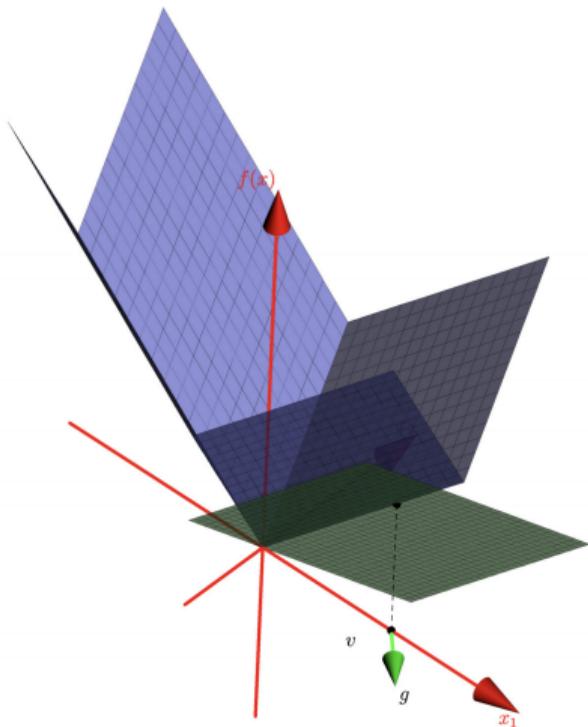
---

Plot courtesy of Brett Bernstein.

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

- Let's find the subdifferential of  $f(x_1, x_2) = |x_1| + 2|x_2|$  and  $(3, 0)$ .
- 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  $g_2 \in [-2, 2]$ .

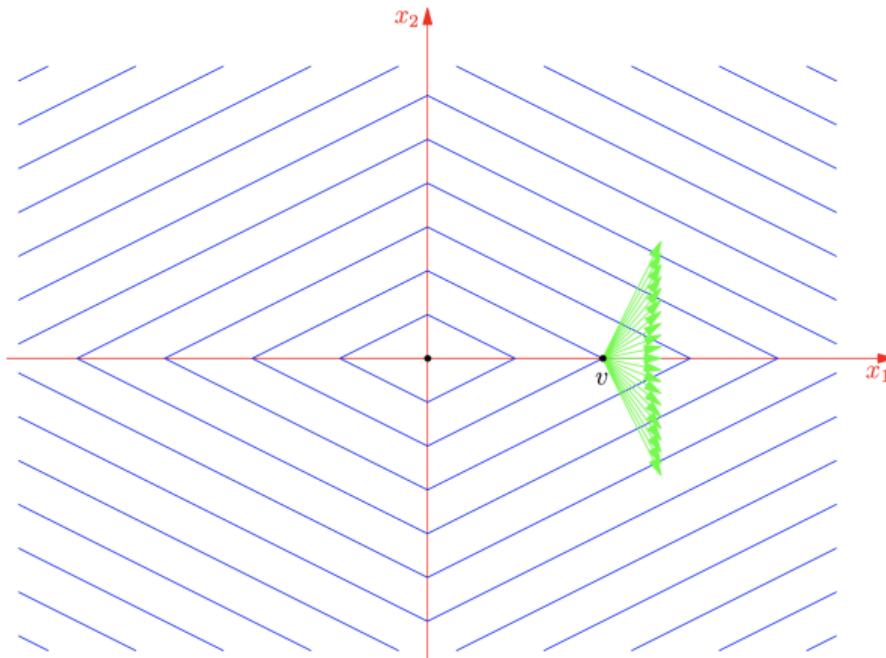
# Underestimating Hyperplane to $f(x_1, x_2) = |x_1| + 2|x_2|$



Plot courtesy of Brett Bernstein.

## Subdifferential on Contour Plot

$$\partial f(3, 0) = \{(1, b)^T \mid b \in [-2, 2]\}$$

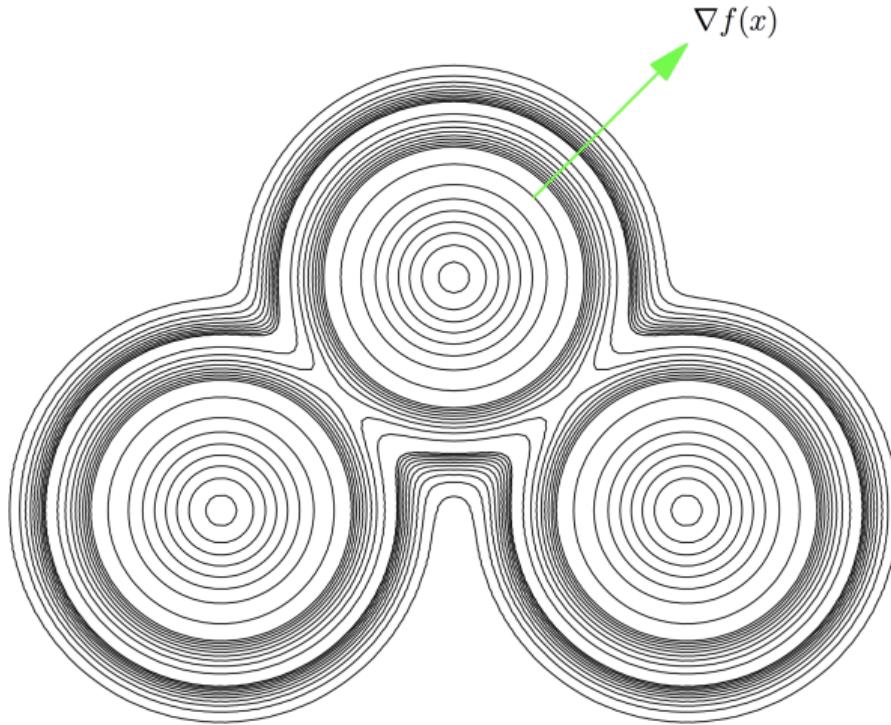


Contour plot of  $f(x_1, x_2) = |x_1| + 2|x_2|$ , with set of subgradients at  $(3, 0)$ . .

## Contour Lines and Gradients

- For function  $f : \mathbf{R}^d \rightarrow \mathbf{R}$ ,
  - **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  $\mathbf{R}^d$ .
- $f : \mathbf{R}^d \rightarrow \mathbf{R}$  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

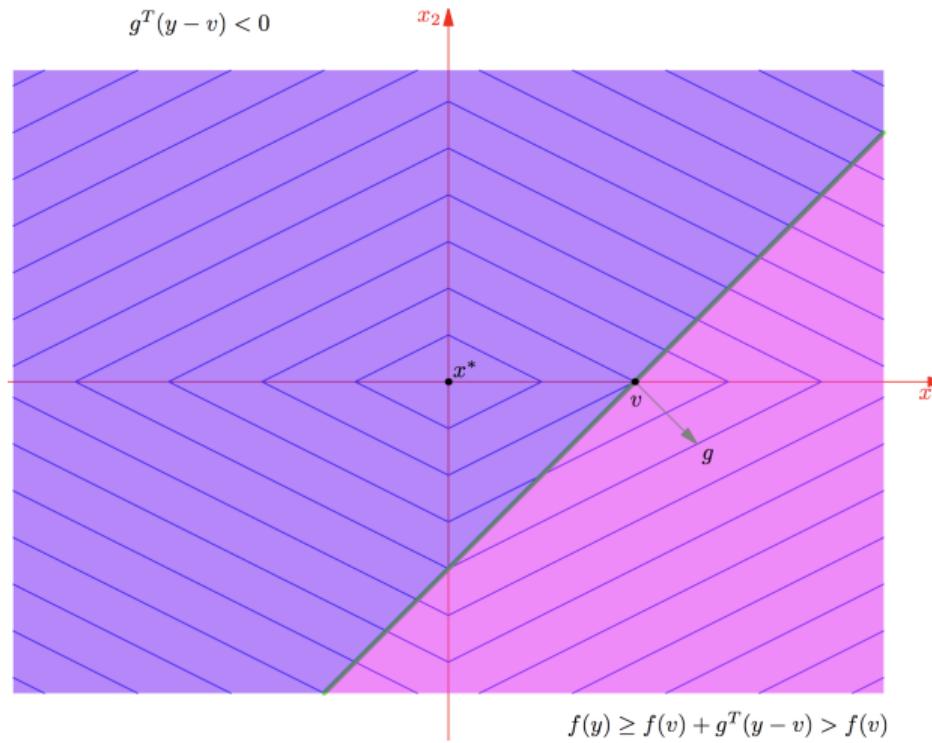
Let  $f : \mathbf{R}^d \rightarrow \mathbf{R}$  have a subgradient  $g$  at  $x_0$ .

- Hyperplane  $H$  orthogonal to  $g$  at  $x_0$  must **support** the level set  $S = \{x \in \mathbf{R}^d \mid f(x) = f(x_0)\}$ .
  - i.e  $H$  contains  $x_0$  and all of  $S$  lies one one side of  $H$ .

Proof:

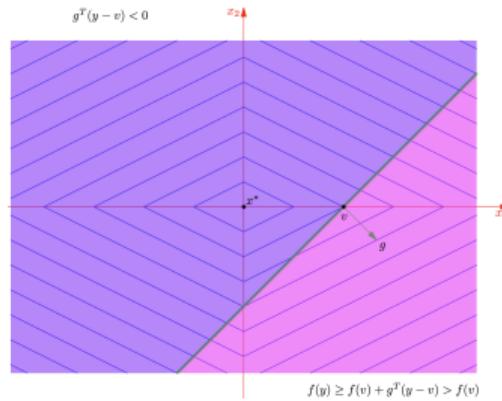
- For any  $y$ , we have  $f(y) \geq f(x) + g^T(y - x)$ . (def of subgradient)
- If  $y$  is strictly on side of  $H$  that  $g$  points in,
  - then  $g^T(y - x) > 0$ .
  - So  $f(y) > f(x)$ .
  - So  $y$  is not in the level set  $S$ .
- $\therefore$  All elements of  $S$  must be on  $H$  or on the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

⇒ 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

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

# 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,

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