

# Optimization for Machine Learning

## Lecture 1: Introduction to Convexity

S.V. N. (vishy) Vishwanathan

Purdue University  
vishy@purdue.edu

July 12, 2012

# Regularized Risk Minimization

## Machine Learning

- We want to build a model which predicts well on data
- A model's performance is quantified by a loss function
  - a sophisticated discrepancy score
- Our model must generalize to unseen data
- Avoid over-fitting by penalizing *complex* models (Regularization)

## More Formally

- Training data:  $\{x_1, \dots, x_m\}$
- Labels:  $\{y_1, \dots, y_m\}$
- Learn a vector:  $w$

$$\underset{w}{\text{minimize}} \quad J(w) := \underbrace{\lambda \Omega(w)}_{\text{Regularizer}} + \underbrace{\frac{1}{m} \sum_{i=1}^m l(x_i, y_i, w)}_{\text{Risk } R_{\text{emp}}}$$

# Regularized Risk Minimization

## Machine Learning

- We want to build a model which predicts well on data
- A model's performance is quantified by a loss function
  - a sophisticated discrepancy score
- Our model must generalize to unseen data
- Avoid over-fitting by penalizing *complex* models (Regularization)

## More Formally

- Training data:  $\{x_1, \dots, x_m\}$
- Labels:  $\{y_1, \dots, y_m\}$
- Learn a vector:  $w$

$$\underset{w}{\text{minimize}} \quad J(w) := \underbrace{\lambda \Omega(w)}_{\text{Regularizer}} + \underbrace{\frac{1}{m} \sum_{i=1}^m l(x_i, y_i, w)}_{\text{Risk } R_{\text{emp}}}$$

# Regularized Risk Minimization

## Machine Learning

- We want to build a model which predicts well on data
- A model's performance is quantified by a loss function
  - a sophisticated discrepancy score
- Our model must generalize to unseen data
- Avoid over-fitting by penalizing *complex* models (Regularization)

## More Formally

- Training data:  $\{x_1, \dots, x_m\}$
- Labels:  $\{y_1, \dots, y_m\}$
- Learn a vector:  $w$

$$\underset{w}{\text{minimize}} \quad J(w) := \underbrace{\lambda \Omega(w)}_{\text{Regularizer}} + \underbrace{\frac{1}{m} \sum_{i=1}^m l(x_i, y_i, w)}_{\text{Risk } R_{\text{emp}}}$$

# Regularized Risk Minimization

## Machine Learning

- We want to build a model which predicts well on data
- A model's performance is quantified by a loss function
  - a sophisticated discrepancy score
- Our model must generalize to unseen data
- Avoid over-fitting by penalizing *complex* models (Regularization)

## More Formally

- Training data:  $\{x_1, \dots, x_m\}$
- Labels:  $\{y_1, \dots, y_m\}$
- Learn a vector:  $w$

$$\underset{w}{\text{minimize}} \quad J(w) := \underbrace{\lambda \Omega(w)}_{\text{Regularizer}} + \underbrace{\frac{1}{m} \sum_{i=1}^m l(x_i, y_i, w)}_{\text{Risk } R_{\text{emp}}}$$

# Regularized Risk Minimization

## Machine Learning

- We want to build a model which predicts well on data
- A model's performance is quantified by a loss function
  - a sophisticated discrepancy score
- Our model must generalize to unseen data
- Avoid over-fitting by penalizing *complex* models (Regularization)

## More Formally

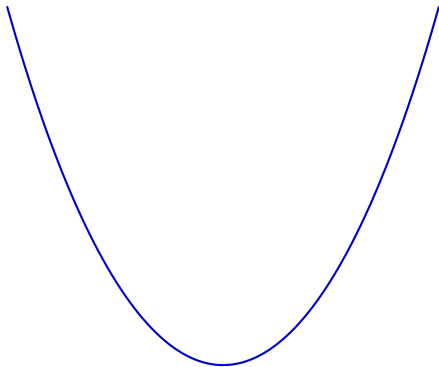
- Training data:  $\{x_1, \dots, x_m\}$
- Labels:  $\{y_1, \dots, y_m\}$
- Learn a vector:  $w$

$$\underset{w}{\text{minimize}} \quad J(w) := \underbrace{\lambda \Omega(w)}_{\text{Regularizer}} + \underbrace{\frac{1}{m} \sum_{i=1}^m l(x_i, y_i, w)}_{\text{Risk } R_{\text{emp}}}$$

# Outline

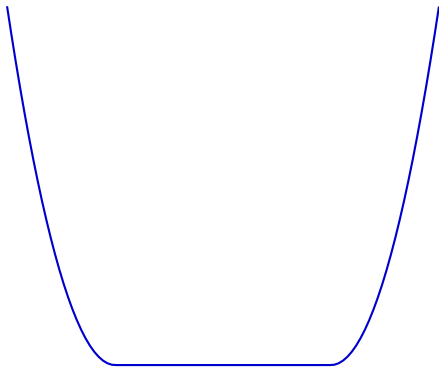
- 1 **Convex Functions and Sets**
- 2 Operations Which Preserve Convexity
- 3 First Order Properties
- 4 Subgradients
- 5 Constraints
- 6 Warmup: Minimizing a 1-d Convex Function
- 7 Warmup: Coordinate Descent

# Focus of my Lectures

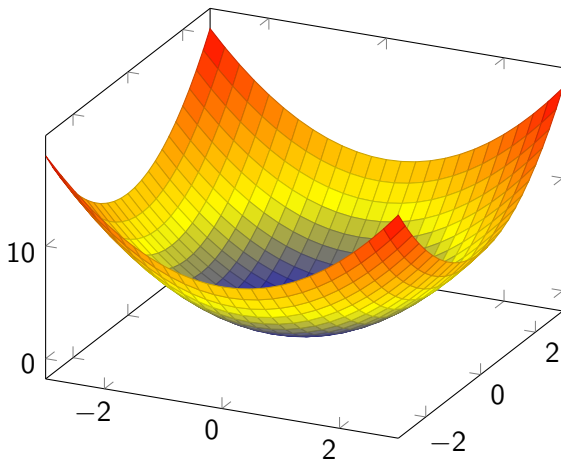




# Focus of my Lectures



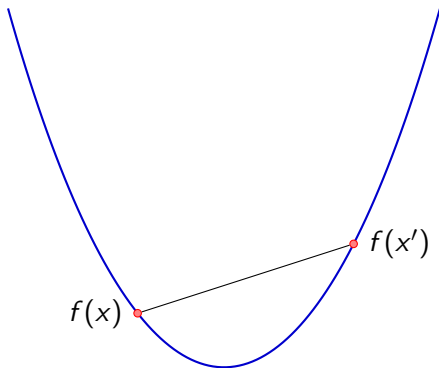
## Focus of my Lectures



# Disclaimer

- My focus is on showing connections between various methods
- I will sacrifice mathematical rigor and focus on intuition

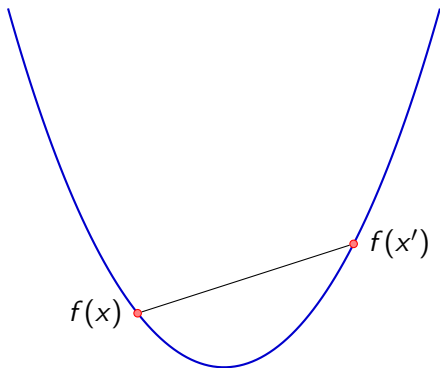
# Convex Function



A function  $f$  is convex if, and only if, for all  $x, x'$  and  $\lambda \in (0, 1)$

$$f(\lambda x + (1 - \lambda)x') \leq \lambda f(x) + (1 - \lambda)f(x')$$

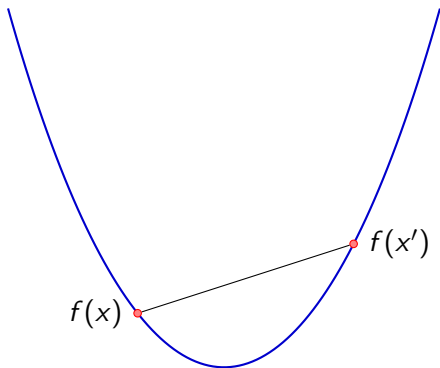
# Convex Function



A function  $f$  is **strictly** convex if, and only if, for all  $x, x'$  and  $\lambda \in (0, 1)$

$$f(\lambda x + (1 - \lambda)x') < \lambda f(x) + (1 - \lambda)f(x')$$

# Convex Function



A function  $f$  is  $\sigma$ -strongly convex if, and only if,  $f(\cdot) - \frac{\sigma}{2} \|\cdot\|^2$  is convex. That is, for all  $x, x'$  and  $\lambda \in (0, 1)$

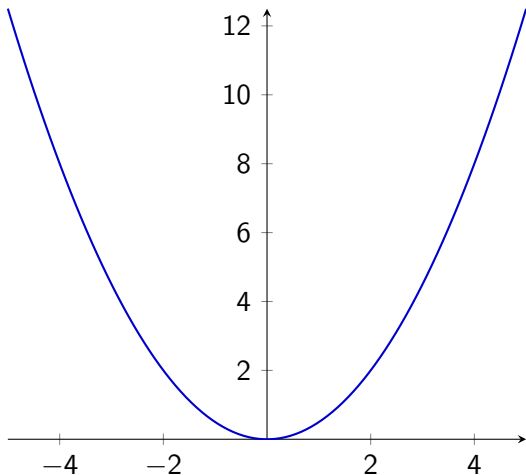
$$f(\lambda x + (1 - \lambda)x') \leq \lambda f(x) + (1 - \lambda)f(x') - \frac{\sigma}{2} \lambda(1 - \lambda) \|x - x'\|^2$$

## Exercise: Jensen's Inequality

- Extend the definition of convexity to show that if  $f$  is convex, then for all  $\lambda_i \geq 0$  such that  $\sum_i \lambda_i = 1$  we have

$$f\left(\sum_i \lambda_i x_i\right) \leq \sum_i \lambda_i f(x_i)$$

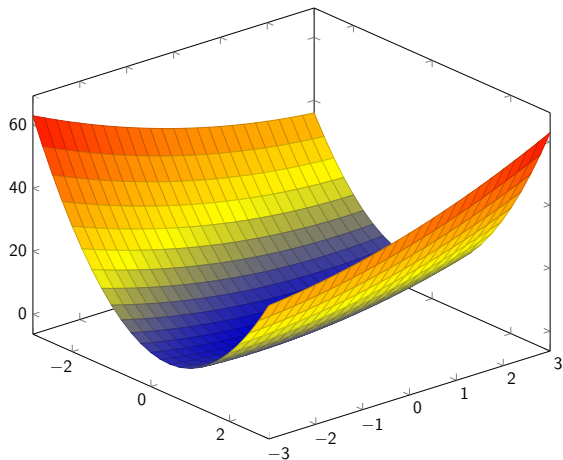
## Some Familiar Examples



$$f(x) = \frac{1}{2}x^2 \text{ (Square norm)}$$

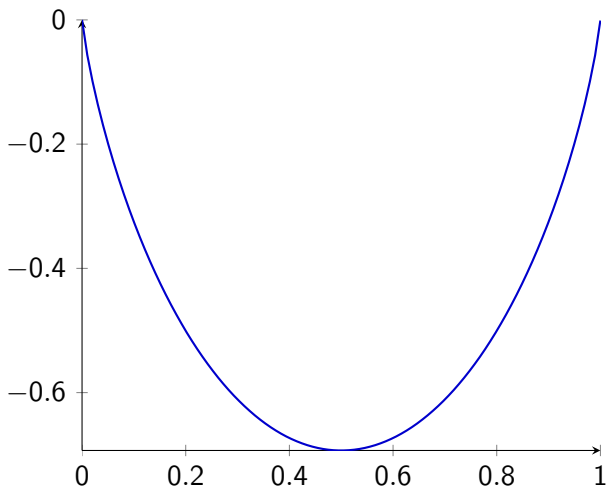


## Some Familiar Examples



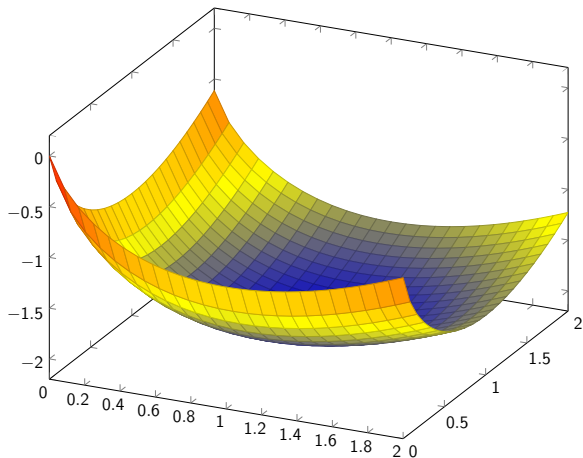
$$f(x, y) = \frac{1}{2} \begin{bmatrix} x, y \end{bmatrix} \begin{bmatrix} 10, 1 \\ 2, 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

## Some Familiar Examples



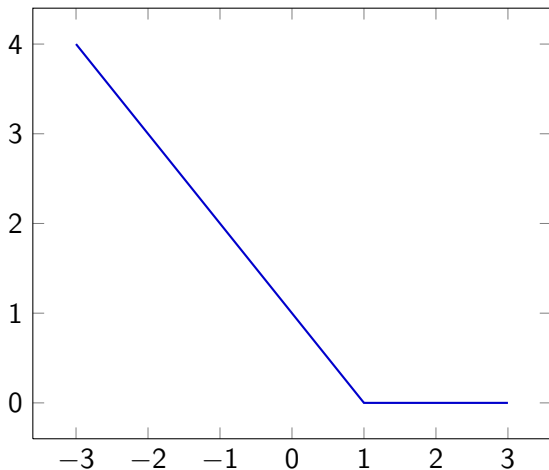
$$f(x) = x \log x + (1 - x) \log(1 - x) \text{ (Negative entropy)}$$

## Some Familiar Examples



$$f(x, y) = x \log x + y \log y - x - y \text{ (Un-normalized negative entropy)}$$

## Some Familiar Examples

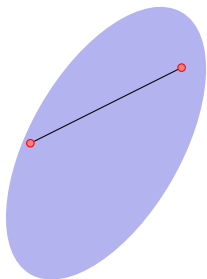


$$f(x) = \max(0, 1 - x) \text{ (Hinge Loss)}$$

## Some Other Important Examples

- Linear functions:  $f(x) = ax + b$
- Softmax:  $f(x) = \log \sum_i \exp(x_i)$
- Norms: For example the 2-norm  $f(x) = \sqrt{\sum_i x_i^2}$

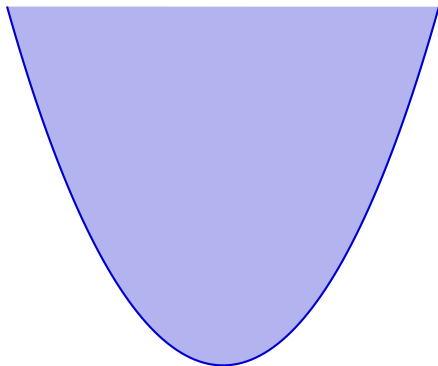
# Convex Sets



A set  $C$  is convex if, and only if, for all  $x, x' \in C$  and  $\lambda \in (0, 1)$  we have

$$\lambda x + (1 - \lambda)x' \in C$$

# Convex Sets and Convex Functions



A function  $f$  is convex if, and only if, its epigraph is a convex set

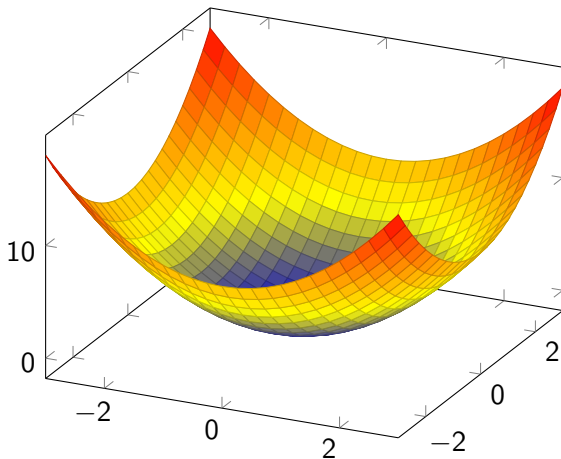
# Convex Sets and Convex Functions

- Indicator functions of convex sets are convex

$$I_C(x) = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{otherwise.} \end{cases}$$

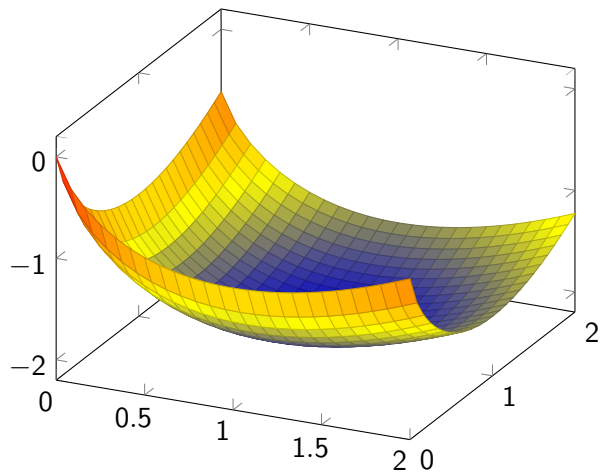


## Below sets of Convex Functions



$$f(x, y) = x^2 + y^2$$

## Below sets of Convex Functions

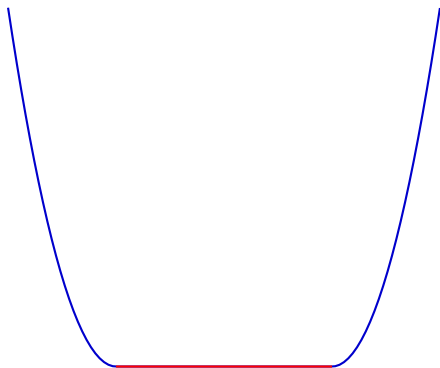


$$f(x, y) = x \log x + y \log y - x - y$$

## Below sets of Convex Functions

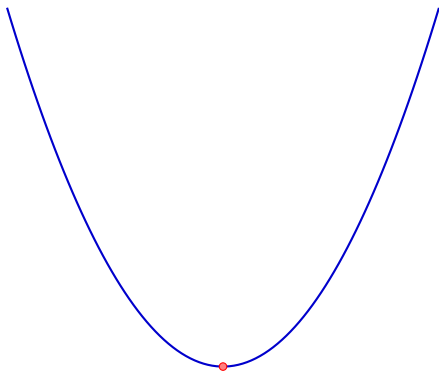
- If  $f$  is convex, then all its level sets are convex
- Is the converse true? (Exercise: construct a counter-example)

# Minima on Convex Sets



- Set of minima of a convex function is a convex set
- Proof: Consider the set  $\{x : f(x) \leq f^*\}$

# Minima on Convex Sets



- Set of minima of a **strictly** convex function is a singleton
- Proof: try this at home!

# Outline

- 1 Convex Functions and Sets
- 2 Operations Which Preserve Convexity**
- 3 First Order Properties
- 4 Subgradients
- 5 Constraints
- 6 Warmup: Minimizing a 1-d Convex Function
- 7 Warmup: Coordinate Descent

# Set Operations

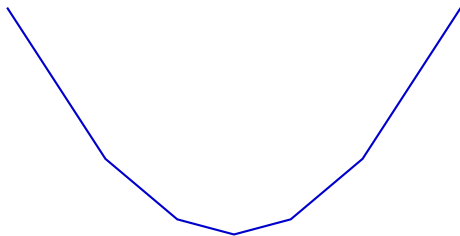
- Intersection of convex sets is convex
- Image of a convex set under a linear transformation is convex
- Inverse image of a convex set under a linear transformation is convex

## Function Operations

- Linear Combination with non-negative weights:  $f(x) = \sum_i w_i f_i(x)$   
s.t.  $w_i \geq 0$
- Pointwise maximum:  $f(x) = \max_i f_i(x)$
- Composition with affine function:  $f(x) = g(Ax + b)$
- Projection along a direction:  $f(\eta) = g(x_0 + \eta d)$
- Restricting the domain on a convex set:  $f(x)$  s.t.  $x \in \mathcal{C}$



## One Quick Example



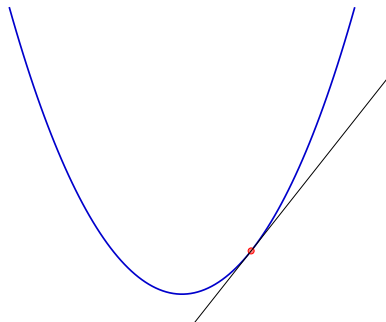
The piecewise linear function  $f(x) := \max_i \langle u_i, x \rangle$  is convex

# Outline

- 1 Convex Functions and Sets
- 2 Operations Which Preserve Convexity
- 3 First Order Properties**
- 4 Subgradients
- 5 Constraints
- 6 Warmup: Minimizing a 1-d Convex Function
- 7 Warmup: Coordinate Descent

# First Order Taylor Expansion

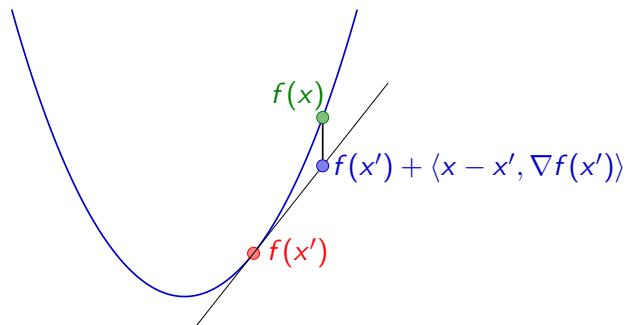
The First Order Taylor approximation globally lower bounds the function



For any  $x$  and  $x'$  we have

$$f(x) \geq f(x') + \langle x - x', \nabla f(x') \rangle$$

# Bregman Divergence



- For any  $x$  and  $x'$  the Bregman divergence defined by  $f$  is given by

$$\Delta_f(x, x') = f(x) - f(x') - \langle x - x', \nabla f(x') \rangle.$$

## Euclidean Distance Squared

### Bregman Divergence

- For any  $x$  and  $x'$  the Bregman divergence defined by  $f$  is given by

$$\Delta_f(x, x') = f(x) - f(x') - \langle x - x', \nabla f(x') \rangle.$$

- Use  $f(x) = \frac{1}{2} \|x\|^2$  and verify that

$$\Delta_f(x, x') = \frac{1}{2} \|x - x'\|^2$$

## Unnormalized Relative Entropy

### Bregman Divergence

- For any  $x$  and  $x'$  the Bregman divergence defined by  $f$  is given by

$$\Delta_f(x, x') = f(x) - f(x') - \langle x - x', \nabla f(x') \rangle.$$

- Use  $f(x) = \sum_i x_i \log x_i - x_i$  and verify that

$$\Delta_f(x, x') = \sum_i x_i \log x_i - x_i - x_i \log x'_i + x'_i$$

## Identifying the Minimum

- Let  $f : X \rightarrow \mathbb{R}$  be a differentiable convex function. Then  $x$  is a minimizer of  $f$ , if, and only if,

$$\langle x' - x, \nabla f(x) \rangle \geq 0 \text{ for all } x'.$$

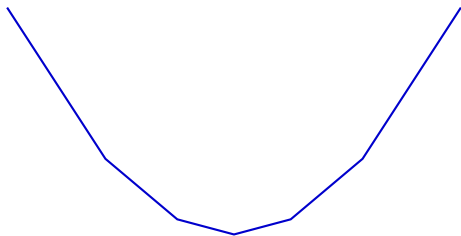
- One way to ensure this is to set  $\nabla f(x) = 0$
- Minimizing a smooth convex function is the same as finding an  $x$  such that  $\nabla f(x) = 0$

# Outline

- 1 Convex Functions and Sets
- 2 Operations Which Preserve Convexity
- 3 First Order Properties
- 4 Subgradients**
- 5 Constraints
- 6 Warmup: Minimizing a 1-d Convex Function
- 7 Warmup: Coordinate Descent



## What if the Function is NonSmooth?

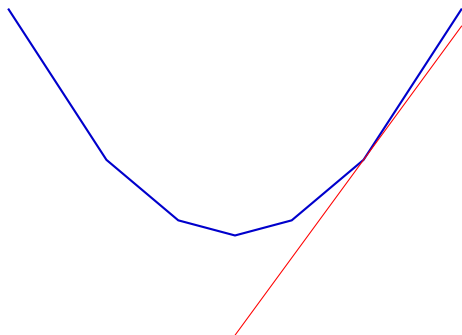


The piecewise linear function

$$f(x) := \max_i \langle u_i, x \rangle$$

is convex but not differentiable at the kinks!

# Subgradients to the Rescue

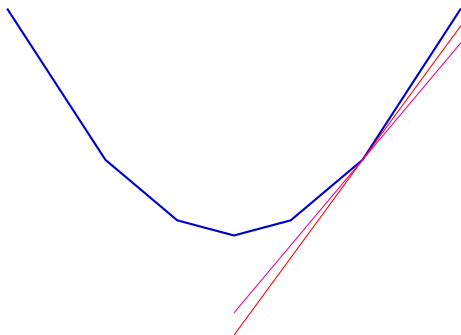


A subgradient at  $x'$  is any vector  $s$  which satisfies

$$f(x) \geq f(x') + \langle x - x', s \rangle \text{ for all } x$$

Set of all subgradients is denoted as  $\partial f(w)$

# Subgradients to the Rescue

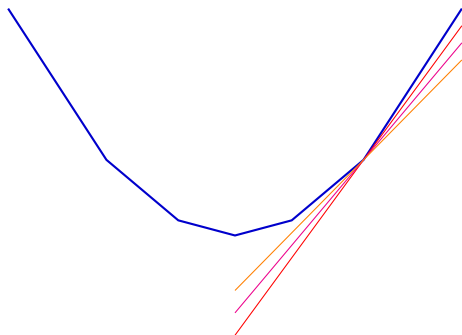


A subgradient at  $x'$  is any vector  $s$  which satisfies

$$f(x) \geq f(x') + \langle x - x', s \rangle \text{ for all } x$$

Set of all subgradients is denoted as  $\partial f(w)$

# Subgradients to the Rescue

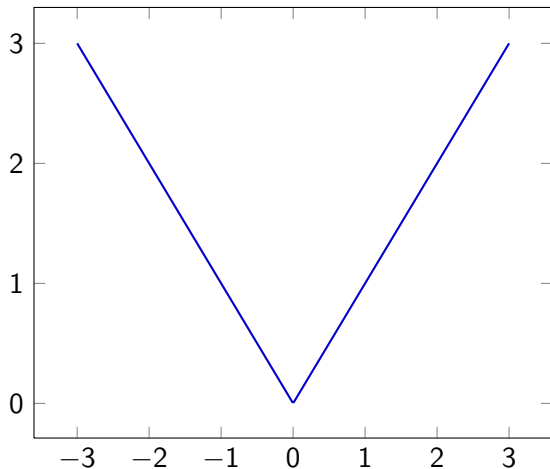


A subgradient at  $x'$  is any vector  $s$  which satisfies

$$f(x) \geq f(x') + \langle x - x', s \rangle \text{ for all } x$$

Set of all subgradients is denoted as  $\partial f(w)$

## Example



- $f(x) = |x|$  and  $\partial f(0) = [-1, 1]$

## Identifying the Minimum

- Let  $f : X \rightarrow \mathbb{R}$  be a convex function. Then  $x$  is a minimizer of  $f$ , if, and only if, there exists a  $\mu \in \partial f(x)$  such that

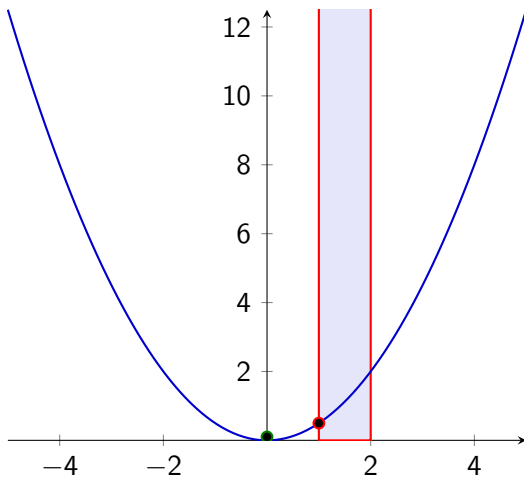
$$\langle x' - x, \mu \rangle \geq 0 \text{ for all } x'.$$

- One way to ensure this is to ensure that  $0 \in \partial f(x)$

# Outline

- 1 Convex Functions and Sets
- 2 Operations Which Preserve Convexity
- 3 First Order Properties
- 4 Subgradients
- 5 Constraints**
- 6 Warmup: Minimizing a 1-d Convex Function
- 7 Warmup: Coordinate Descent

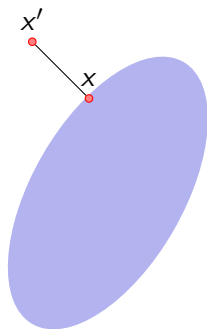
## A Simple Example



- Minimize  $\frac{1}{2}x^2$  s.t.  $1 \leq w \leq 2$



# Projection



$$P_{\mathcal{C}}(x') := \min_{x \in \mathcal{C}} \|x - x'\|^2$$

- Assignment: Compute  $P_{\mathcal{C}}(x')$  when  $\mathcal{C} = \{x \text{ s.t. } l \leq x_i \leq u\}$

# First Order Conditions For Constrained Problems

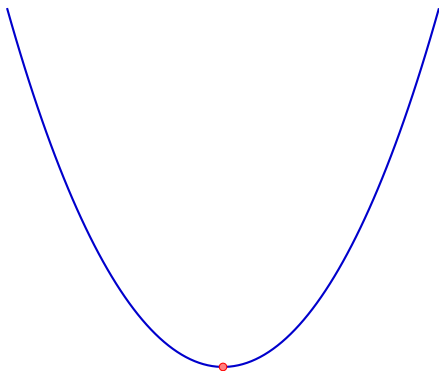
$$x = P_{\mathcal{C}}(x - \nabla f(x))$$

- If  $x - \nabla f(x) \in \mathcal{C}$  then  $P_{\mathcal{C}}(x - \nabla f(x)) = x$  implies that  $\nabla f(x) = 0$
- Otherwise, it shows that the constraints are preventing further progress in the direction of descent

# Outline

- 1 Convex Functions and Sets
- 2 Operations Which Preserve Convexity
- 3 First Order Properties
- 4 Subgradients
- 5 Constraints
- 6 Warmup: Minimizing a 1-d Convex Function**
- 7 Warmup: Coordinate Descent

## Problem Statement



- Given a black-box which can compute  $J: \mathbb{R} \rightarrow \mathbb{R}$  and  $J': \mathbb{R} \rightarrow \mathbb{R}$  find the minimum value of  $J$

## Increasing Gradients

- From the first order conditions

$$J(w) \geq J(w') + (w - w') \cdot J'(w')$$

and

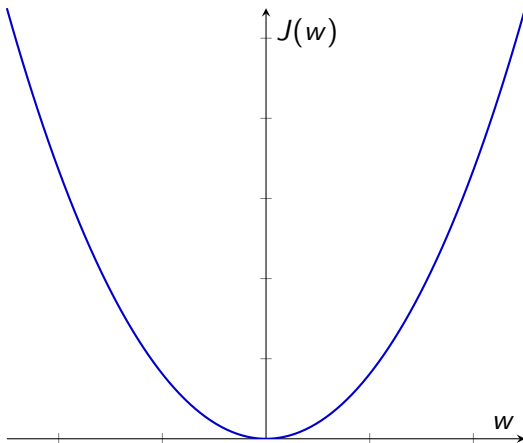
$$J(w') \geq J(w) + (w' - w) \cdot J'(w)$$

- Add the two

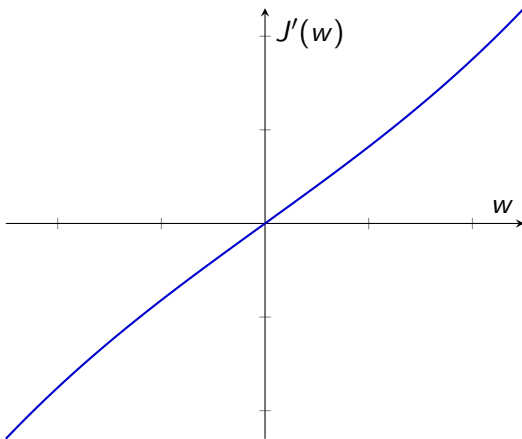
$$(w - w') \cdot (J'(w) - J'(w')) \geq 0$$

$w \geq w'$  implies that  $J'(w) \geq J'(w')$

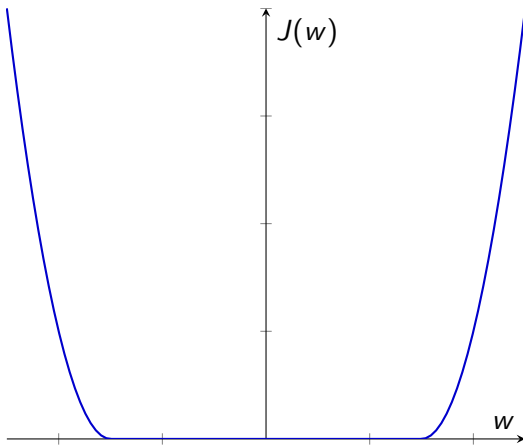
# Increasing Gradients



# Increasing Gradients

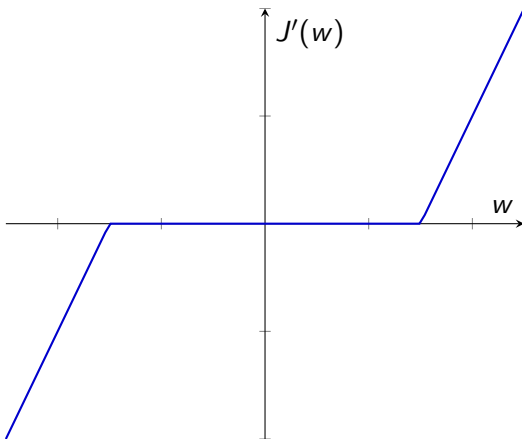


# Increasing Gradients

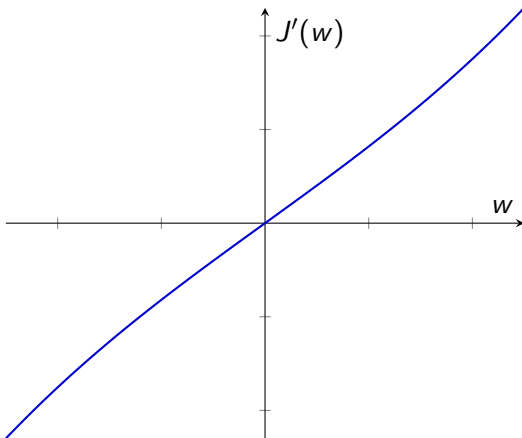




# Increasing Gradients

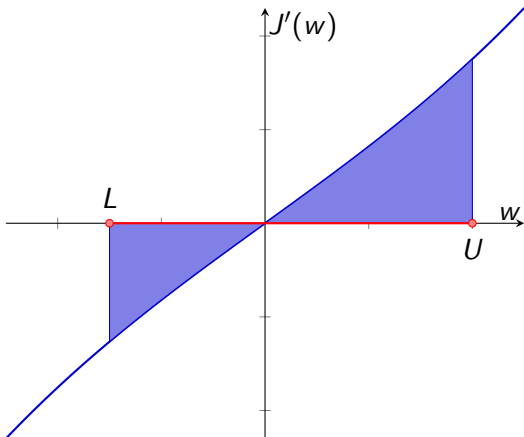


## Problem Restatement

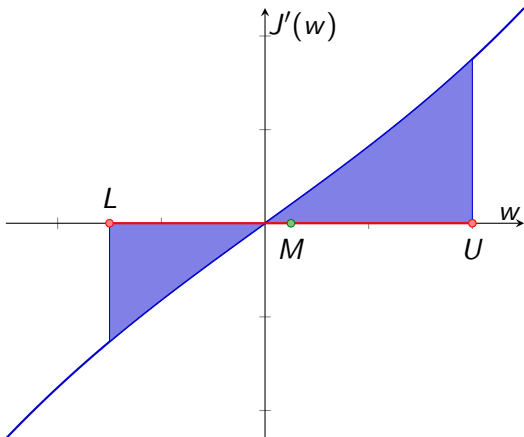


- Identify the point where the increasing function  $J'$  crosses zero

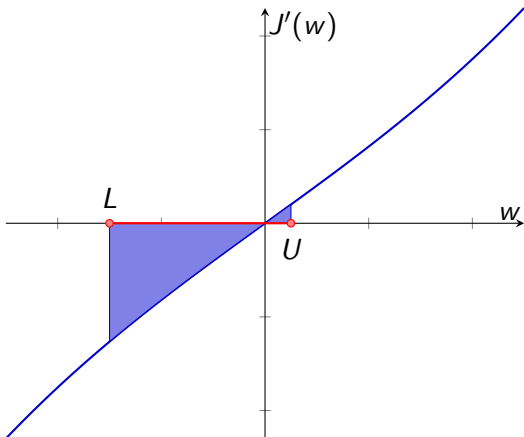
# Bisection Algorithm



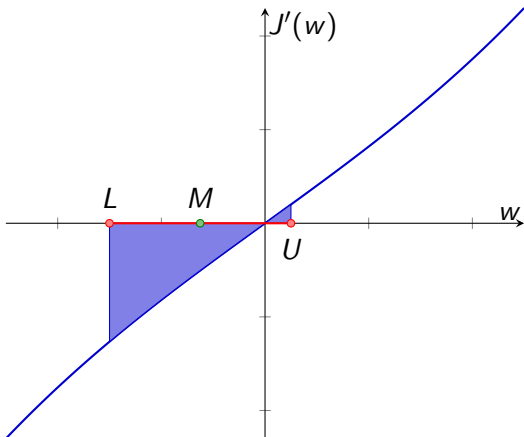
# Bisection Algorithm



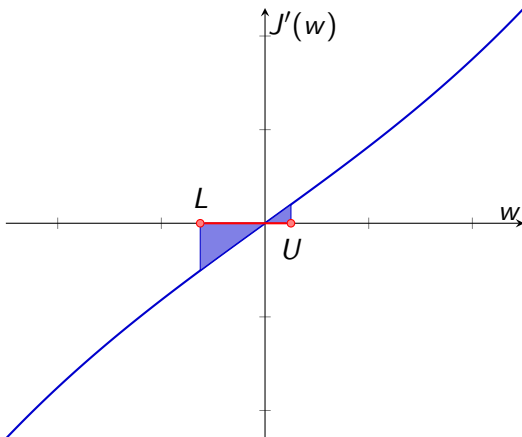
# Bisection Algorithm



# Bisection Algorithm



# Bisection Algorithm



## Interval Bisection

**Require:**  $L, U, \epsilon$

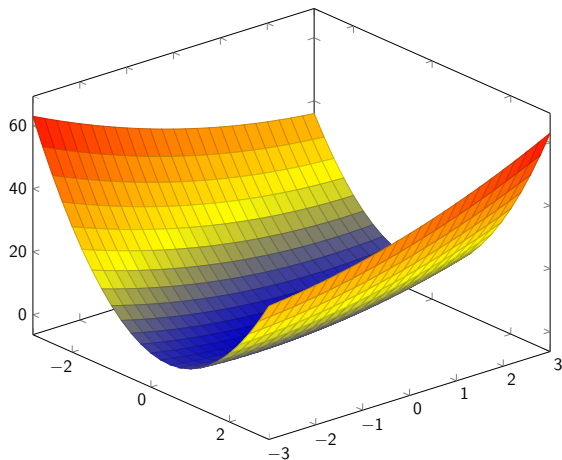
```
1:  $maxgrad \leftarrow J'(U)$ 
2: while  $(U - L) \cdot maxgrad > \epsilon$  do
3:    $M \leftarrow \frac{U+L}{2}$ 
4:   if  $J'(M) > 0$  then
5:      $U \leftarrow M$ 
6:   else
7:      $L \leftarrow M$ 
8:   end if
9: end while
10: return  $\frac{U+L}{2}$ 
```



# Outline

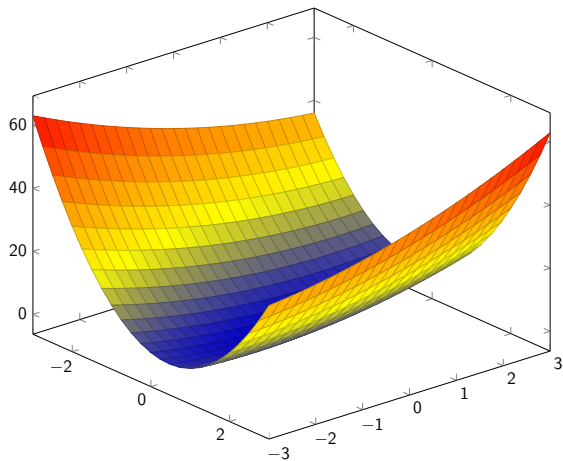
- 1 Convex Functions and Sets
- 2 Operations Which Preserve Convexity
- 3 First Order Properties
- 4 Subgradients
- 5 Constraints
- 6 Warmup: Minimizing a 1-d Convex Function
- 7 Warmup: Coordinate Descent**

# Problem Statement



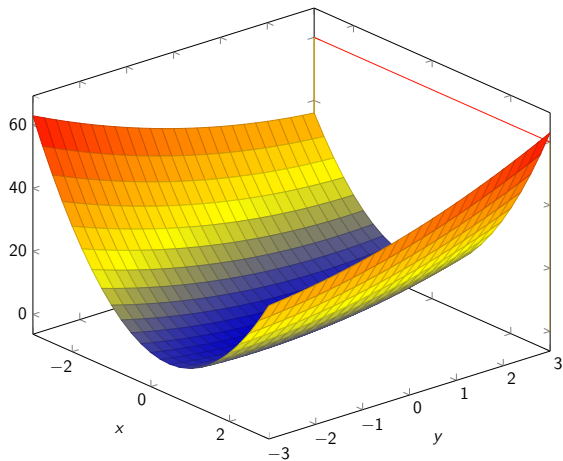
- Given a black-box which can compute  $J: \mathbb{R}^n \rightarrow \mathbb{R}$  and  $J': \mathbb{R}^n \rightarrow \mathbb{R}^n$  find the minimum value of  $J$

## Concrete Example



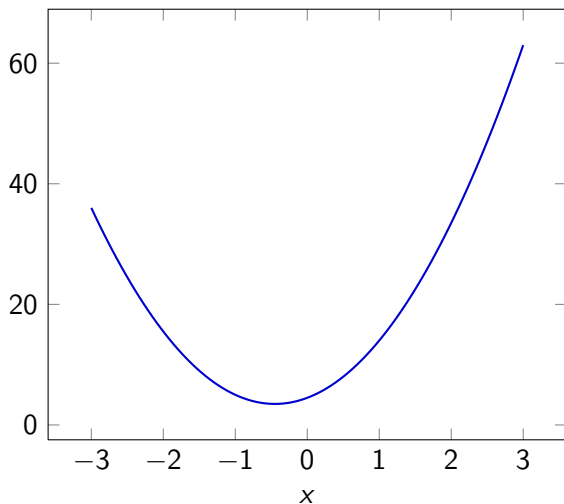
$$f(x, y) = \frac{1}{2} \begin{bmatrix} x, y \end{bmatrix} \begin{bmatrix} 10, 1 \\ 2, 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

## Concrete Example



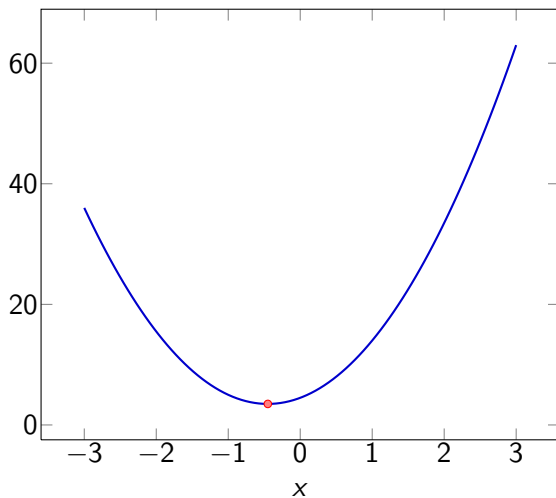
$$f(x, 3) = \frac{1}{2} \begin{bmatrix} x, 3 \end{bmatrix} \begin{bmatrix} 10, 1 \\ 2, 1 \end{bmatrix} \begin{bmatrix} x \\ 3 \end{bmatrix}$$

## Concrete Example



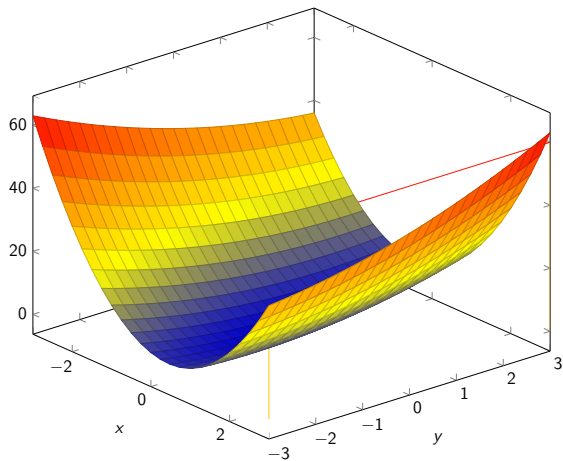
$$f(x, 3) = 5x^2 + \frac{9}{2}x + \frac{9}{2}$$

## Concrete Example

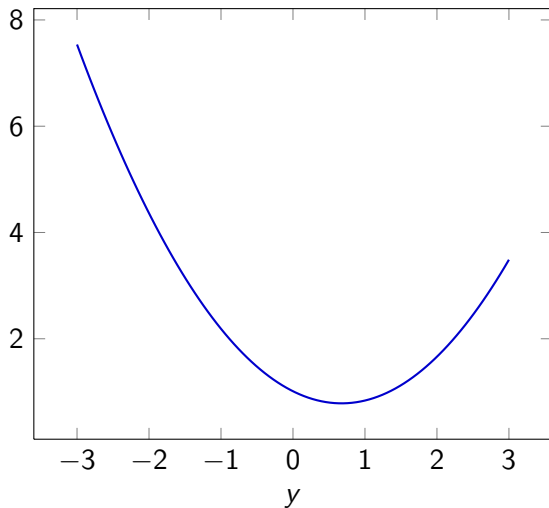


$$f(x, 3) = 5x^2 + \frac{9}{2}x + \frac{9}{2} \quad \text{Minima: } x = -\frac{9}{20}$$

# Concrete Example



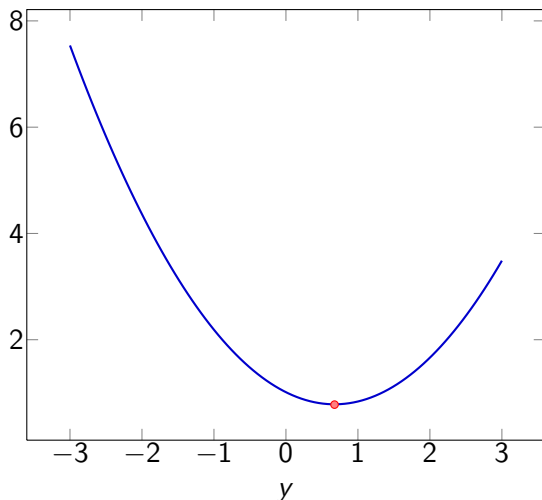
## Concrete Example



$$f\left(-\frac{9}{20}, y\right) = \frac{1}{2}y^2 - \frac{27}{40}y + \frac{81}{80}$$

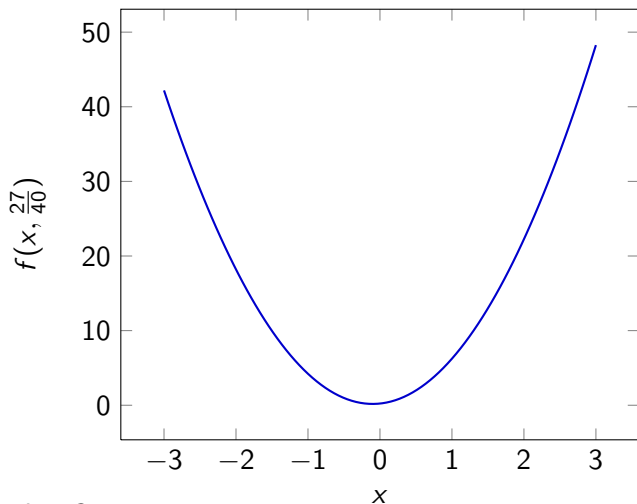


## Concrete Example



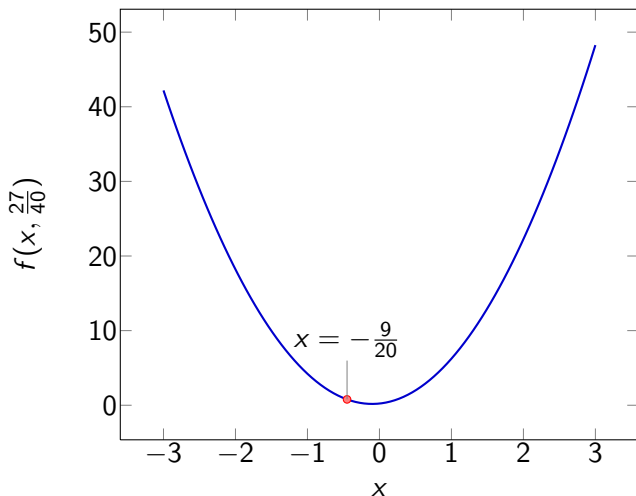
$$f\left(-\frac{9}{20}, y\right) = \frac{1}{2}y^2 - \frac{27}{40}y + \frac{81}{80} \quad \text{Minima: } y = \frac{27}{40}$$

## Concrete Example



- Are we done?

## Concrete Example



- Are we done?