Center for Data Science
NYU

DS-GA 3001.007
Introduction to Machine Learning

Lecture 10

Support Vector Machines – Classifying with Hinge Loss

Extending Perceptron Algorithm by Incorporating Margins

DS-GA 3001.007

Introduction to Machine Learning

Lecture 10

Support Vector Machines – Classifying with Hinge Loss

Announcements

- Homework 4 extended to **Wednesday November 13** at 11:59pm
- Survey 3 due **Sunday November 10** at 11:59pm
- Project
  - Milestone due **Thursday November 28** at 11:59pm
  - Background
  - Plans
    - Description of Methodology
    - Proposed Experiments
    - Some Relevant Datasets

Review: Loss Functions for Classification

- Notation

  Outcome space $\mathcal{Y} = \{-1, 1\}$

  Action space $\mathcal{A} = \{-1, 1\}$

- 0-1 Loss

$$\ell(f(x), y) = 1(f(x) \neq y)$$

Review: Loss Functions for Classification

- Notation

  Outcome space $\mathcal{Y} = \{-1, 1\}$

  Action space $\mathcal{A} = \{-1, 1\}$

- 0-1 Loss

$$\ell(f(x), y) = 1(f(x) \neq y)$$

Does not capture certainty about the classification

- Notation

  Output space $\mathcal{Y} = \{-1, 1\}$

  Action space $\mathcal{A} = \mathbf{R}$

Review: Loss Functions for Classification

- Notation

  Outcome space $\mathcal{Y} = \{-1, 1\}$

  Action space $\mathcal{A} = \{-1, 1\}$

- 0-1 Loss

$$\ell(f(x), y) = 1(f(x) \neq y)$$

Does not capture certainty about the classification

- Notation

  Output space $\mathcal{Y} = \{-1, 1\}$

  Action space $\mathcal{A} = \mathbf{R}$

- Margin
  - For prediction f(x) and label $y \in \{-1, 1\}$ is f(x) y
  - Same sign means positive value. Different sign means negative
  - Positive means correct. Negative means incorrect.

Review: Loss Functions for Classification

- Notation

  Outcome space $\mathcal{Y} = \{-1, 1\}$

  Action space $\mathcal{A} = \{-1, 1\}$

- 0-1 Loss

$$\ell(f(x), y) = 1(f(x) \neq y)$$

Does not capture certainty about the classification

- Notation

  Output space $\mathcal{Y} = \{-1, 1\}$

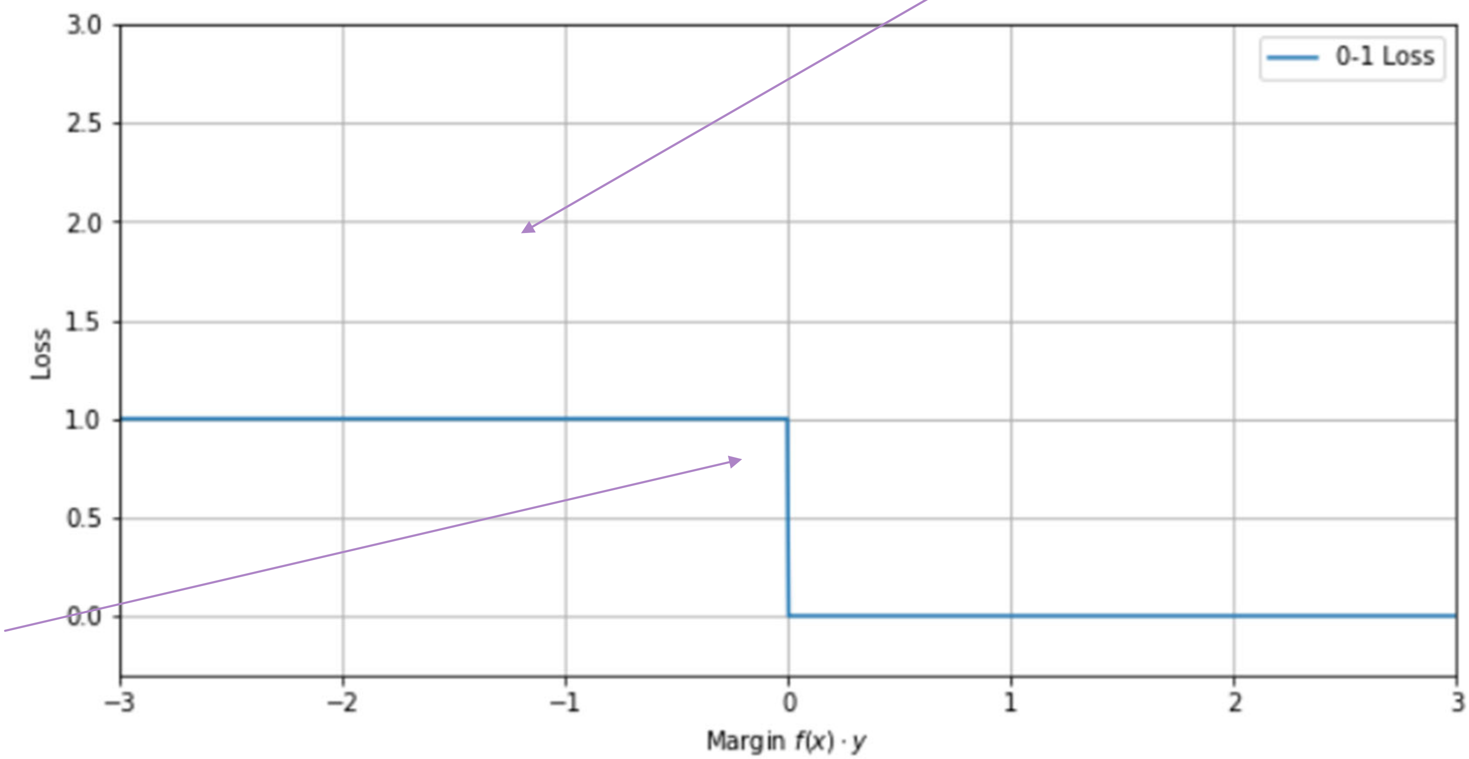
  Action space $\mathcal{A} = \mathbf{R}$

Functional Margin
not Geometric Margin

- Margin
  - For prediction f(x) and label $y \in \{-1, 1\}$ is f(x) y
  - Same sign means positive value. Different sign means negative
  - Positive means correct. Negative means incorrect.

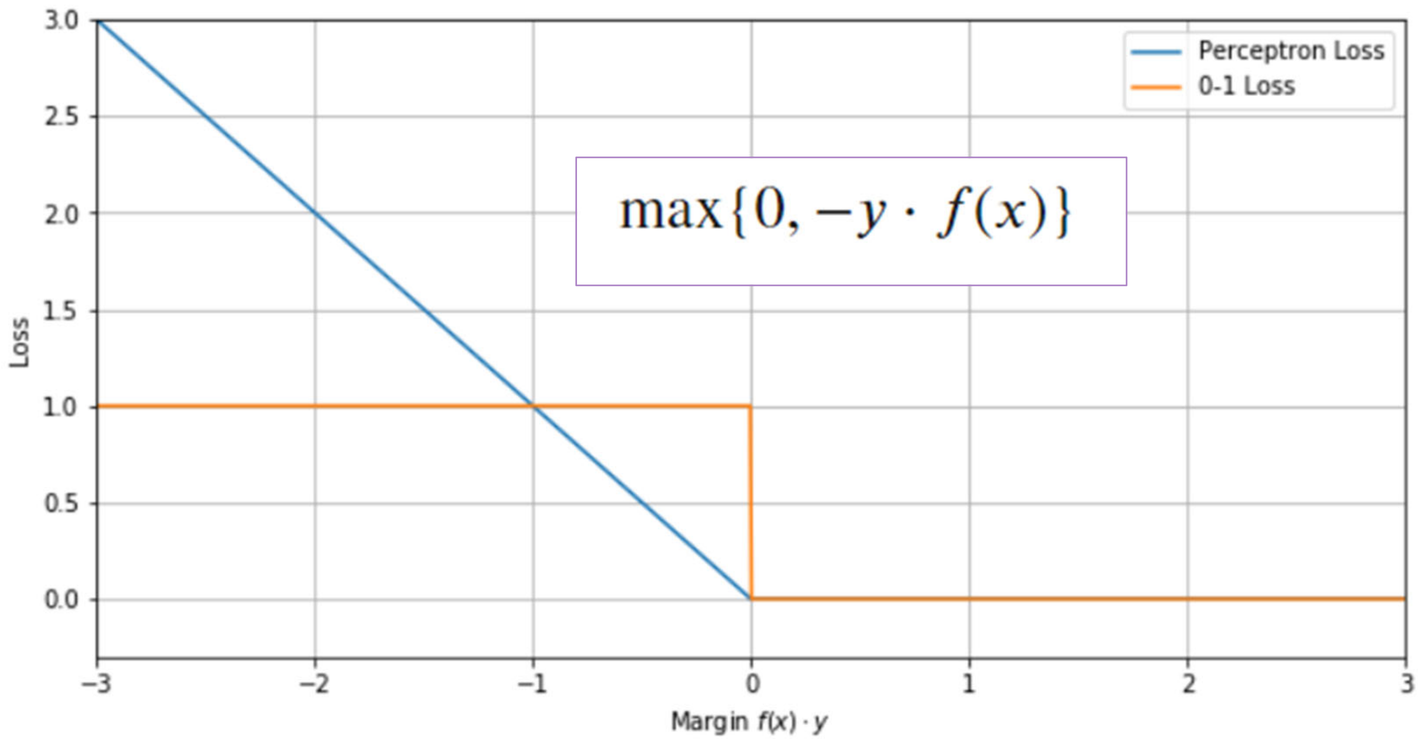
Review: Loss Functions for Classification

Not convex meaning no **subgradient** at decision boundary

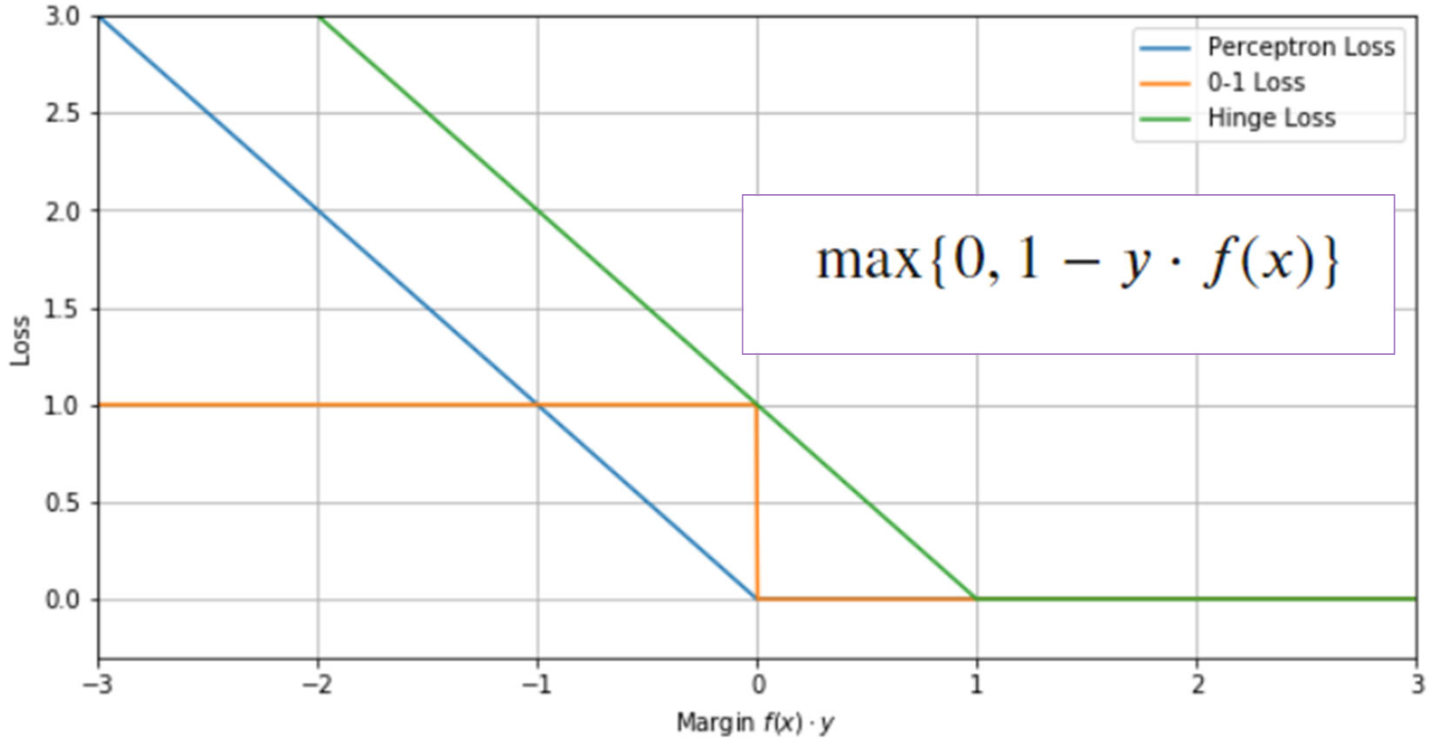


No gradient at decision boundary

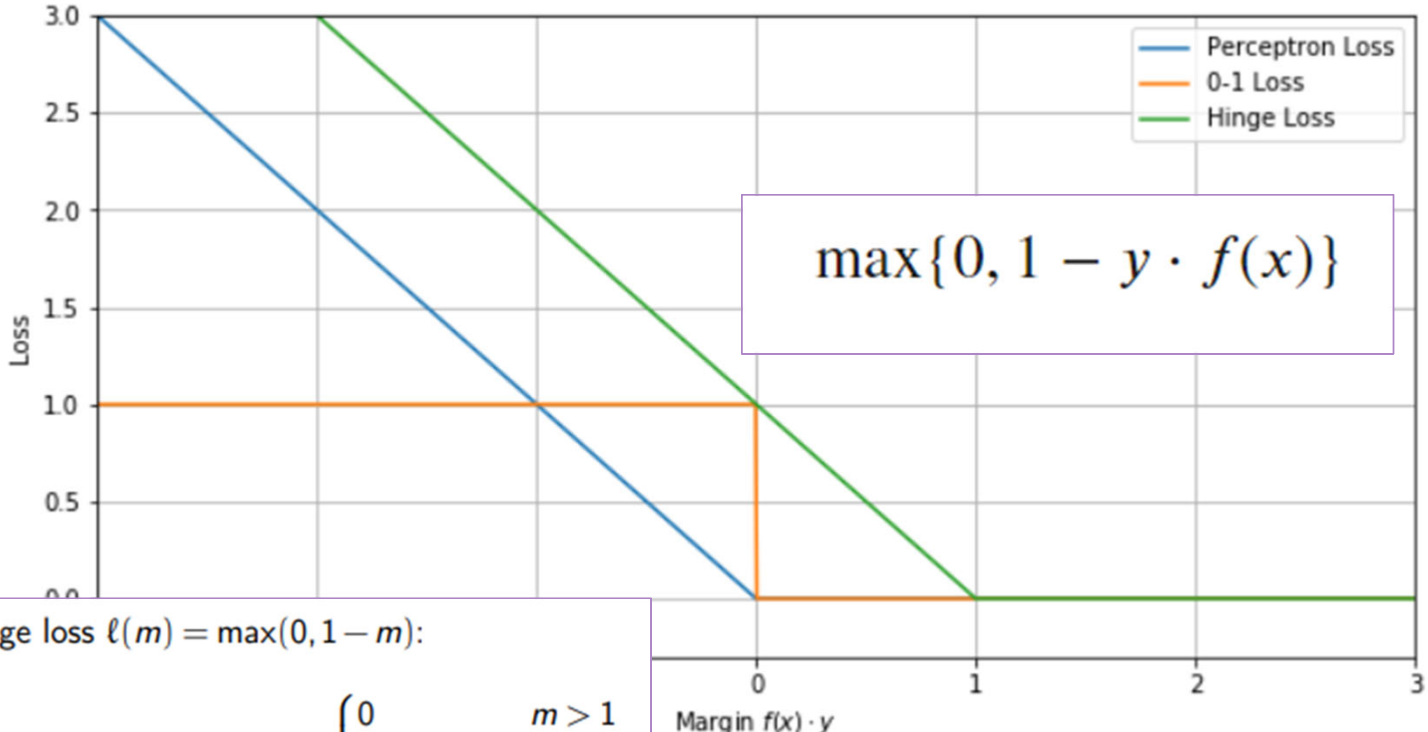
Review: Loss Functions for Classification



Review: Loss Functions for Classification



Review: Loss Functions for Classification

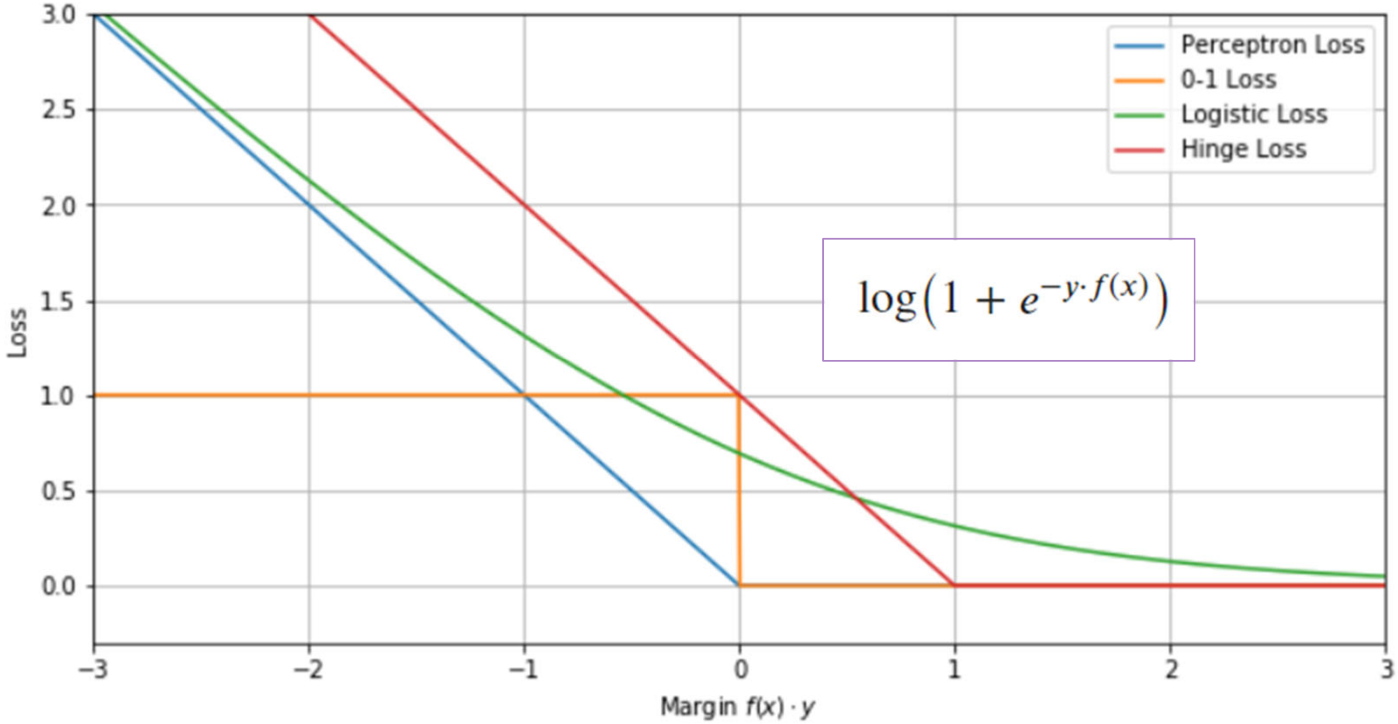


$$\max\{0, 1 - y \cdot f(x)\}$$

Derivative of hinge loss $\ell(m) = \max(0, 1 - m)$:

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

Review: Loss Functions for Classification



Agenda

- Lesson
  - Support Vector Machines
    - Hard Margin
    - Soft Margin
  - Convexity and Subgradients
  - Rearranging Optimization Problems
- Demo
  - Classifying Images with SVM

Objectives

- How can we generalize derivatives to non-differentiable functions
- What is the geometric way to understand SVM?
- How can we combine objective and constraint in a minimization problem?
- **Readings:**
  - Shalev-Schwarz Chapter 9
  - Boyd notes, Murphy Chapter 8.3

Agenda

We will cover next week

- Lesson
  - Support Vector Machines
    - Hard Margin
    - Soft Margin
  - Convexity and Subgradients
  - Rearranging Optimization Problems
- Demo
  - Classifying Images with SVM

Serkan will discuss in Section

Objectives

- How can we generalize derivatives to non-differentiable functions?
- What is the geometric way to understand SVM?
- How can we combine objective and constraint in a minimization problem?
- **Readings:**
  - Shalev-Schwarz Chapter 9
  - Boyd notes, Murphy Chapter 8.3

Agenda

Useful for classification and outlier detection

- Lesson
  - Support Vector Machines
    - Hard Margin
    - Soft Margin
  - Convexity and Subgradients
  - Rearranging Optimization Problems
- Demo
  - Classifying Images with SVM

Serkan will discuss in Section

Objectives

- How can we generalize derivatives to non-differentiable functions
- What is the geometric way to understand SVM?
- How can we combine objective and constraint in a minimization problem?
- **Readings:**
  - Shalev-Schwarz Chapter 9
  - Boyd notes, Murphy Chapter 8.3

Agenda

Useful for working with absolute value

- Lesson
  - Support Vector Machines
    - Hard Margin
    - Soft Margin
  - Convexity and Subgradients
  - Rearranging Optimization Problems
- Demo
  - Classifying Images with SVM

Serkan will discuss in Section

Objectives

- How can we generalize derivatives to non-differentiable functions
- What is the geometric way to understand SVM?
- How can we combine objective and constraint in a minimization problem?
- **Readings:**
  - Shalev-Schwarz Chapter 9
  - Boyd notes, Murphy Chapter 8.3

Agenda

- Lesson
  - Support Vector Machines
    - Hard Margin
    - Soft Margin
  - Convexity and Subgradients
  - Rearranging Optimization Problems
- Demo
  - Classifying Images with SVM

Useful for determining features

Serkan will discuss in Section

Objectives

- How can we generalize derivatives to non-differentiable functions
- What is the geometric way to understand SVM?
- How can we combine objective and constraint in a minimization problem?
- **Readings:**
  - Shalev-Schwarz Chapter 9
  - Boyd notes, Murphy Chapter 8.3

Support Vector Machines

Soft Margin

$$\min_{w \in \mathbf{R}^d, b \in \mathbf{R}} \frac{1}{2}||w||^2 + \frac{c}{n} \sum_{i=1}^{n} \max\left(0, 1 - y_i \left[w^T x_i + b\right]\right).$$

Support Vector Machines

Soft Margin

$$\min_{w \in \mathbf{R}^d, b \in \mathbf{R}} \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i\left[w^T x_i + b\right]\right).$$

Penalization form not constraint form with l2 regularization not l1 regularization

Support Vector Machines

Soft Margin

$$\min_{w \in \mathbf{R}^d, b \in \mathbf{R}} \frac{1}{2}||w||^2 + \frac{c}{n} \sum_{i=1}^{n} \max\left(0, 1 - y_i \left[w^T x_i + b\right]\right).$$

c not lambda

Penalization form not constraint form with l2 regularization not l1 regularization

Support Vector Machines

Soft Margin

b is intercept term in line…for classification with lines b is threshold

$$\min_{w \in \mathbf{R}^d, b \in \mathbf{R}} \frac{1}{2}||w||^2 + \frac{c}{n} \sum_{i=1}^{n} \max\left(0, 1 - y_i \left[w^T x_i + b\right]\right).$$

c not lambda

Penalization form not constraint form with l2 regularization not l1 regularization

Support Vector Machines

b is intercept term in line…for classification with lines b is threshold

Soft Margin

$$\min_{w \in \mathbf{R}^d, b \in \mathbf{R}} \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i\left[w^T x_i + b\right]\right).$$

c not lambda

Penalization form not constraint form with l2 regularization not l1 regularization

While w and b are unconstrained, the objective is not differentiable…so use make sense of gradient or rearrange

Support Vector Machines

Soft Margin

$$\min_{w \in \mathbf{R}^d, b \in \mathbf{R}} \frac{1}{2}||w||^2 + \frac{c}{n} \sum_{i=1}^{n} \max\left(0, 1 - y_i \left[w^T x_i + b\right]\right).$$

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}||w||^2 + \frac{c}{n} \sum_{i=1}^{n} \xi_i \\ \text{subject to} & \xi_i \geqslant \max\left(0, 1 - y_i \left[w^T x_i + b\right]\right). \end{array}$$

Support Vector Machines

Soft Margin

$$\min_{w \in \mathbf{R}^d, b \in \mathbf{R}} \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i\left[w^T x_i + b\right]\right).$$

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n}\xi_i \\ \text{subject to} & \xi_i \geqslant \max\left(0, 1 - y_i\left[w^T x_i + b\right]\right). \end{array}$$

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n}\xi_i \\ \text{subject to} & \xi_i \geqslant \left(1 - y_i\left[w^T x_i + b\right]\right) \text{ for } i = 1, \ldots, n \\ & \xi_i \geqslant 0 \text{ for } i = 1, \ldots, n \end{array}$$

Support Vector Machines

Soft Margin

$$\min_{w \in \mathbf{R}^d, b \in \mathbf{R}} \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i\left[w^T x_i + b\right]\right).$$

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i \\ \text{subject to} & \xi_i \geqslant \max\left(0, 1 - y_i\left[w^T x_i + b\right]\right). \end{array}$$

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i \\ \text{subject to} & \xi_i \geqslant \left(1 - y_i\left[w^T x_i + b\right]\right) \text{ for } i = 1, \ldots, n \\ & \xi_i \geqslant 0 \text{ for } i = 1, \ldots, n \end{array}$$

Differentiable with n + d + 1 unknowns

Support Vector Machines

Soft Margin

$$\min_{w \in \mathbf{R}^d, b \in \mathbf{R}} \frac{1}{2}||w||^2 + \frac{c}{n} \sum_{i=1}^{n} \max\left(0, 1 - y_i \left[w^T x_i + b\right]\right).$$

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n}\xi_i \\ \text{subject to} & \xi_i \geqslant \max\left(0, 1 - y_i\left[w^T x_i + b\right]\right). \end{array}$$

Quadratic Programming Problem…could solve with [CVXOPT]

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n}\xi_i \\ \text{subject to} & \xi_i \geqslant \left(1 - y_i\left[w^T x_i + b\right]\right) \text{ for } i = 1, \ldots, n \\ & \xi_i \geqslant 0 \text{ for } i = 1, \ldots, n \end{array}$$

Differentiable with n + d + 1 unknowns

Support Vector Machines

Derivative of hinge loss $\ell(m) = \max(0, 1 - m)$:

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

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

Support Vector Machines

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

Support Vector Machines

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

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

Support Vector Machines

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

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

Does it make sense to check this on the computer…with floating point numbers

Demo

- Support Vector Machine
  - Iris Dataset
  - Features
    - Petal Width
    - Petal Length
  - Classification
    - Iris-Versicolor
    - Iris-Setosa

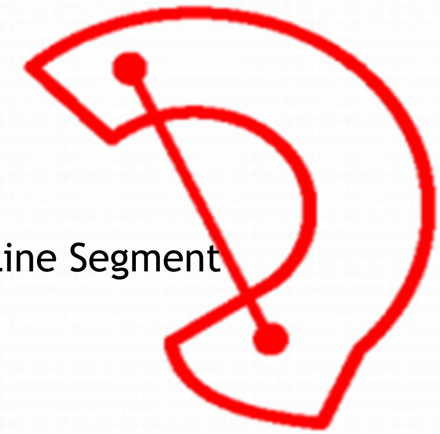
Take-Aways

- Why is SVM affected by scaling?
- How can soft margin SVM be used to detect outliers?
- How does changing C affect the classification? What prevents against overfitting.
- How do we use SVM in sklearn?

Convexity

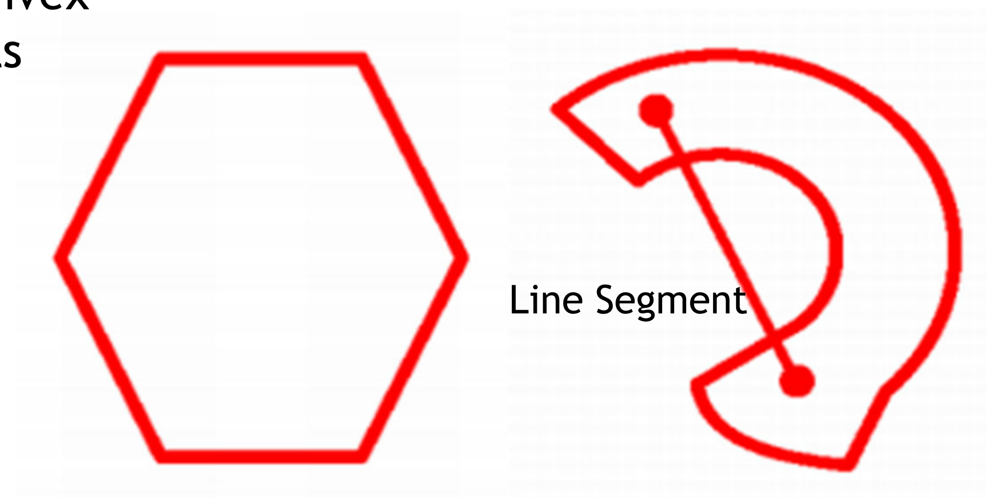
Convex
Sets

Line Segment

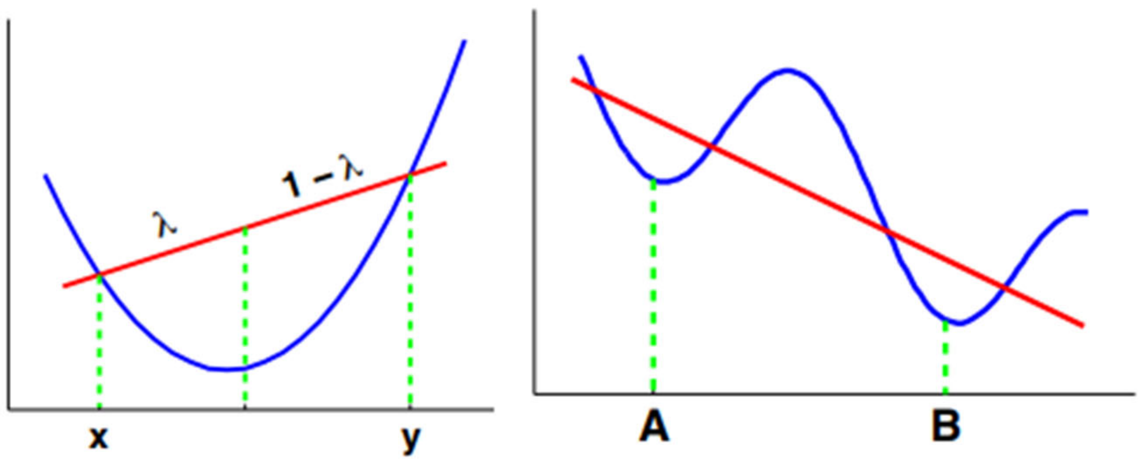


Convexity

Convex Sets



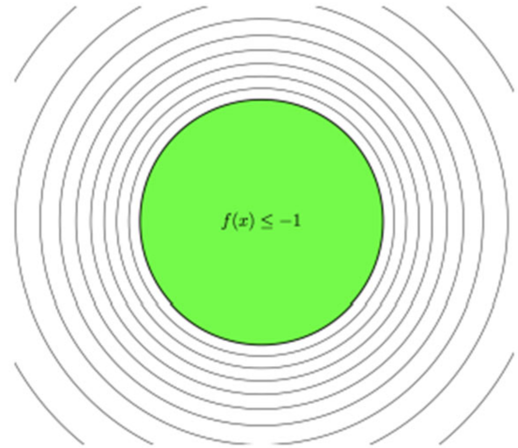
Convex Function



Question

- What is a concave function?
- Can a function be both convex and concave?

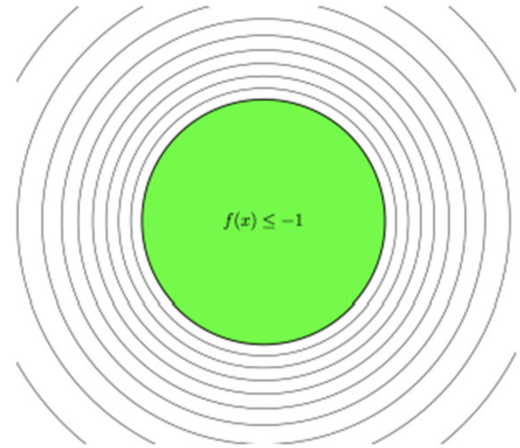
Convexity



$f : \mathbf{R}^d \to \mathbf{R}$ be a function.

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

Convexity

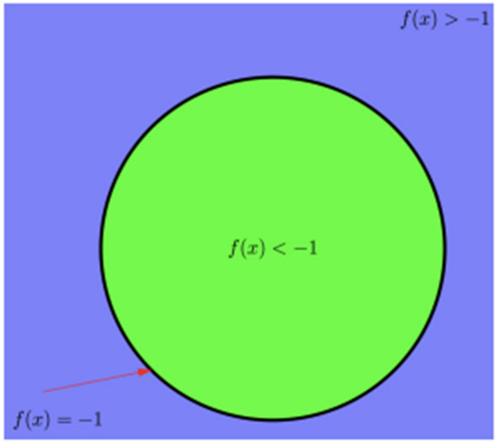


$f : \mathbf{R}^d \to \mathbf{R}$ be a function.

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

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

Convexity



$f : \mathbf{R}^d \to \mathbf{R}$ be a function.

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

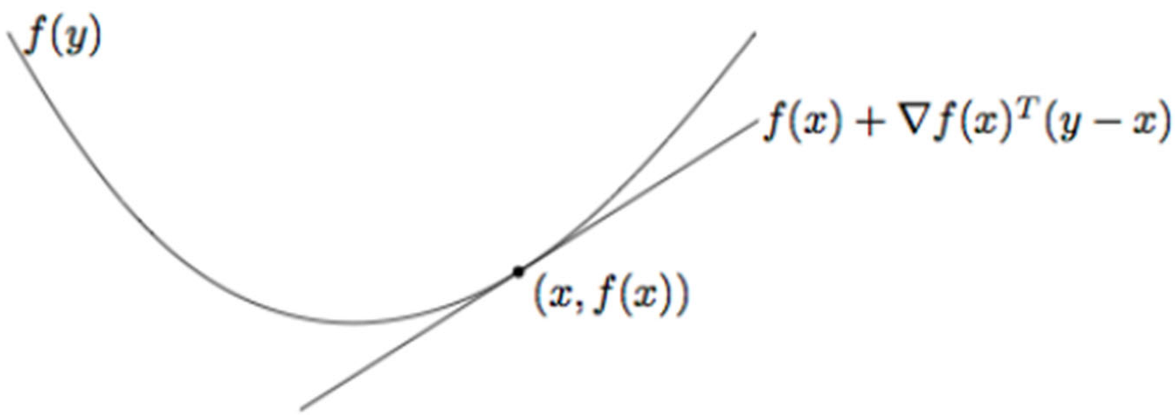
A **sublevel** set for the value c is the set of points $x \in \mathbf{R}^d$ for which $f(x) \leqslant c$.

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

Convexity

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:

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

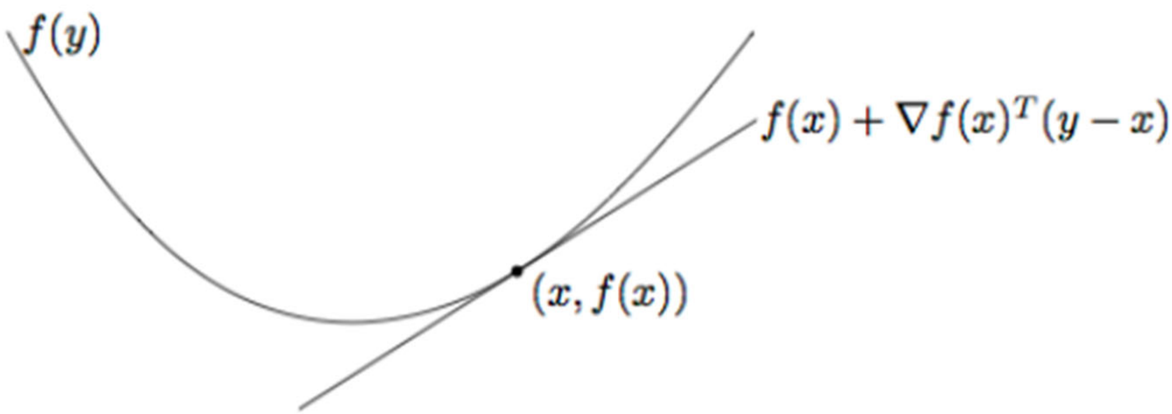


Convexity

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:

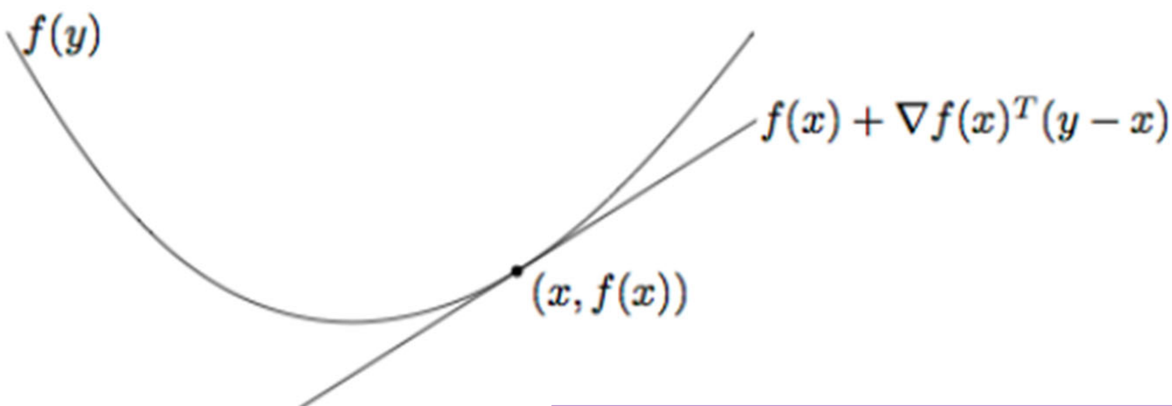


Convexity

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:

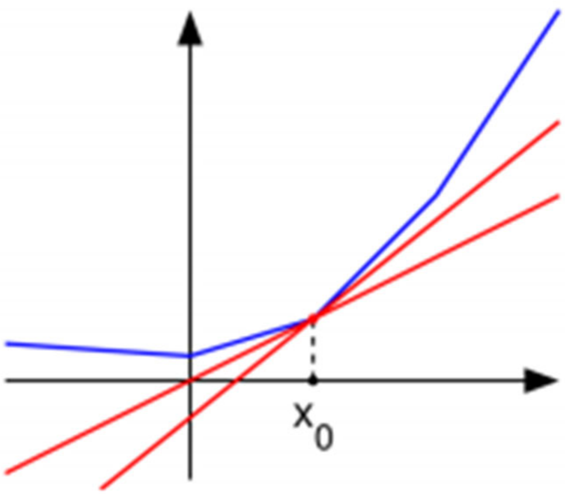


If $\nabla f(x) = 0$ then x is a global minimizer of f.

Subgradients

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

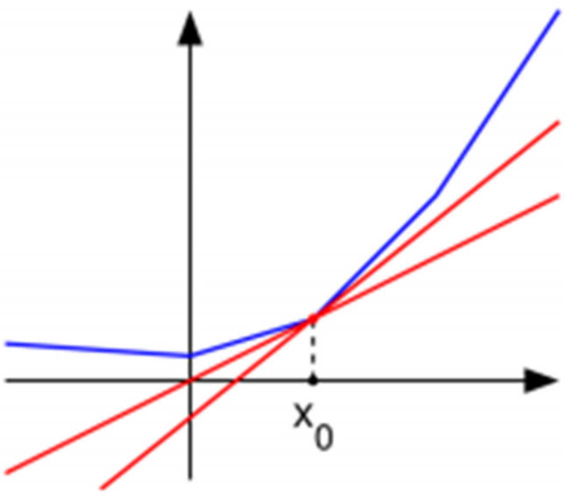
$$f(z) \geqslant f(x) + g^T (z - x).$$



Subgradients

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

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



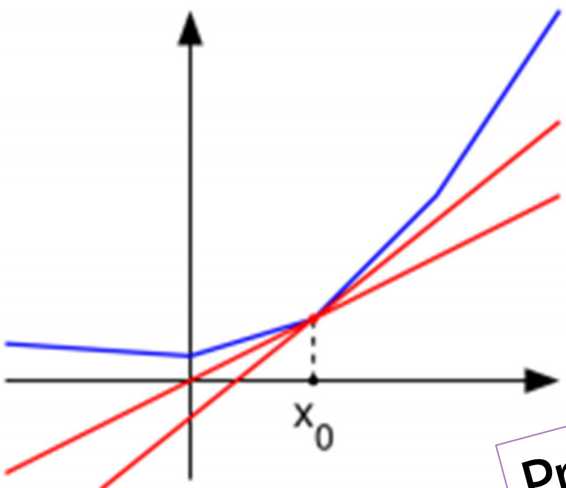
f is **subdifferentiable** at x if $\exists$ at least one subgradient at x.

The set of all subgradients at x is called the **subdifferential**: $\partial f(x)$

Subgradients

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

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



f is **subdifferentiable** at x if $\exists$ at least one subgradient at x.

The set of all subgradients at x is called the **subdifferential:** $\partial f(x)$

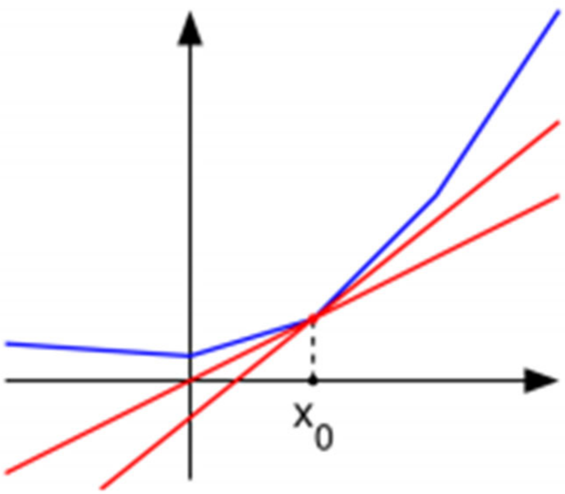
Properties

- 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.

Subgradients

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

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



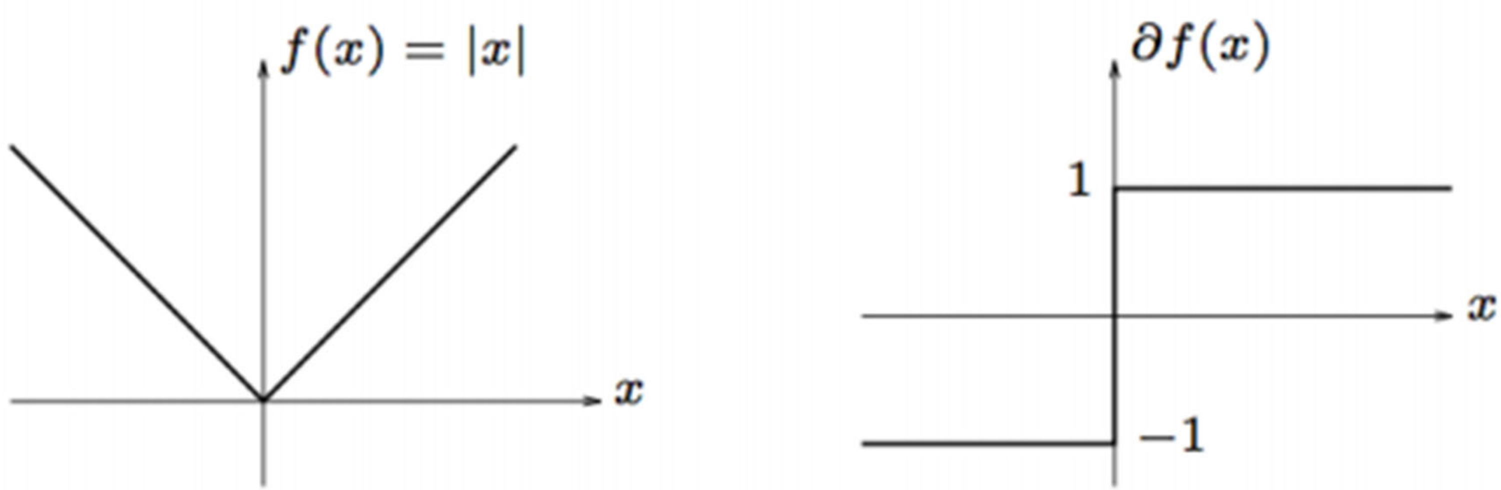
f is **subdifferentiable** at x if $\exists$ at least one subgradient at x.
The set of all subgradients at x is called the **subdifferential:** $\partial f(x)$

What if

$$0 \in \partial f(x)$$

- 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.

Subgradients



Subgradients

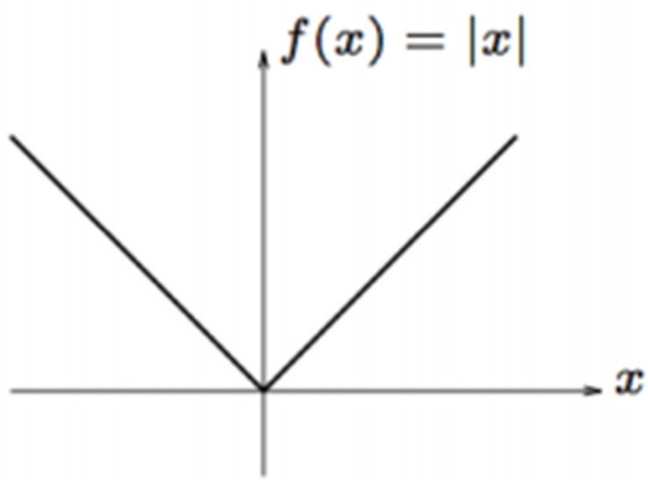
Question

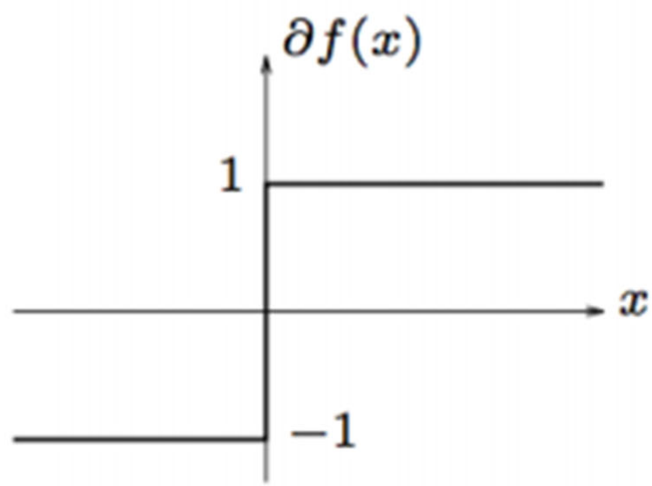
Let $\mathcal{X} = \{1, \ldots, 10\}$, let $\mathcal{Y} = \{1, \ldots, 10\}$, and let $A = \mathcal{Y}$. Suppose the data generating distribution, P, has marginal $X \sim \text{Unif}\{1, \ldots, 10\}$ and conditional distribution $Y|X = x \sim \text{Unif}\{1, \ldots, x\}$. For each loss function below give a target function

(a) $\ell(a, y) = (a - y)^2$,

(b) $\ell(a, y) = |a - y|$,

(c) $\ell(a, y) = \mathbb{1}(a \neq y)$.

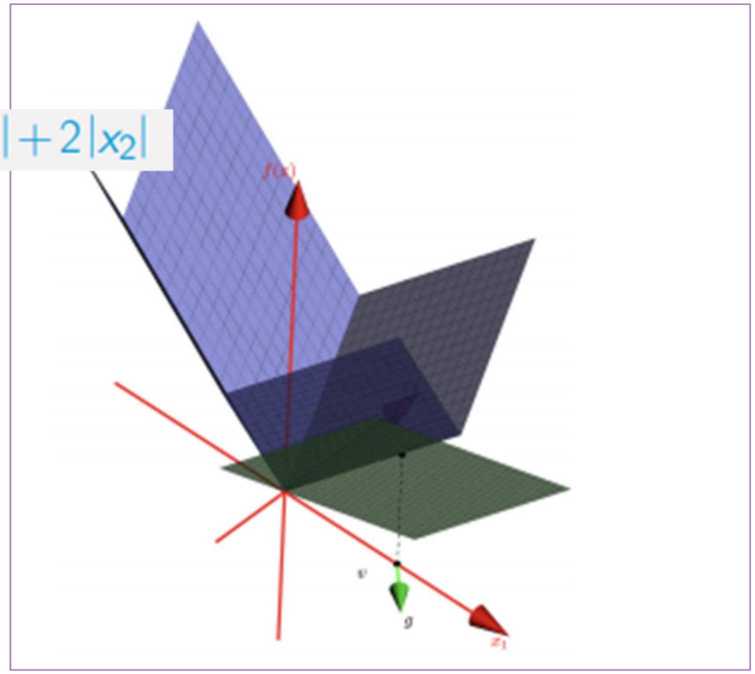




Subgradients

Example

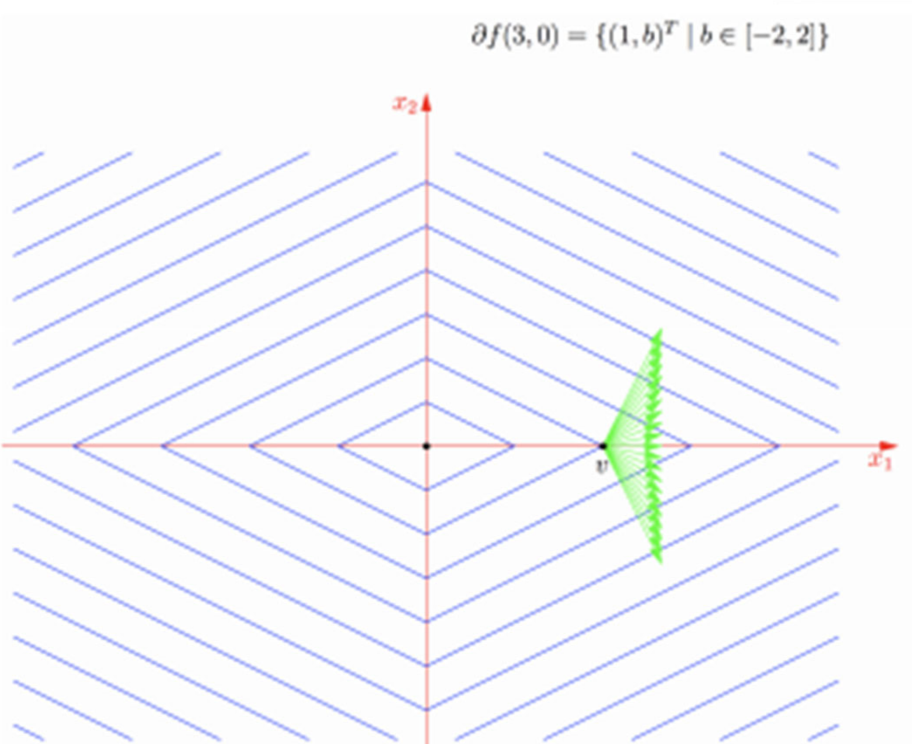
$$f(x_1, x_2) = |x_1| + 2|x_2|$$

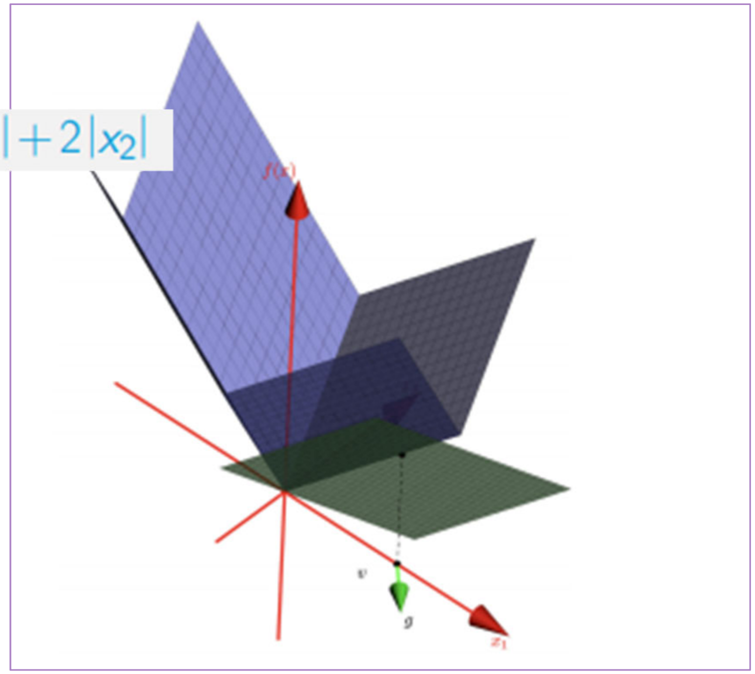


Subgradients

Example

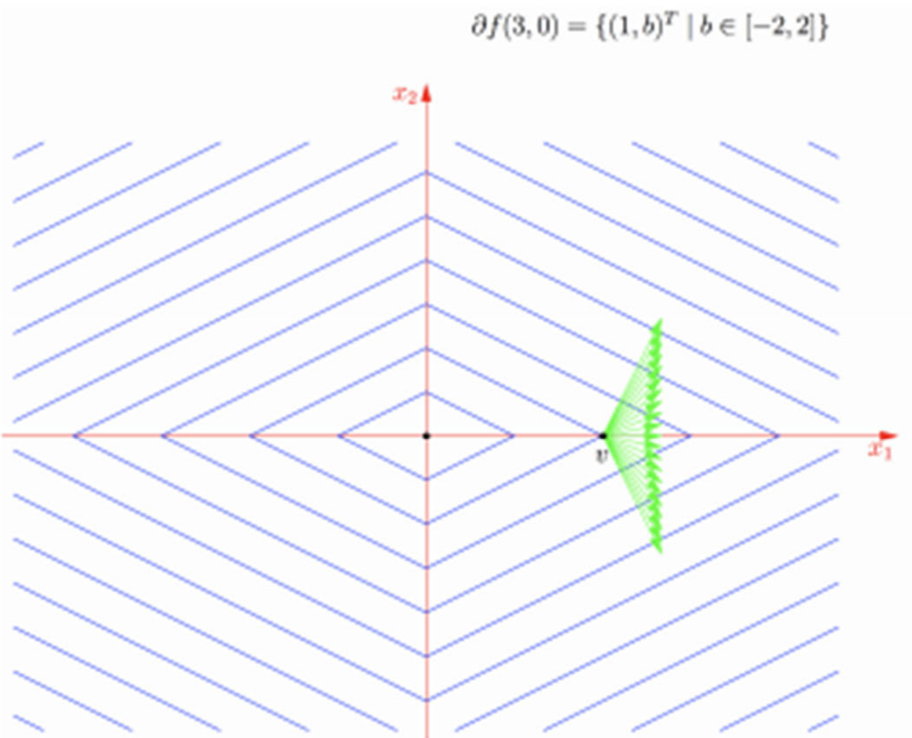
$= |x_1| + 2|x_2|$

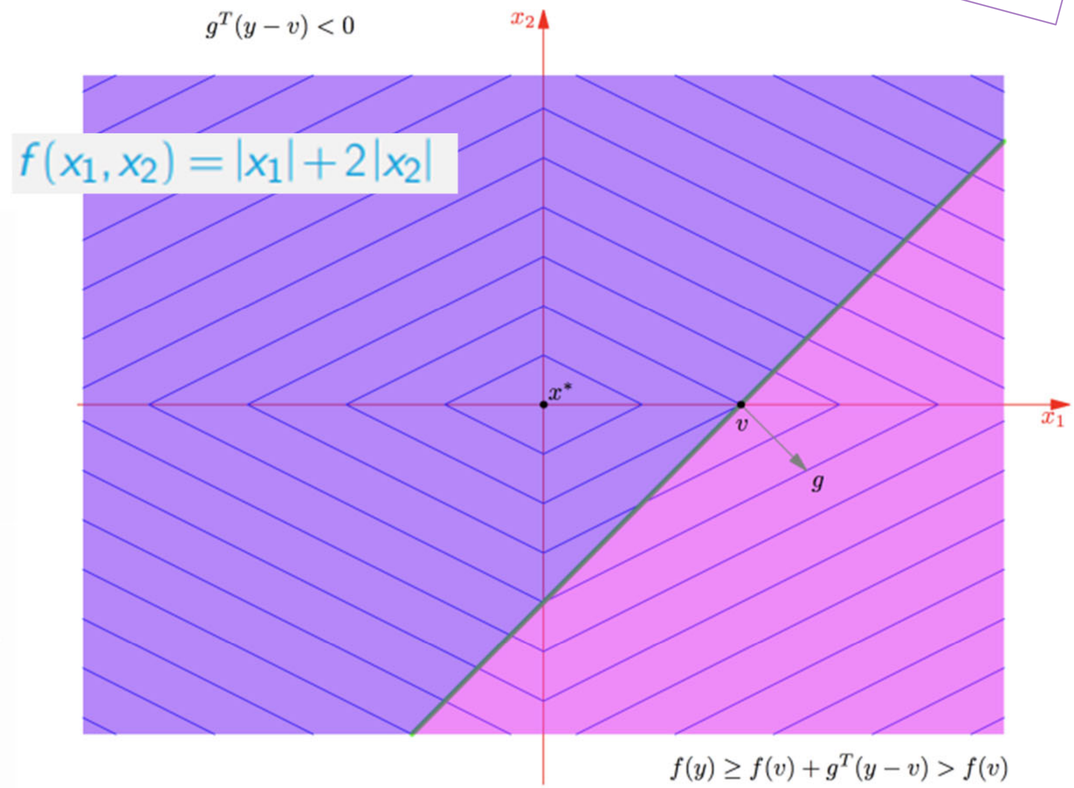




Subgradients

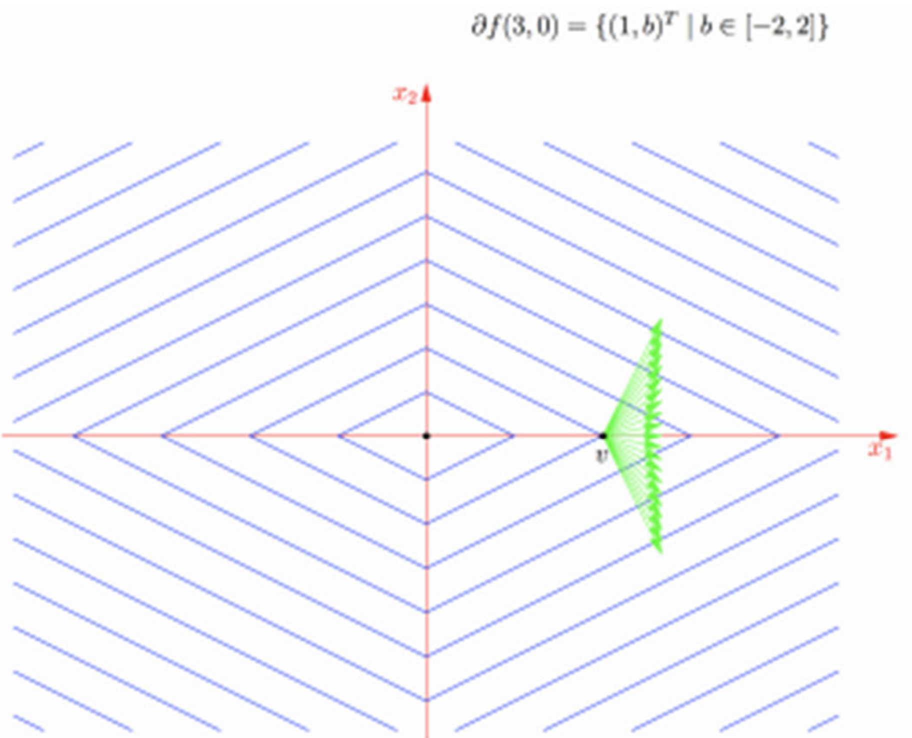
Example

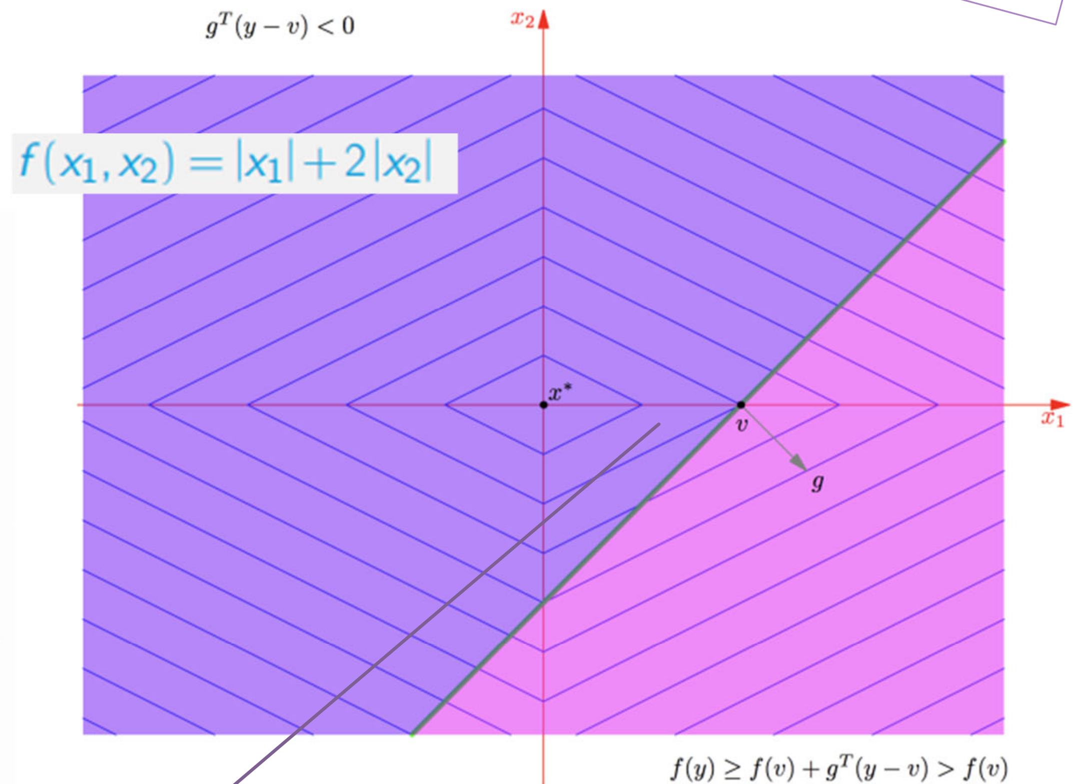




Subgradients

Example





-*g* is not a descent direction…the function might not decrease

Subgradient Descent

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 $t > 0$,*

$$\|x - z\|_2 < \|x_0 - z\|_2.$$

Subgradient Descent

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

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 $t > 0$,*

$$\|x - z\|_2 < \|x_0 - z\|_2.$$

Subgradient Descent

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 $t > 0$,*

$$\|x - z\|_2 < \|x_0 - z\|_2.$$

- 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

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

When are these terms negative?

Subgradient Descent

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 $t > 0$,*

$$\|x - z\|_2 < \|x_0 - z\|_2.$$

- 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

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

When are these terms negative?

- 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

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

Rearranging Optimization Problems

Example

- How to solve the Lasso?

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

- $\|w\|_1 = |w_1| + |w_2|$ is not differentiable!

Rearranging Optimization Problems

Example

- How to solve the Lasso?

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

- $\|w\|_1 = |w_1| + |w_2|$ is not differentiable!

- Replace each w_i by $w_i^+ - w_i^-$.
- Write $w^+ = (w_1^+, \ldots, w_d^+)$ and $w^- = (w_1^-, \ldots, w_d)$.

Rearranging Optimization Problems

Example

- How to solve the Lasso?

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

- $\|w\|_1 = |w_1| + |w_2|$ is not differentiable!

- Replace each w_i by $w_i^+ - w_i^-$.
- Write $w^+ = (w_1^+, \ldots, w_d^+)$ and $w^- = (w_1^-, \ldots, w_d^-)$.

$$\boxed{\begin{aligned} &\min_{w^+, w^- \in \mathbf{R}^d} \sum_{i=1}^{n} \left(\left(w^+ - w^-\right)^T x_i - y_i\right)^2 + \lambda 1^T \left(w^+ + w^-\right) \\ &\text{subject to } w_i^+ \geqslant 0 \text{ for all } i \\ &\qquad\qquad\quad w_i^- \geqslant 0 \text{ for all } i \end{aligned}}$$

Rearranging Optimization Problems

Example

- How to solve the Lasso?

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

- $\|w\|_1 = |w_1| + |w_2|$ is not differentiable!

Switching the order is helpful operation

$x_i \cdot$

$w \in \mathbf{R}^d \quad i=1$

- Replace each w_i by $w_i^+ - w_i^-$.
- Write $w^+ = (w_1^+, \ldots, w_d^+)$ and $w^- = (w_1^-, \ldots, w_d^-)$.

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

subject to $w_i^+ \geqslant 0$ for all i

$w_i^- \geqslant 0$ for all i

Rearranging Optimization Problems

Example

- Suppose you want to minimize penalization form

$$\min_x f(x) + \lambda g(x)$$

Rearranging Optimization Problems

Example

- Suppose you want to minimize penalization form

$$\min_x f(x) + \lambda g(x)$$

- Question: Is it equivalent to constraint form?

$$\min_x f(x)$$

$$\text{s.t. } g(x) \le R$$

Rearranging Optimization Problems

Example

- Suppose you want to minimize penalization form

$$\min_x f(x) + \lambda g(x)$$

- Question: Is it equivalent to constraint form?

$$\min_x f(x)$$

$$\text{s.t. } g(x) \leq R$$

- Suppose minimizer for penalization form is not minimizer for constraint form

$$\text{Set } R = g(x^*).$$

Rearranging Optimization Problems

Example

- Suppose you want to minimize penalization form

$$\min_x f(x) + \lambda g(x)$$

- Question: Is it equivalent to constraint form?

$$\min_x f(x)$$

$$\text{s.t. } g(x) \leq R$$

- Suppose minimizer for penalization form is not minimizer for constraint form

$$\text{Set } R = g(x^*).$$

- Note

$$f(x') < f(x^*)$$

Rearranging Optimization Problems

Example

- Suppose you want to minimize penalization form

$$\min_x f(x) + \lambda g(x)$$

- Question: Is it equivalent to constraint form?

$$\min_x f(x)$$

$$\text{s.t. } g(x) \leq R$$

- Suppose minimizer for penalization form is not minimizer for constraint form

$$\text{Set } R = g(x^*).$$

- Note

$$f(x') < f(x^*)$$

- Therefore

$$f(x') + \lambda g(x') < f(x^*) + \lambda g(x^*)$$

Rearranging Optimization Problems

Example

- Suppose you want to minimize penalization form

$$\min_x f(x) + \lambda g(x)$$

- Question: Is it equivalent to constraint form?

$$\min_x f(x)$$

$$\text{s.t. } g(x) \leq R$$

Combining objective function and constraint is helpful

- Suppose minimizer for penalization form is not minimizer for constraint form

$$\text{Set } R = g(x^*).$$

- Note

$$f(x') < f(x^*)$$

- Therefore

$$f(x') + \lambda g(x') < f(x^*) + \lambda g(x^*)$$

Rearranging Optimization Problems

Question

Suppose we have two functions $f : \mathbf{R}^d \to \mathbf{R}$ and $g : \mathbf{R}^d \to \mathbf{R}$. Now consider the following optimization problem:

$$\boxed{\min_{x \in \mathbf{R}^d} f(x) + g(x).}$$

Rearranging Optimization Problems

Question

Suppose we have two functions $f : \mathbf{R}^d \to \mathbf{R}$ and $g : \mathbf{R}^d \to \mathbf{R}$. Now consider the following optimization problem:

$$\boxed{\min_{x \in \mathbf{R}^d} f(x) + g(x).}$$

This is an unconstrained optimization problem. Let's also consider the following constrained optimization problem:

$$\boxed{\begin{array}{ll} \text{minimize} & f(x) + \xi \\ \text{subject to} & \xi \geq g(x). \end{array}}$$

Rearranging Optimization Problems

Question

Suppose we have two functions $f : \mathbf{R}^d \to \mathbf{R}$ and $g : \mathbf{R}^d \to \mathbf{R}$. Now consider the following optimization problem:

$$\boxed{\min_{x \in \mathbf{R}^d} f(x) + g(x).}$$

This is an unconstrained optimization problem. Let's also consider the following constrained optimization problem:

$$\boxed{\begin{array}{ll} \text{minimize} & f(x) + \xi \\ \text{subject to} & \xi \geq g(x). \end{array}}$$

Need to go both ways to have equivalent problem…there cannot be a gap

Rearranging Optimization Problems

Example

$$A = \begin{bmatrix} 5 & 5 & 5 & 5 & 5 \\ 8 & 8 & 1 & 8 & 8 \\ +\infty & +\infty & +\infty & 0 & +\infty \end{bmatrix}$$

Rearranging Optimization Problems

Example

$$A = \begin{bmatrix} 5 & 5 & 5 & 5 & 5 \\ 8 & 8 & 1 & 8 & 8 \\ +\infty & +\infty & +\infty & 0 & +\infty \end{bmatrix}$$

- We always have

$$\boxed{\max_j \min_i a_{ij} = d^* \leq p^* = \min_i \max_j a_{ij}.}$$

$$p^* = \min_i \max_j a_{ij}$$

$$d^* = \max_j \min_i a_{ij}$$

Rearranging Optimization Problems

Example

$$A = \begin{bmatrix} 5 & 5 & 5 & 5 & 5 \\ 8 & 8 & 1 & 8 & 8 \\ +\infty & +\infty & +\infty & 0 & +\infty \end{bmatrix}$$

$$p^* = \min_i \max_j a_{ij}$$

$$d^* = \max_j \min_i a_{ij}$$

- We always have

$$\boxed{\max_j \min_i a_{ij} = d^* \leq p^* = \min_i \max_j a_{ij}.}$$

because

$$d^* = a_{i_d j_d} \leq a_{i_p j_d} \leq a_{i_p j_p} = p^*.$$

Rearranging Optimization Problems

Example

$$A = \begin{bmatrix} 5 & 5 & 5 & 5 & 5 \\ 8 & 8 & 1 & 8 & 8 \\ +\infty & +\infty & +\infty & 0 & +\infty \end{bmatrix}$$

Primal Problem and Dual Problem may not be equal meaning you cannot switch max and min

$$p^* = \min_i \max_j a_{ij}$$

$$d^* = \max_j \min_i a_{ij}$$

- We always have

$$\boxed{\max_j \min_i a_{ij} = d^* \leq p^* = \min_i \max_j a_{ij}.}$$

because

$$d^* = a_{i_d j_d} \leq a_{i_p j_d} \leq a_{i_p j_p} = p^*.$$

Rearranging Optimization Problems

Example

- Minimize $x + y$ subject to constraint $x^2 + y^2 = 1$

Rearranging Optimization Problems

Example

- Minimize $x + y$ subject to constraint $x^2 + y^2 = 1$
- We can combine the objective and constraint into a single function called the Lagrangian

$$\boxed{L(x, y, \lambda) = x + y + \lambda(x^2 + y^2 - 1)}$$

Rearranging Optimization Problems

Example

- Minimize $x + y$ subject to constraint $x^2 + y^2 = 1$
- We can combine the objective and constraint into a single function called the Lagrangian

$$\boxed{L(x, y, \lambda) = x + y + \lambda(x^2 + y^2 - 1)}$$

- Take derivative to find minimum

$$\nabla L = \begin{pmatrix} \frac{\partial L}{\partial x} \\ \frac{\partial L}{\partial y} \\ \frac{\partial L}{\partial \lambda} \end{pmatrix} = \begin{pmatrix} 1 + 2\lambda x \\ 1 + 2\lambda y \\ x^2 + y^2 - 1 \end{pmatrix}$$

Rearranging Optimization Problems

Example

- Minimize $x + y$ subject to constraint $x^2 + y^2 = 1$
- We can combine the objective and constraint into a single function called the Lagrangian

$$\boxed{L(x, y, \lambda) = x + y + \lambda(x^2 + y^2 - 1)}$$

- Take derivative to find minimum

$$\nabla L = \begin{pmatrix} \frac{\partial L}{\partial x} \\ \frac{\partial L}{\partial y} \\ \frac{\partial L}{\partial \lambda} \end{pmatrix} = \begin{pmatrix} 1 + 2\lambda x \\ 1 + 2\lambda y \\ x^2 + y^2 - 1 \end{pmatrix}$$

- Solutions at

$$(\tfrac{\sqrt{2}}{2}, \tfrac{\sqrt{2}}{2}) \text{ and } (-\tfrac{\sqrt{2}}{2}, -\tfrac{\sqrt{2}}{2}).$$

Rearranging Optimization Problems

Example

- Minimize $x + y$ subject to constraint $x^2 + y^2 \leq 1$

Rearranging Optimization Problems

Example

- Minimize $x + y$ subject to constraint $x^2 + y^2 \leq 1$
- We can combine the objective and constraint into a single function called the Lagrangian

$$\boxed{L(x, y, \lambda) = x + y + \lambda\left(1 - (x^2 + y^2)\right)}$$

- Here $\lambda > 0$

Rearranging Optimization Problems

Example

- Minimize $x + y$ subject to constraint $x^2 + y^2 \leq 1$
- We can combine the objective and constraint into a single function called the Lagrangian

$$L(x, y, \lambda) = x + y + \lambda\left(1 - (x^2 + y^2)\right)$$

- Here $\lambda > 0$

Take max over the dual variables and min over the primal variables

Rearranging Optimization Problems

Example

- Minimize $x + y$ subject to constraint $x^2 + y^2 \leq 1$
- We can combine the objective and constraint into a single function called the Lagrangian

$$\boxed{L(x, y, \lambda) = x + y + \lambda\left(1 - (x^2 + y^2)\right)}$$

- Here $\lambda > 0$

So penalization form and constraint form are definitely the same!

Example

Rearranging Optimization Problems

- Minimize $x + y$ subject to constraint $x^2 + y^2 \leq 1$
- We can combine the objective and constraint into a single function called the Lagrangian

$$\boxed{L(x, y, \lambda) = x + y + \lambda\left(1 - (x^2 + y^2)\right)}$$

- Here $\lambda > 0$. Take derivative to find minimum

$$\nabla L = \begin{pmatrix} \frac{\partial L}{\partial x} \\ \frac{\partial L}{\partial y} \end{pmatrix} = \begin{pmatrix} 1 + 2\lambda x \\ 1 + 2\lambda y \end{pmatrix}$$

Example

Rearranging Optimization Problems

- Minimize $x + y$ subject to constraint $x^2 + y^2 \leq 1$
- We can combine the objective and constraint into a single function called the Lagrangian

$$\boxed{L(x, y, \lambda) = x + y + \lambda\left(1 - (x^2 + y^2)\right)}$$

- Here $\lambda > 0$. Take derivative to find minimum

$$\nabla L = \begin{pmatrix} \frac{\partial L}{\partial x} \\ \frac{\partial L}{\partial y} \end{pmatrix} = \begin{pmatrix} 1 + 2\lambda x \\ 1 + 2\lambda y \end{pmatrix}$$

- Solutions at

$$(\tfrac{\sqrt{2}}{2}, \tfrac{\sqrt{2}}{2}) \text{ and } (-\tfrac{\sqrt{2}}{2}, -\tfrac{\sqrt{2}}{2}).$$

Rearranging Optimization Problems

Example

- Minimize $x + y$ subject to constraint $x^2 + y^2 \leq 1$
- We can combine the objective and constraint into a single function called the Lagrangian

$$\boxed{L(x, y, \lambda) = x + y + \lambda\left(1 - (x^2 + y^2)\right)}$$

- Here $\lambda > 0$. Take derivative to find minimum

$$\nabla L = \begin{pmatrix} \frac{\partial L}{\partial x} \\ \frac{\partial L}{\partial y} \end{pmatrix} = \begin{pmatrix} 1 + 2\lambda x \\ 1 + 2\lambda y \end{pmatrix}$$

Note that solutions not unique. Is the objective convex? Is the objective concave?

- Solutions at

$$\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) \text{ and } \left(-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right).$$

Rearranging Optimization Problems

Example

- Minimize $x + y$ subject to constraint $x^2 + y^2 \leq 1$
- We can combine the objective and constraint into a single function called the Lagrangian

$$\boxed{L(x, y, \lambda) = x + y + \lambda\left(1 - (x^2 + y^2)\right)}$$

- Here $\lambda > 0$. Take derivative to find minimum

$$\nabla L = \begin{pmatrix} \frac{\partial L}{\partial x} \\ \frac{\partial L}{\partial y} \end{pmatrix} = \begin{pmatrix} 1 + 2\lambda x \\ 1 + 2\lambda y \end{pmatrix}$$

- Solutions at

$$(\tfrac{\sqrt{2}}{2}, \tfrac{\sqrt{2}}{2}) \text{ and } (-\tfrac{\sqrt{2}}{2}, -\tfrac{\sqrt{2}}{2}).$$

At the minimizer the constraint is satisfied…this is example of complementary slackness

Summary

- Support Vector Machines
  - Hard Margin: Only applies to linearly separable data
  - Soft Margin: Allows for slack variables. Useful for outlier detection
- Subgradients
  - Useful for convex functions. Takes any vector with properties of gradient.
  - Subgradient Descent variant of Gradient Descent
- Rearranging Optimization Problems
  - Combine objective and constraint
  - Switch order of minimization / maximization
  - Lagrangians, First Order Conditions and Complementary Slackness