NYU Center for Data Science

DS-GA 3001.007
Introduction to Machine Learning

Lecture 9

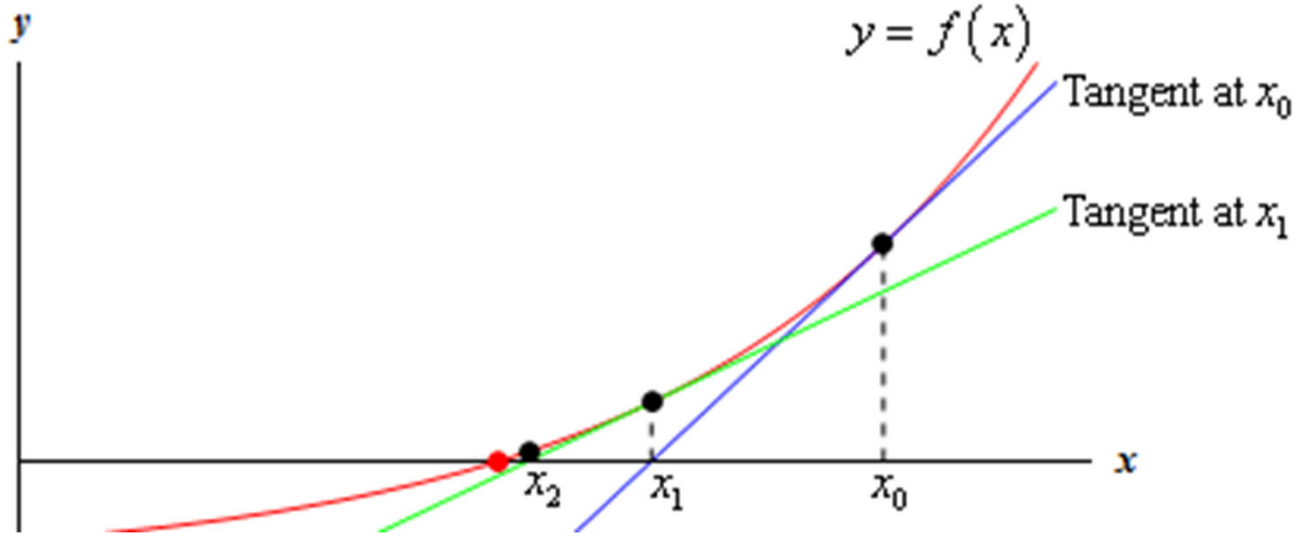
Loss Functions – Working With Approximations

Announcements

- Homework 4 due **Monday November 11** at 11:59pm
- Survey 3 due **Sunday November 10**
- Project 1 Proposal due **Sunday November 3.**
- Midterm
  - Grade determined by better of two
    - 15% Final
    - 10% Midterm or 25% Final

Review

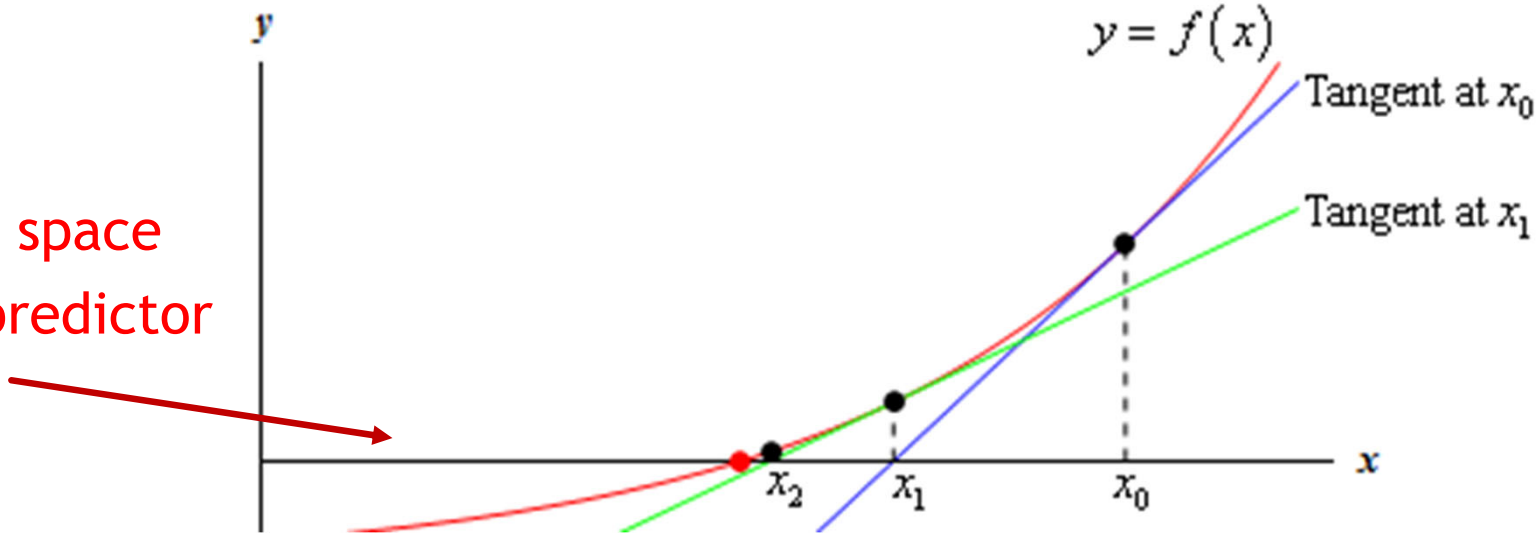
- Loss Function l(x)
  - Set f(x) = l'(x)
  - Find r such that f(r) = 0



Review

- Loss Function l(x)
  - Set f(x) = l'(x)
  - Find r such that f(r) = 0

In hypothesis space
r is optimal predictor

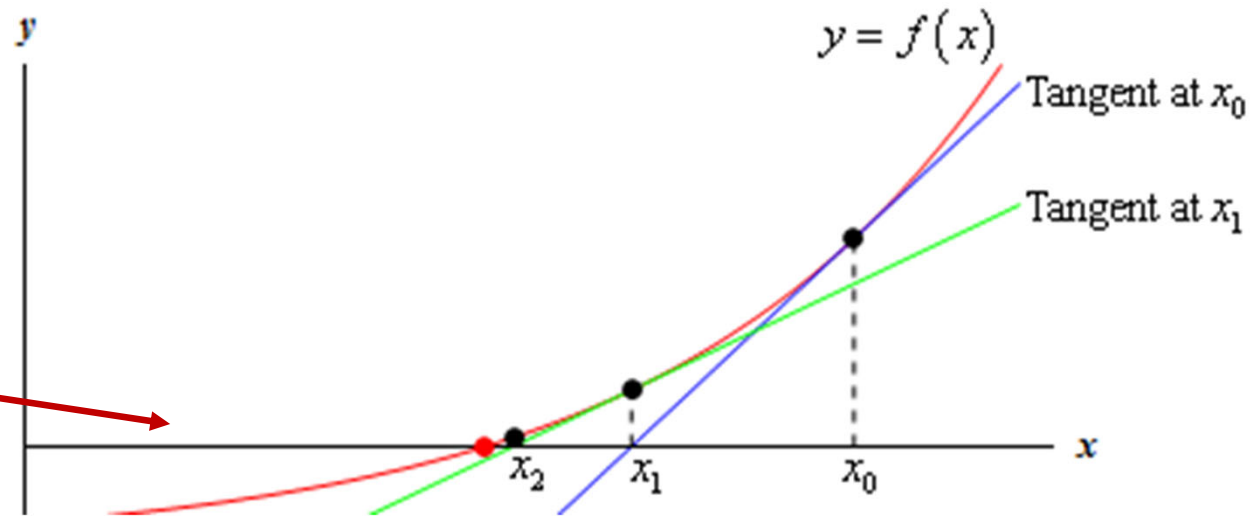


Review

- Loss Function l(x)
  - Set f(x) = l'(x)
  - Find r such that f(r) = 0
- Optimization
  - Set g(x) = x - (f(x) / f'(x))
  - Take x_{t+1} = g(x_t)

In hypothesis space

r is optimal predictor



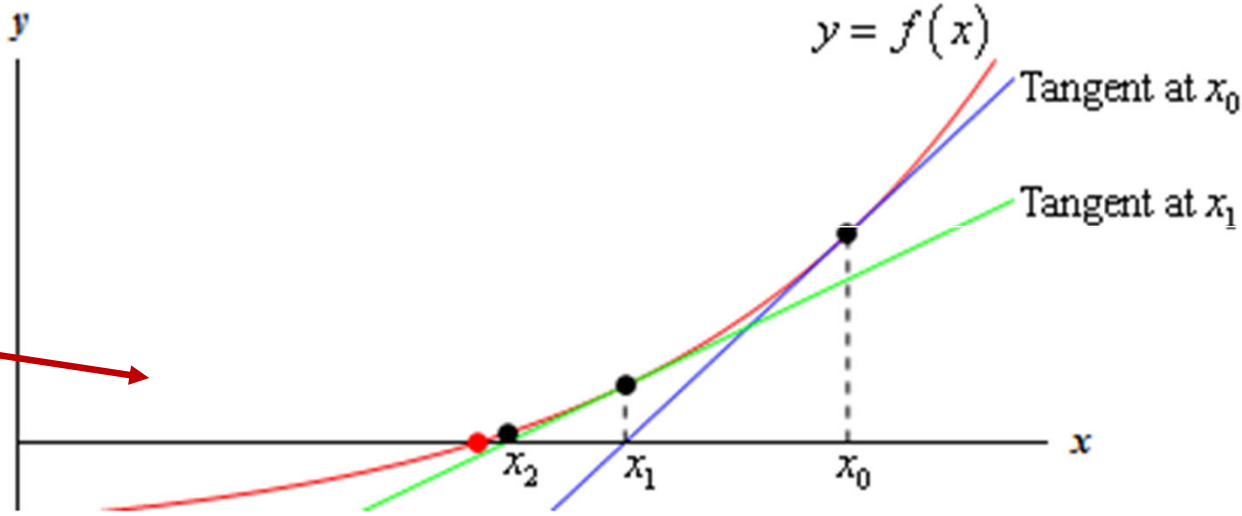
Review

- Loss Function l(x)
  - Set f(x) = l'(x)
  - Find r such that f(r) = 0
- Optimization
  - Set g(x) = x – m (f(x) / f'(x))
  - Take x_{t+1} = g(x_t)

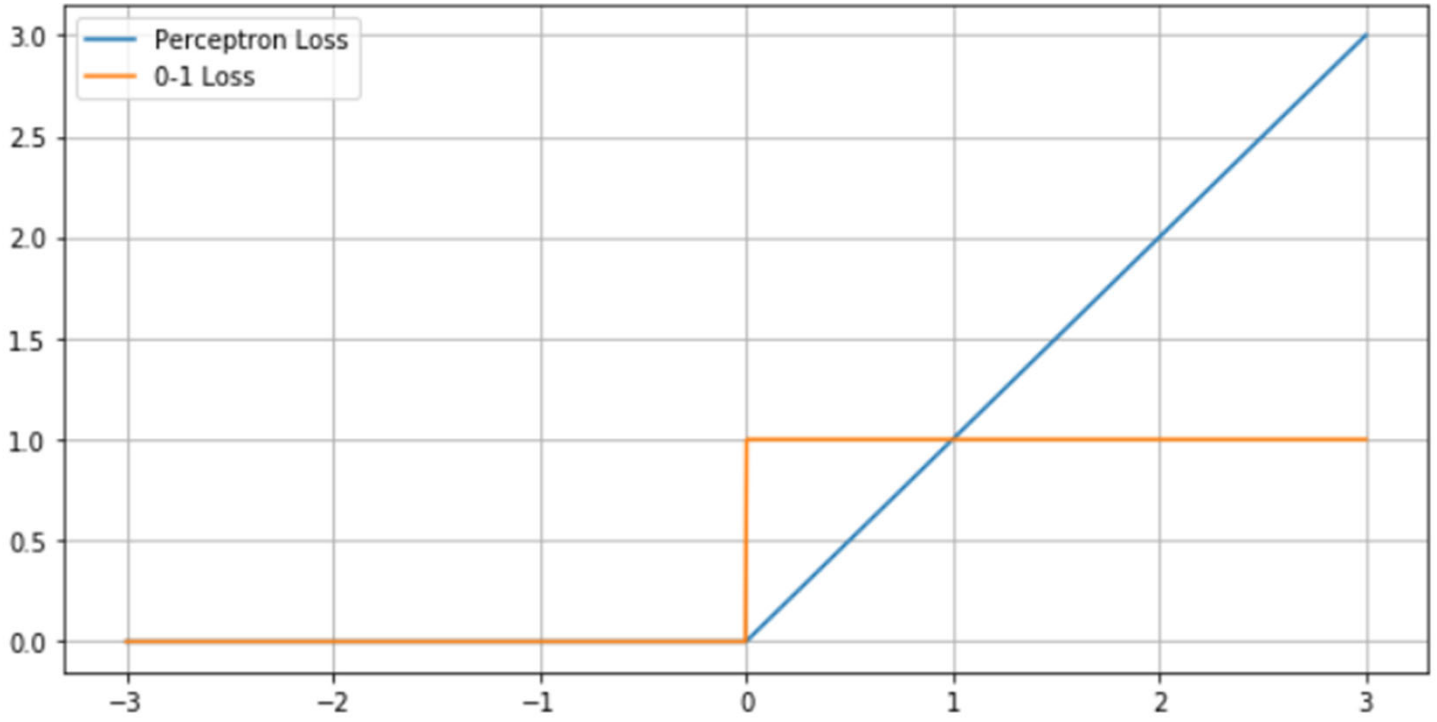
Regularization

In hypothesis space

r is optimal predictor



Review



Agenda

- Lesson
  - How can we work with loss functions?
    - Find a replacement
    - Approximate by different derivative
- Demos
  - How can we quantify our uncertainty in a classification?
    - Logistic Regression

Objectives

- What is distance based loss and margin based loss?
- What is a subdifferential? Can we use it for minimizing objective functions?
- **Readings**:
  - Shalev-Schwarz Chapter 9,

    Boyd notes, Murphy Chapter 8.3

Serkan will discuss in Section 09

Loss Functions for Regression

- Notation

  Input space $\mathcal{X} = \mathbf{R}^d$
  Action space $\mathcal{A} = \mathbf{R}$
  Outcome space $\mathcal{Y} = \mathbf{R}$.

Loss Functions for Regression

- Notation

  Input space $\mathcal{X} = \mathbf{R}^d$
  Action space $\mathcal{A} = \mathbf{R}$
  Outcome space $\mathcal{Y} = \mathbf{R}$.

- The steps are
  - Observe input x
  - Take action a
  - Observe outcome y
  - Evaluate loss l(a,y)

Loss Functions for Regression

- Notation

  Input space $\mathcal{X} = \mathbf{R}^d$
  Action space $\mathcal{A} = \mathbf{R}$
  Outcome space $\mathcal{Y} = \mathbf{R}$.

- The steps are
  - Observe input x
  - Take action a
  - Observe outcome y
  - Evaluate loss l(a,y)

- Prediction Function

$$\begin{array}{rccc} f: & \mathcal{X} & \to & \mathcal{A} \\ & x & \mapsto & f(x) \end{array}$$

Loss Functions for Regression

- Notation

  Input space $\mathcal{X} = \mathbf{R}^d$
  Action space $\mathcal{A} = \mathbf{R}$
  Outcome space $\mathcal{Y} = \mathbf{R}$.

- The steps are
  - Observe input x
  - Take action a
  - Observe outcome y
  - Evaluate loss l(a,y)

- Prediction Function

$$\begin{array}{rccc} f: & \mathcal{X} & \to & \mathcal{A} \\ & x & \mapsto & f(x) \end{array}$$

- Loss Function

$$\begin{array}{rccc} \ell: & \mathcal{A} \times \mathcal{Y} & \to & \mathbf{R} \\ & (a, y) & \mapsto & \ell(a, y) \end{array}$$

Loss Functions for Regression

- Notation

  Input space $\mathcal{X} = \mathbf{R}^d$
  Action space $\mathcal{A} = \mathbf{R}$
  Outcome space $\mathcal{Y} = \mathbf{R}$.

- The steps are
  - Observe input x
  - Take action a
  - Observe outcome y
  - Evaluate loss l(a,y)

- Prediction Function

$$\begin{array}{rccc} f: & \mathcal{X} & \to & \mathcal{A} \\ & x & \mapsto & f(x) \end{array}$$

- Loss Function

$$\begin{array}{rccc} \ell: & \mathcal{A} \times \mathcal{Y} & \to & \mathbf{R} \\ & (a, y) & \mapsto & \ell(a, y) \end{array}$$

- Here A = Y so

$\hat{y}$ is the predicted value (the action)
y is the actual observed value (the outcome)

Loss Functions for Regression

- **Distance based** loss means function

$$(\hat{y}, y) \mapsto \ell(\hat{y}, y) \in \mathbf{R}$$

only depends on the

difference

$$\textbf{residual } r = y - \hat{y}.$$

Loss Functions for Regression

- **Distance based** loss means function

$$(\hat{y}, y) \mapsto \ell(\hat{y}, y) \in \mathbf{R}$$

only depends on the difference

$$\textbf{residual } r = y - \hat{y}.$$

- Examples

Residual: $r = y - \hat{y}$

Square or ℓ_2 Loss: $\ell(r) = r^2$

Absolute or ℓ_1 Loss: $\ell(r) = |r|$

Loss Functions for Regression

- **Distance based** loss means function

$$(\hat{y}, y) \mapsto \ell(\hat{y}, y) \in \mathbf{R}$$

only depends on the difference

$$\textbf{residual } r = y - \hat{y}.$$

- Examples

Residual: $r = y - \hat{y}$

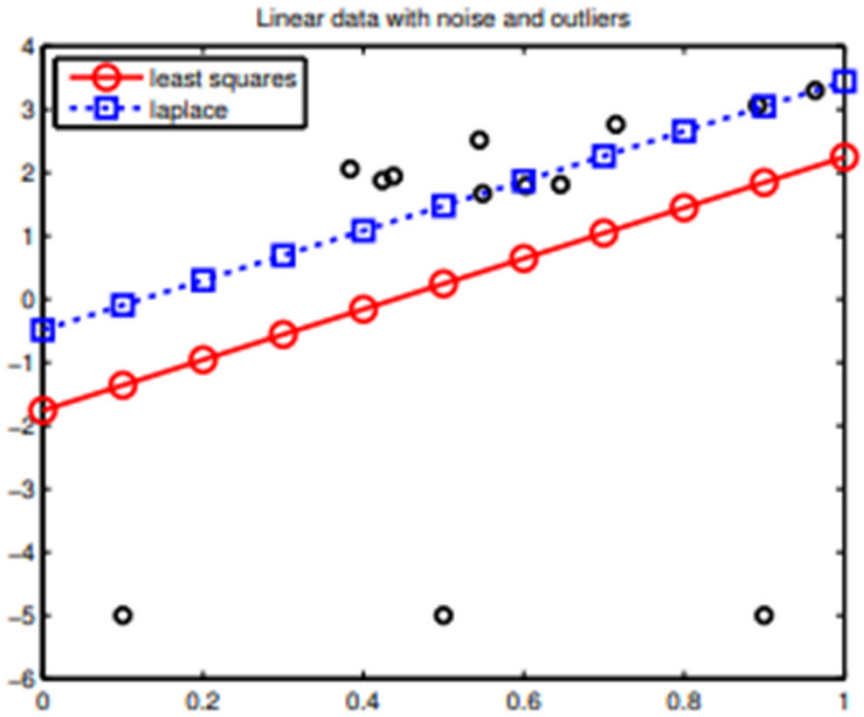
Square or ℓ_2 Loss: $\ell(r) = r^2$

Absolute or ℓ_1 Loss: $\ell(r) = |r|$

All distance based losses are translation invariant meaning

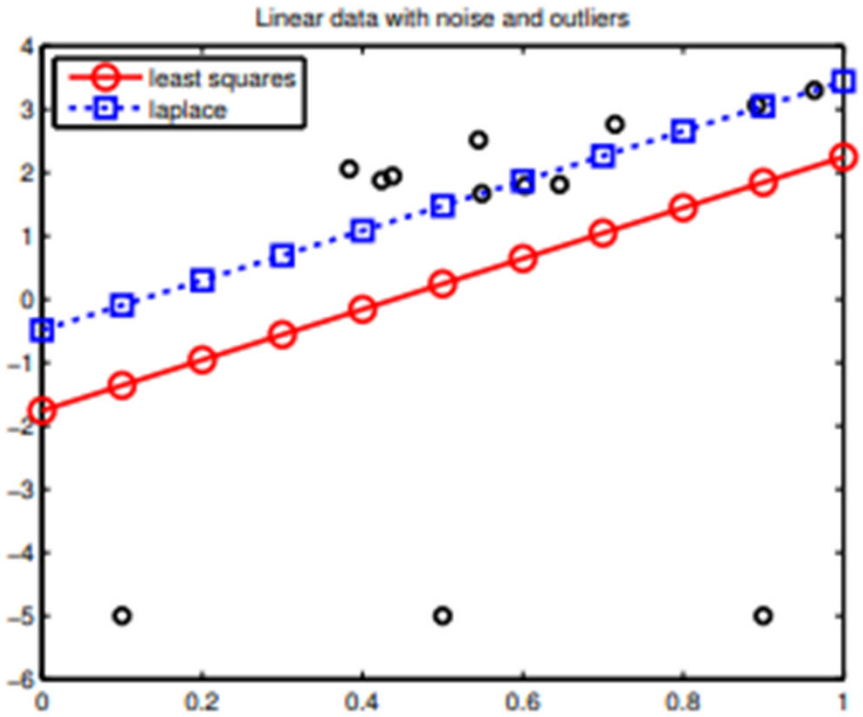
$$\ell(\hat{y} + b, y + b) = \ell(\hat{y}, y)$$

Differences between Regression Loss Functions



Differences between Regression Loss Functions

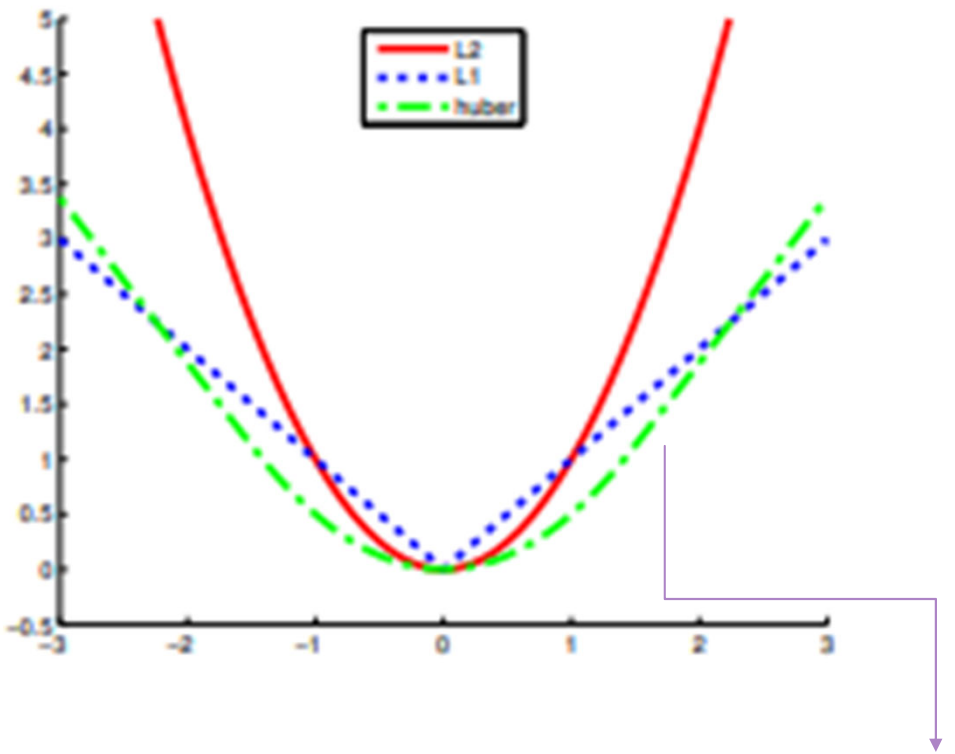
y	$\hat{y}$	$\lvert r\rvert = \lvert y-\hat{y}\rvert$	$r^2 = (y-\hat{y})^2$
1	0	1	1
5	0	5	25
10	0	10	100
50	0	50	2500



Differences between Regression Loss Functions

y	$\hat{y}$	$\lvert r\rvert = \lvert y-\hat{y}\rvert$	$r^2 = (y-\hat{y})^2$
1	0	1	1
5	0	5	25
10	0	10	100
50	0	50	2500

- How to match square loss and absolute loss?



Fix some point a

Loss Functions for Classification

- Notation

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

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

- 0-1 Loss

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

Loss Functions for Classification

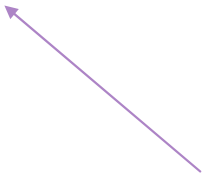
- Notation

  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

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

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.

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

Functional Margin
not Geometric Margin

- Notation

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

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

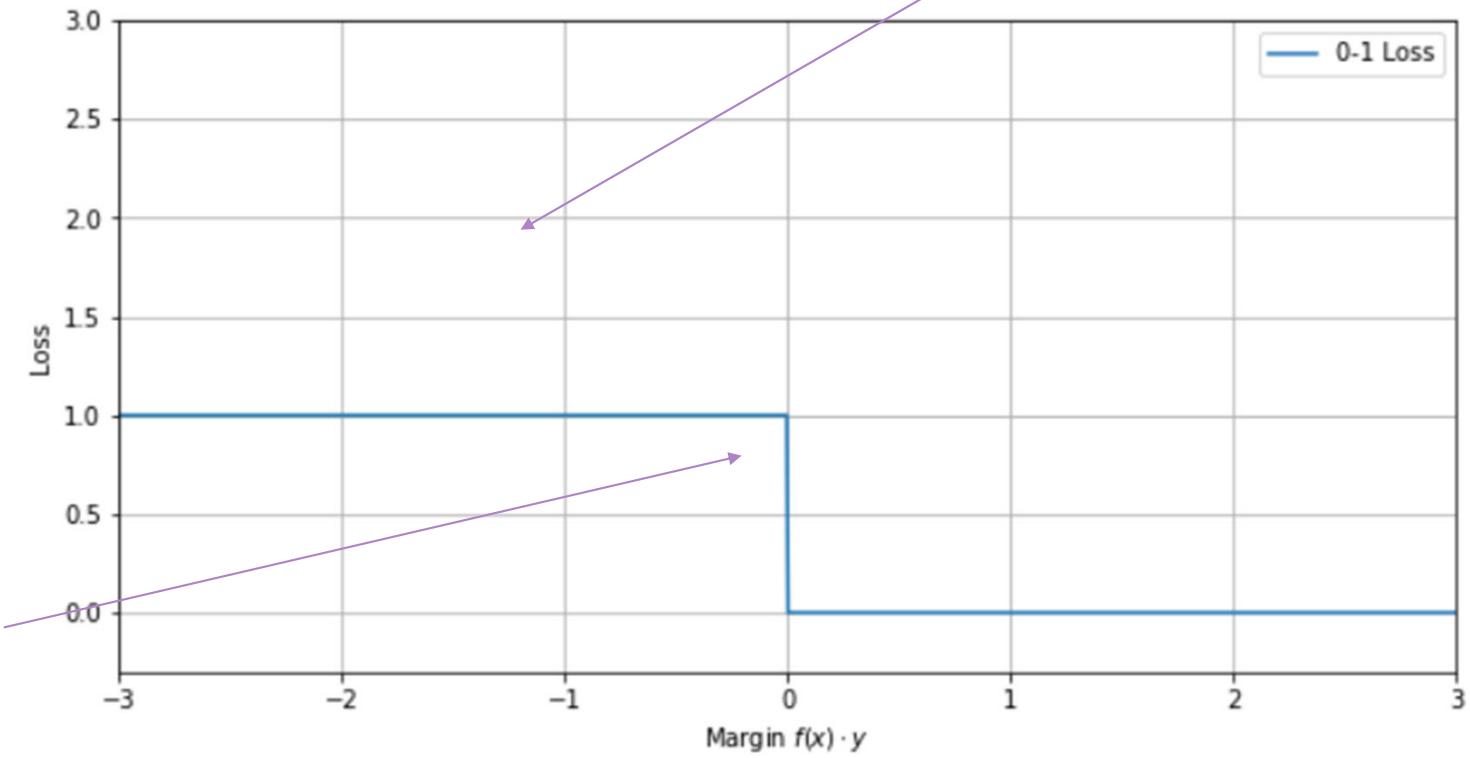
- Margin
  - For prediction f(x) and label $y \in \{-1, 1\}$

    is f(x) y
  - Same sign means positive value. Different sign means negative
  - Positive means correct. Negative means incorrect.

Loss Functions for Classification

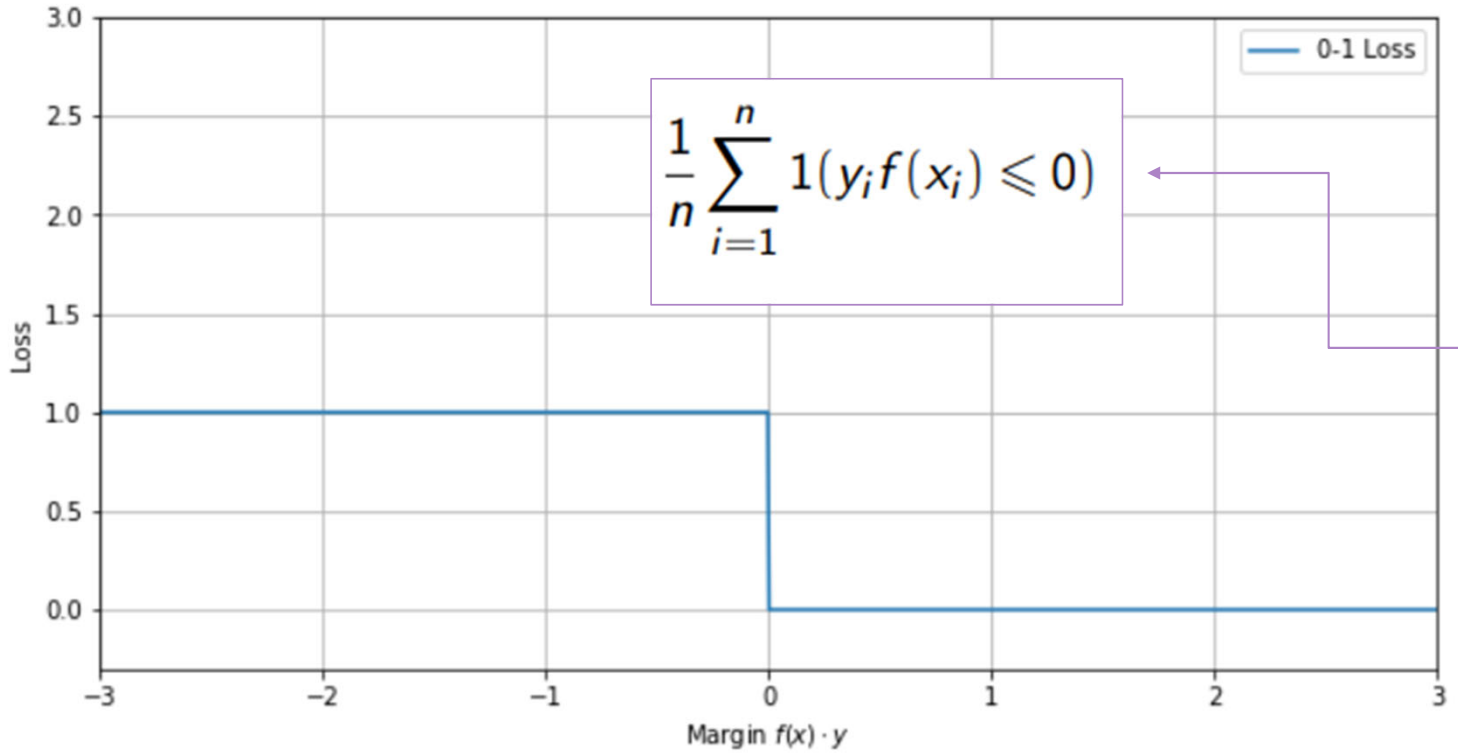
Not convex meaning no **subgradient** at decision boundary

No generalized gradient at 0



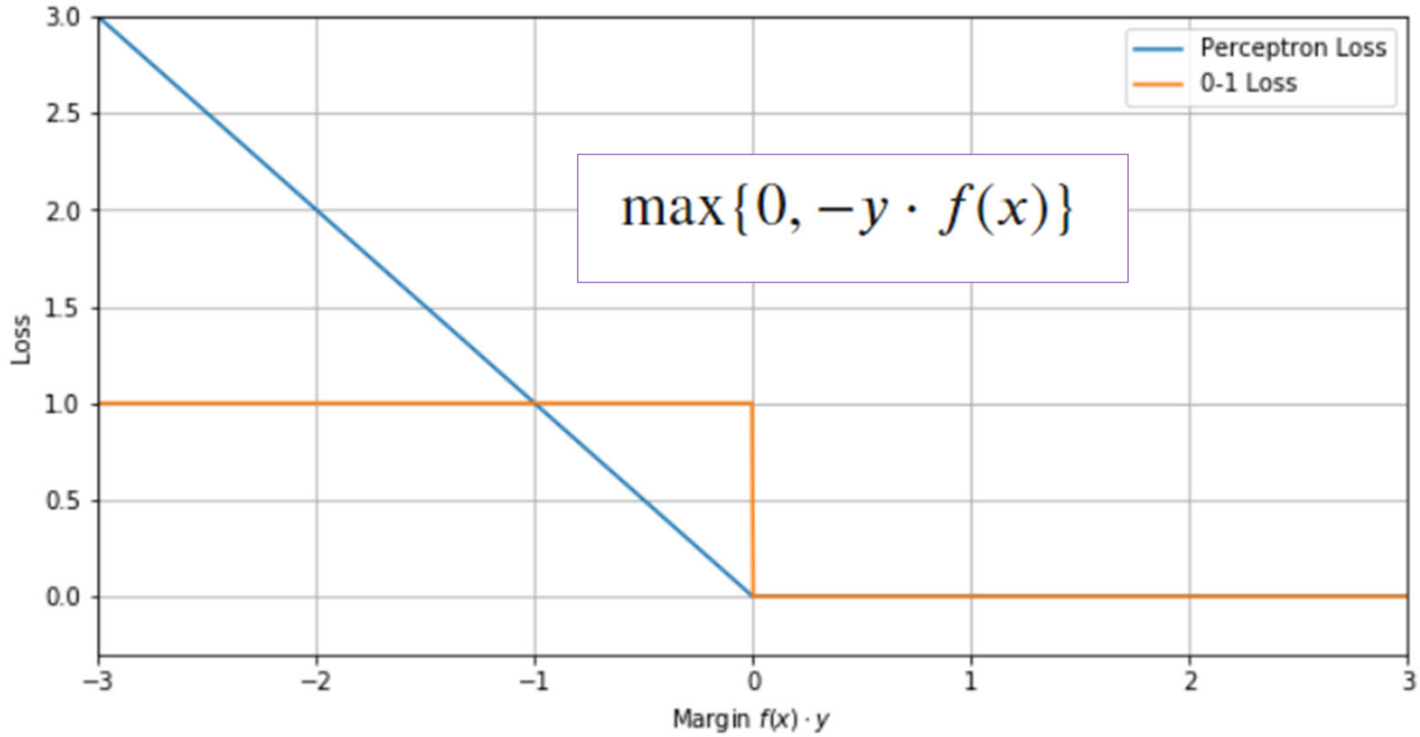
No gradient at decision boundary

Loss Functions for Classification

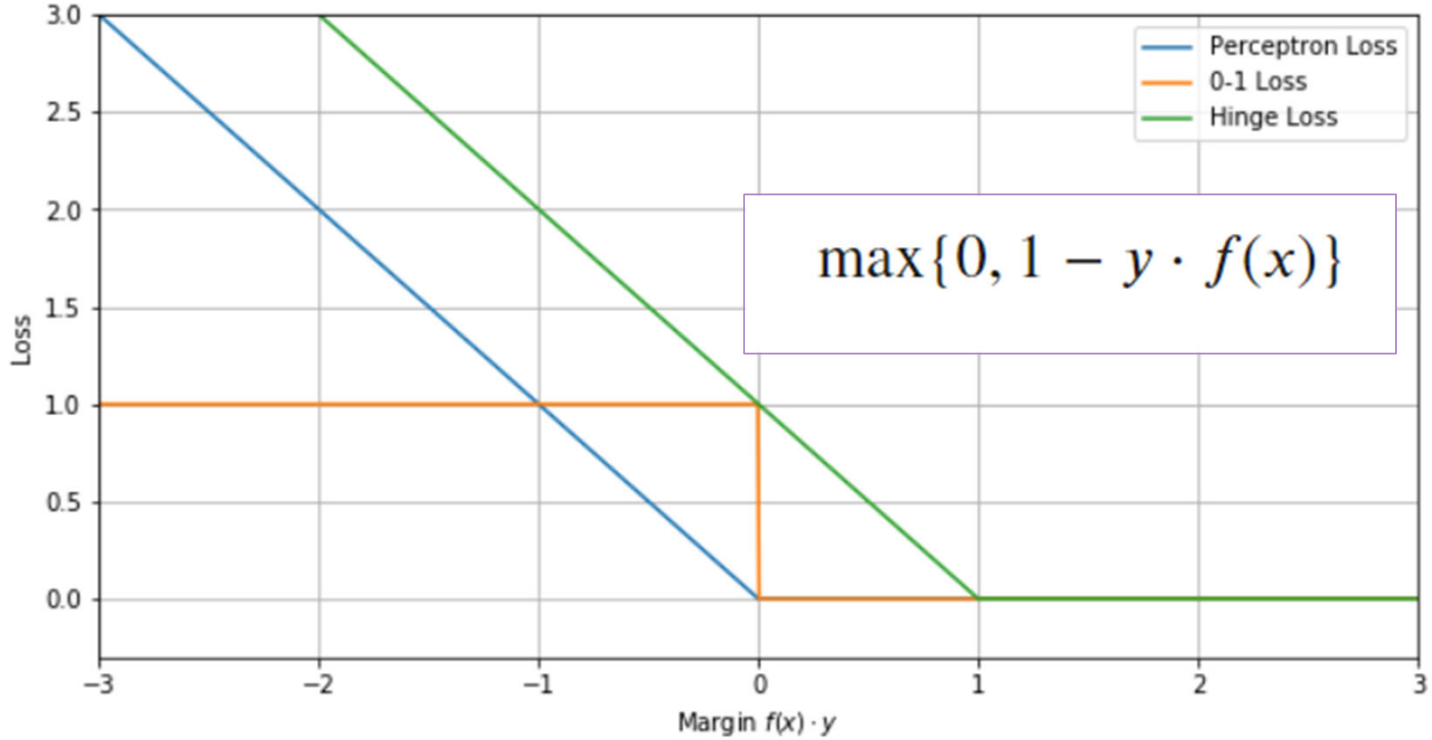


Not Computationally Tractable

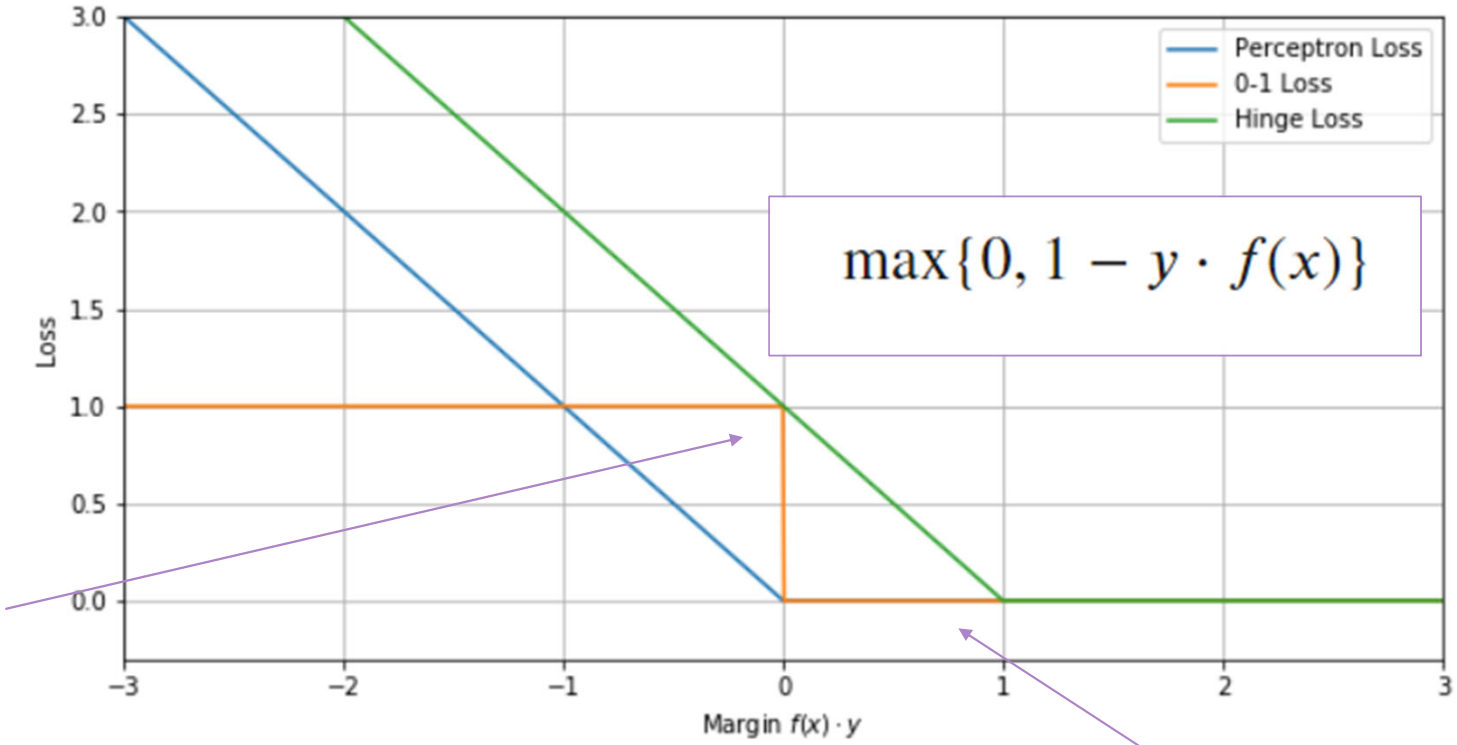
Loss Functions for Classification



Loss Functions for Classification



Loss Functions for Classification

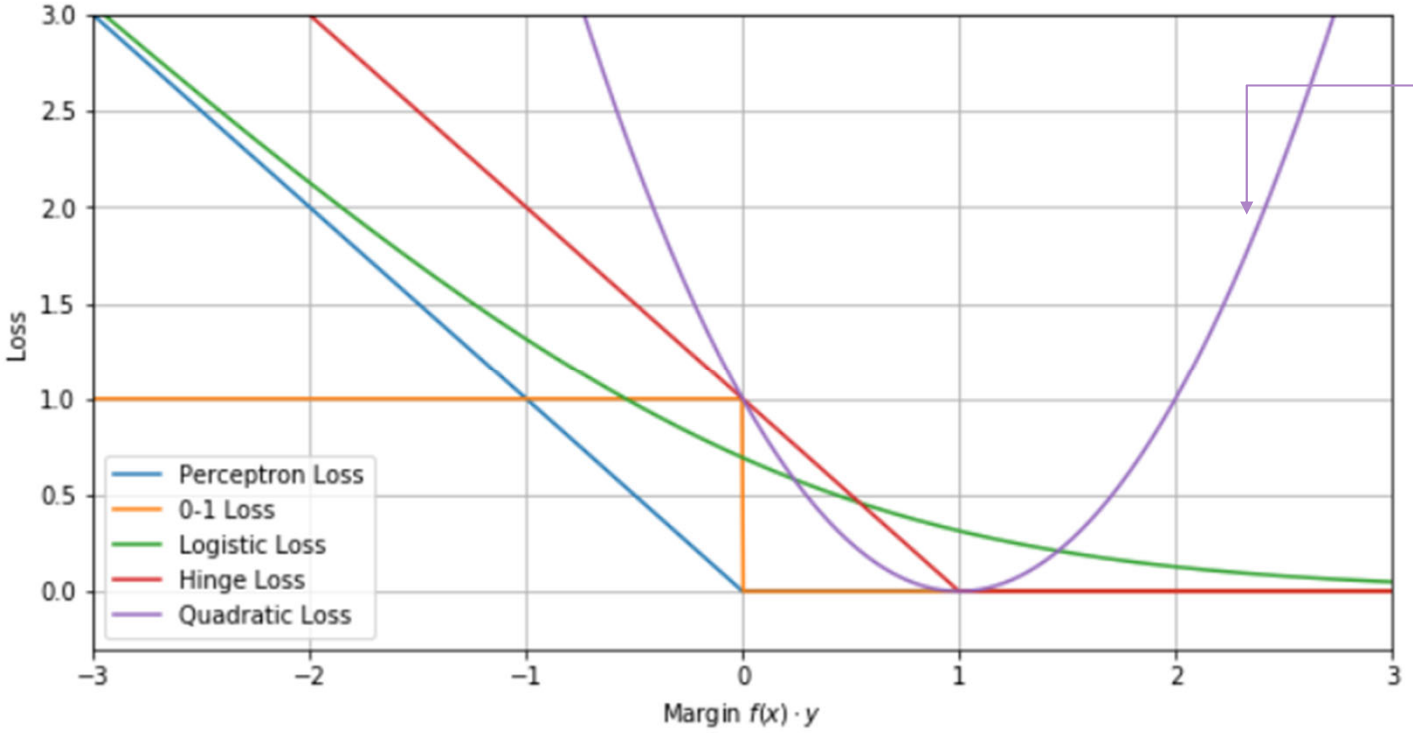


Note that hinge loss lies above 0-1 loss

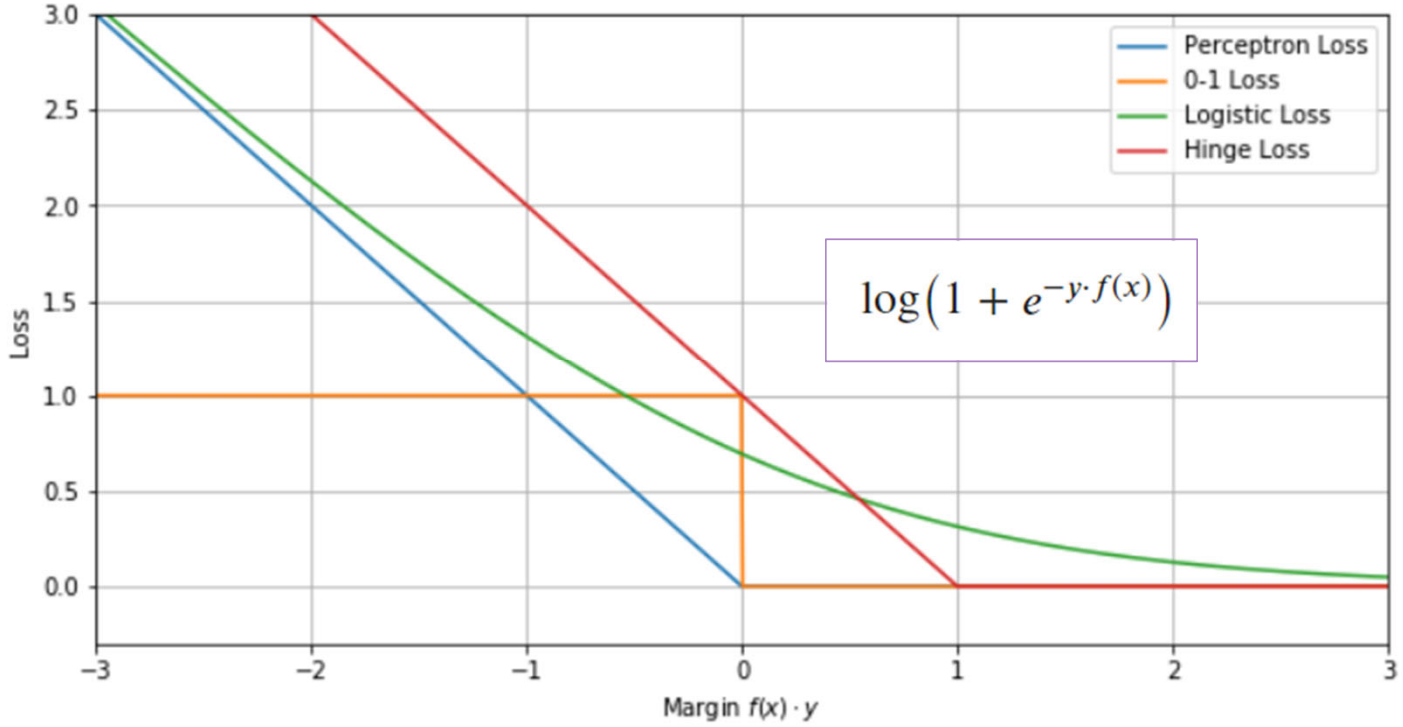
Why is hinge loss positive here?

Loss Functions for Classification

Check out ADALINE algorithm for classification that uses square loss



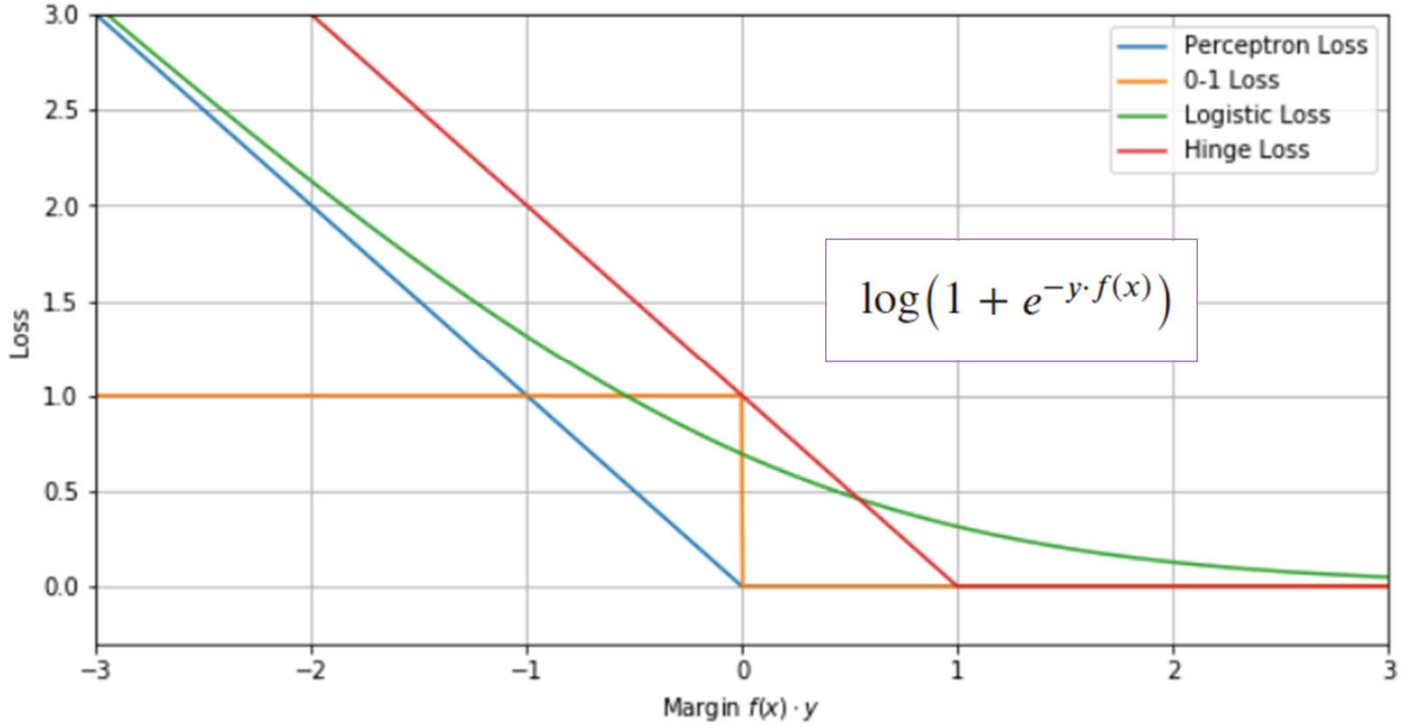
Loss Functions for Classification



Loss Functions for Classification

Question: Show that

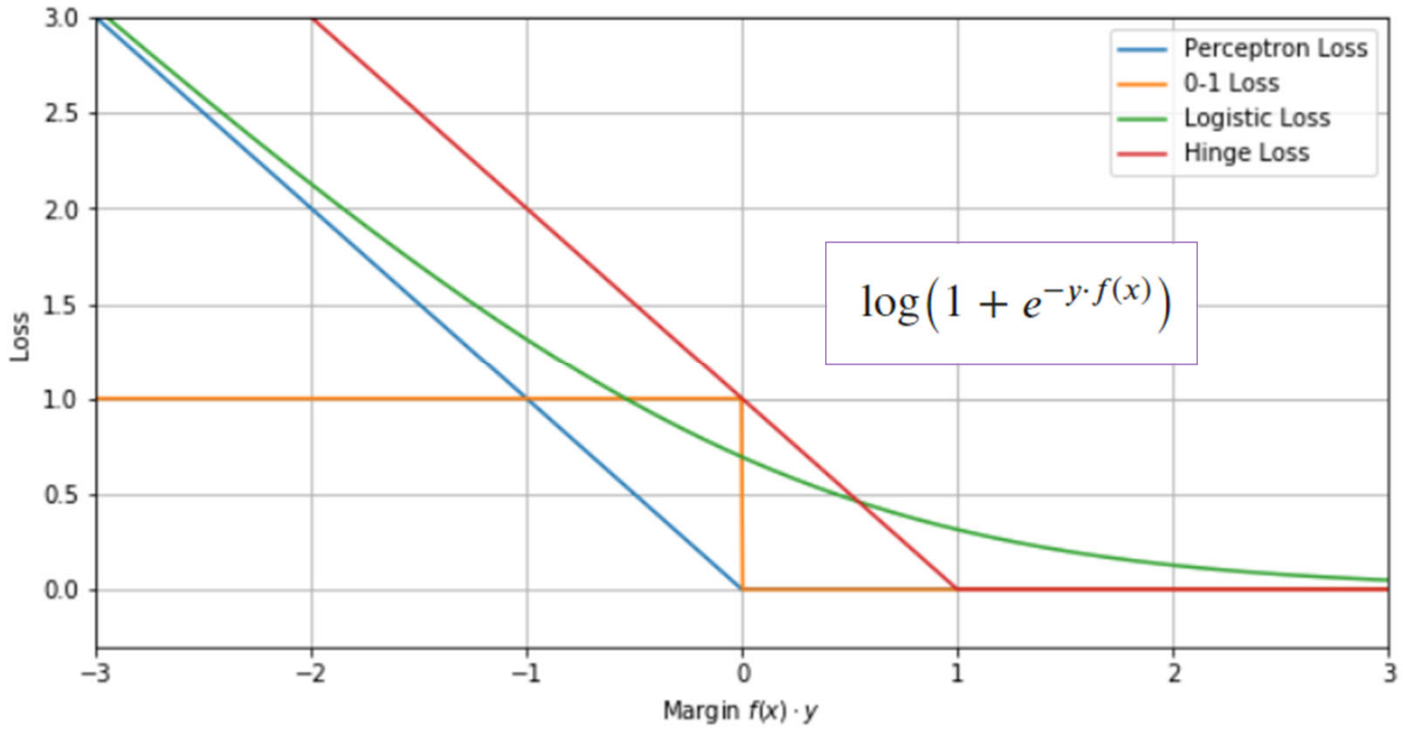
$$\max(a, b) \leq a + b \leq 2\max(a, b)$$



Loss Functions for Classification

Question: Show that

$$\max(a, b) \leq a + b \leq 2\max(a, b)$$



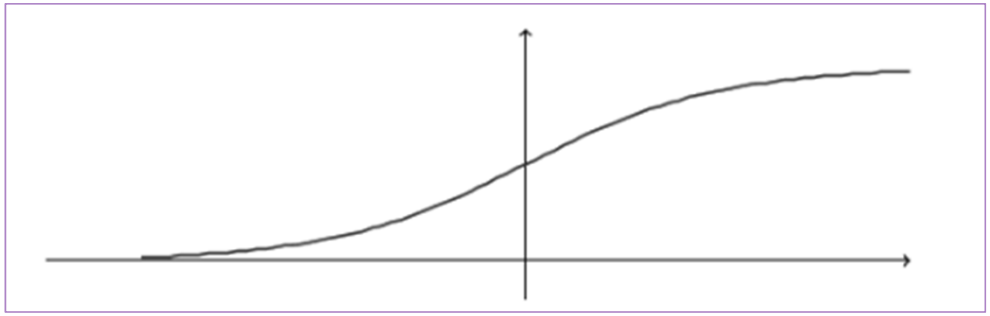
Use to show this inequality

$$\max\{x_1, \ldots, x_n\} \leq \log\left(e^{x_1} + \cdots + e^{x_n}\right) \leq \max\{x_1, \ldots, x_n\} + \log n.$$

Demo

- Logistic Regression
  - Sigmoid function

$$\phi_{\text{sig}}(z) = \frac{1}{1+\exp(-z)}.$$



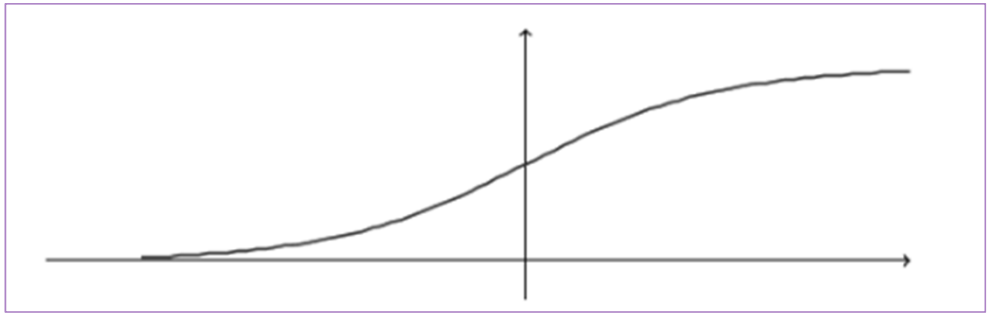
Demo

- Logistic Regression
  - Sigmoid function

$$\phi_{\text{sig}}(z) = \frac{1}{1+\exp(-z)}.$$

  - Hypotheses combine linear predictor with sigmoid

$$\{\mathbf{x} \mapsto \phi_{\text{sig}}(\langle \mathbf{w}, \mathbf{x} \rangle) : \mathbf{w} \in \mathbb{R}^d\}.$$



Demo

$$1 - \frac{1}{1+\exp(-\langle \mathbf{w}, \mathbf{x} \rangle)} = \frac{\exp(-\langle \mathbf{w}, \mathbf{x} \rangle)}{1+\exp(-\langle \mathbf{w}, \mathbf{x} \rangle)} = \frac{1}{1+\exp(\langle \mathbf{w}, \mathbf{x} \rangle)}.$$

- Logistic Regression
  - Sigmoid function

$$\phi_{\text{sig}}(z) = \frac{1}{1+\exp(-z)}.$$

  - Hypotheses combine linear predictor with sigmoid

$$\{\mathbf{x} \mapsto \phi_{\text{sig}}(\langle \mathbf{w}, \mathbf{x} \rangle) : \mathbf{w} \in \mathbb{R}^d\}.$$

  - The output interpreted as probability. Used for classification

Demo

$$1-\frac{1}{1+\exp(-\langle \mathbf{w},\mathbf{x}\rangle)}=\frac{\exp(-\langle \mathbf{w},\mathbf{x}\rangle)}{1+\exp(-\langle \mathbf{w},\mathbf{x}\rangle)}=\frac{1}{1+\exp(\langle \mathbf{w},\mathbf{x}\rangle)}.$$

- Logistic Regression
  - Sigmoid function

$$\phi_{\text{sig}}(z)=\frac{1}{1+\exp(-z)}.$$

  - Hypotheses combine linear predictor with sigmoid

$$\{\mathbf{x}\mapsto \phi_{\text{sig}}(\langle \mathbf{w},\mathbf{x}\rangle):\mathbf{w}\in\mathbb{R}^d\}.$$

  - The output interpreted as probability. Used for classification

- Choose w as maximum likelihood estimator

$$P(Y \mid X;\, w)=\prod_{i=1}^{m}\frac{1}{1+e^{-y_i(w\cdot x_i)}}$$

Demo

- Logistic Regression
  - Sigmoid function

$$\phi_{\text{sig}}(z) = \frac{1}{1+\exp(-z)}.$$

  - Hypotheses combine linear predictor with sigmoid

$$\{\mathbf{x} \mapsto \phi_{\text{sig}}(\langle \mathbf{w}, \mathbf{x} \rangle) : \mathbf{w} \in \mathbb{R}^d\}.$$

  - The output interpreted as probability. Used for classification

$$1 - \frac{1}{1+\exp(-\langle \mathbf{w}, \mathbf{x} \rangle)} = \frac{\exp(-\langle \mathbf{w}, \mathbf{x} \rangle)}{1+\exp(-\langle \mathbf{w}, \mathbf{x} \rangle)} = \frac{1}{1+\exp(\langle \mathbf{w}, \mathbf{x} \rangle)}.$$

  - Choose w as maximum likelihood estimator

$$P(Y \mid X;\, w) = \prod_{i=1}^{m} \frac{1}{1 + e^{-y_i (w \cdot x_i)}}$$

  - Equivalently minimize log loss

$$-\log(P(Y \mid X;\, w)) = \sum_{i=1}^{m} \log\left(1 + e^{-y_i (w \cdot x_i)}\right)$$

Demo

- Logistic Regression
  - Sigmoid function

$$\phi_{\text{sig}}(z) = \frac{1}{1+\exp(-z)}.$$

  - Hypotheses combine linear predictor with sigmoid

$$\{\mathbf{x} \mapsto \phi_{\text{sig}}(\langle \mathbf{w}, \mathbf{x} \rangle) : \mathbf{w} \in \mathbb{R}^d\}.$$

  - The output interpreted as probability. Used for classification

$$-\log(P(Y \mid X;\, w)) = \sum_{i=1}^{m} \log\left(1 + e^{-y_i(w \cdot x_i)}\right)$$

Take-Aways

- How is logistic loss part classification and part regression ?
- What Python packages can be used for logistic regression?

Absolute Value

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!

Absolute Value

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!

- Consider any number $a \in \mathbf{R}$.
- Let the **positive part** of a be

$$a^+ = a1(a \geqslant 0).$$

- Let the **negative part** of a be

$$a^- = -a1(a \leqslant 0).$$

- Do you see why $a^+ \geqslant 0$ and $a^- \geqslant 0$?
- How do you write a in terms of a^+ and a^-?
- How do you write $|a|$ in terms of a^+ and a^-?

Absolute Value

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 $w^- = (w_1^-, \ldots, w_d^-)$.

- Consider any number $a \in \mathbf{R}$.
- Let the **positive part** of a be

$$a^+ = a1(a \geqslant 0).$$

- Let the **negative part** of a be

$$a^- = -a1(a \leqslant 0).$$

- Do you see why $a^+ \geqslant 0$ and $a^- \geqslant 0$?
- How do you write a in terms of a^+ and a^-?
- How do you write $|a|$ in terms of a^+ and a^-?

Absolute Value

We **will show:** substituting $w = w^+ - w^-$ and $|w| = w^+ + w^-$ gives an **equivalent** problem:

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

- Objective is **differentiable** (in fact, **convex and quadratic**)
- $2d$ variables vs d variables and $2d$ constraints vs no constraints

Absolute Value?

Lasso problem is trivially equivalent to the following:

$$\begin{aligned} \min_{w} \min_{a,b} \quad & \sum_{i=1}^{n} \left((a-b)^T x_i - y_i \right)^2 + \lambda 1^T (a+b) \\ \text{subject to} \quad & a_i \geqslant 0 \text{ for all } i \qquad b_i \geqslant 0 \text{ for all } i, \\ & a - b = w \\ & a + b = |w| \end{aligned}$$

- Claim: Don't need constraint $a+b=|w|$.
- $a' \leftarrow a - \min(a,b)$ and $b' \leftarrow b - \min(a,b)$ at least as good
- So if a and b are minimizers, at least one is 0.
- Since $a-b=w$, we must have $a=w^+$ and $b=w^-$. So also $a+b=|w|$.

Absolute Value

- So lasso optimization problem is equivalent to

$$\begin{aligned} \min_{a,b} \quad & \sum_{i=1}^{n} \left((a-b)^T x_i - y_i \right)^2 + \lambda 1^T (a+b) \\ \text{subject to} \quad & a_i \geqslant 0 \text{ for all } i \qquad b_i \geqslant 0 \text{ for all } i, \end{aligned}$$

  where at the end we take $w^* = a^* - b^*$ (and we've shown above that a^* and b^* are positive and negative parts of w^*, respectively.)

- Has constraints – how do we optimize?

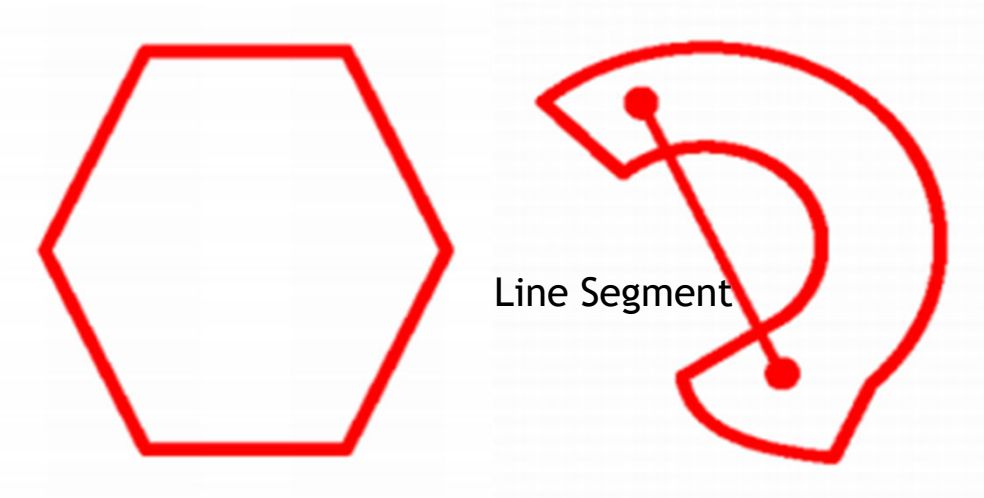
Projected Gradient Descent

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

- Just like SGD, but after each step
  - Project w^+ and w^- into the constraint set.
  - In other words, if any component of w^+ or w^- becomes negative, set it back to 0.

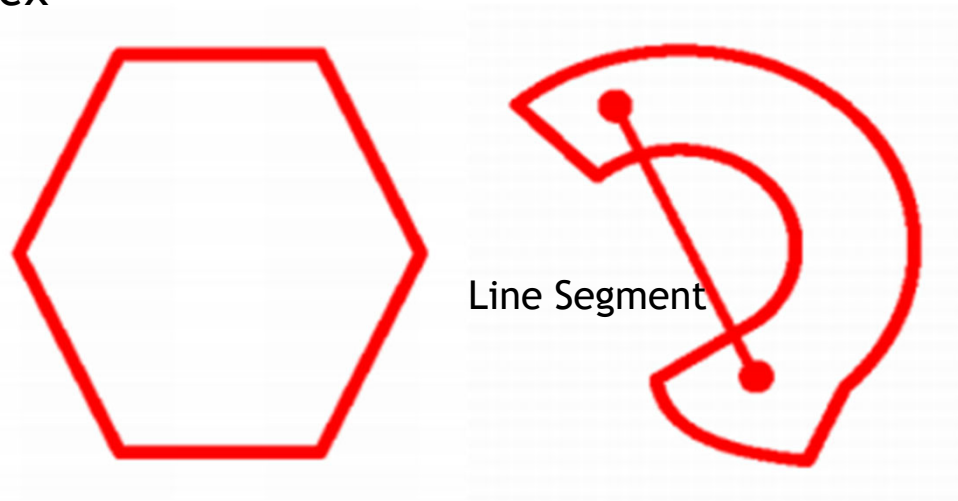
Convexity

Convex
Sets

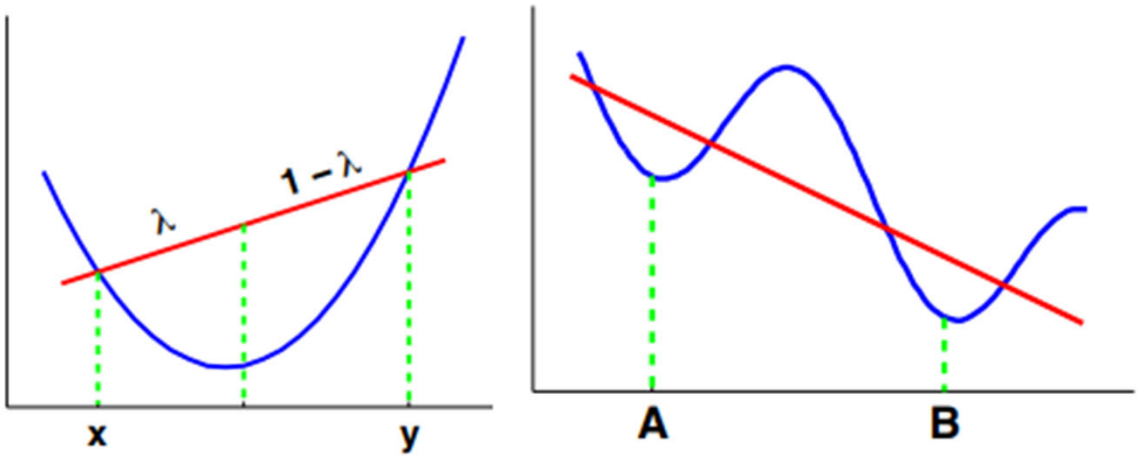


Convexity

Convex Sets



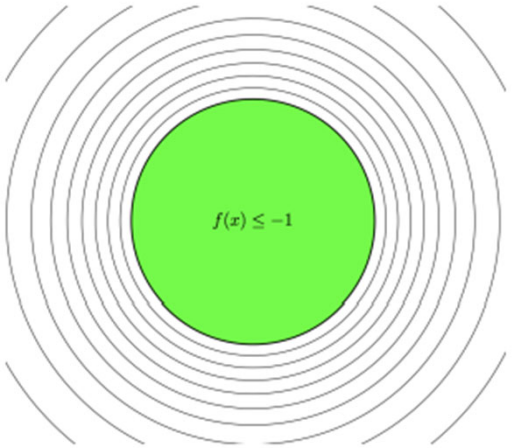
Convex Function



What is a concave function?

Can a function be both convex and concave?

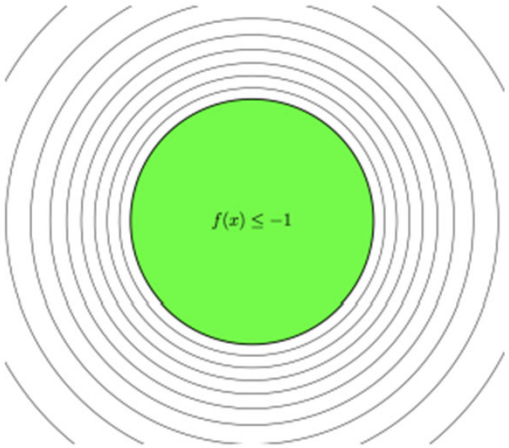
Convexity



$f : \mathbf{R}^d \to \mathbf{R}$ 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

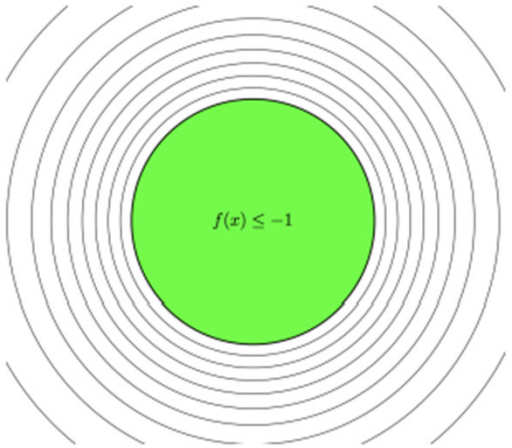


$f : \mathbf{R}^d \to \mathbf{R}$ be a function.

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

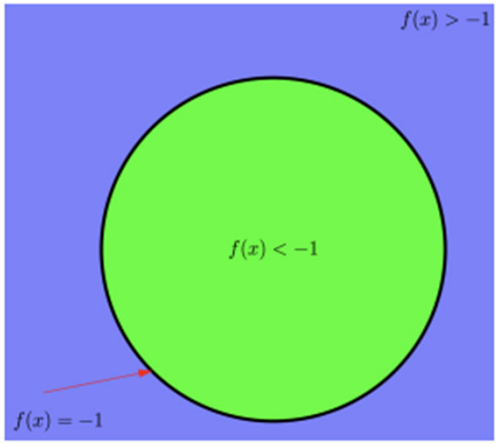


$f : \mathbf{R}^d \to \mathbf{R}$ be a function.

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 $f(x) \leqslant c$.

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 $f(x) = c$.

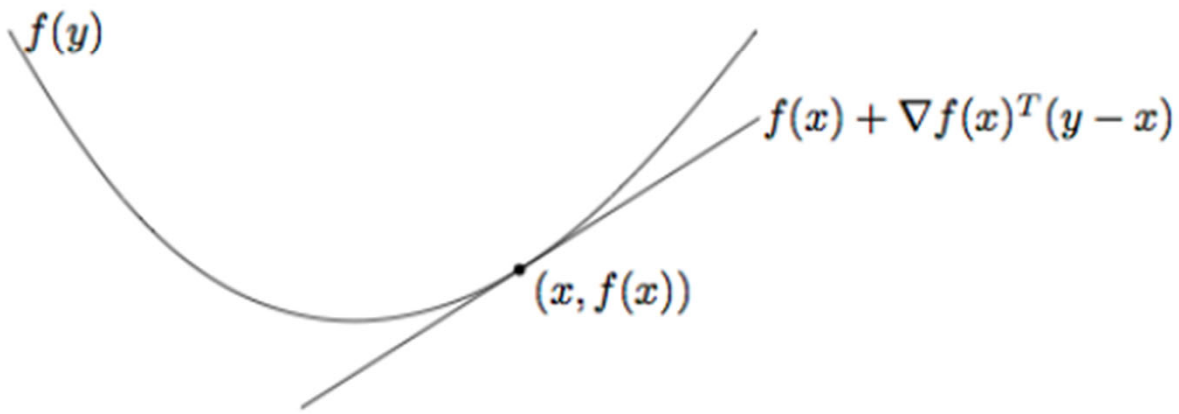
A **sublevel** set for the value c is the set of points $x \in \mathbf{R}^d$ for which $f(x) \leqslant c$.

*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:

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



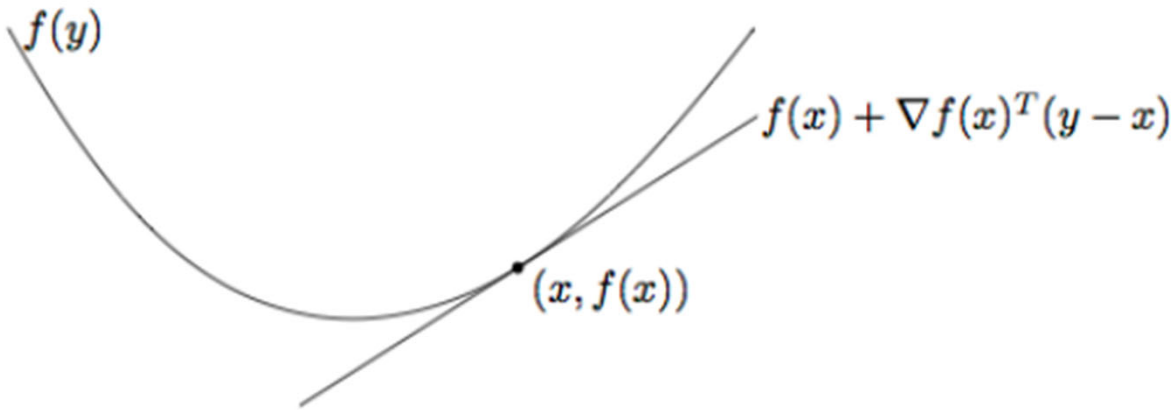
Convexity

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:



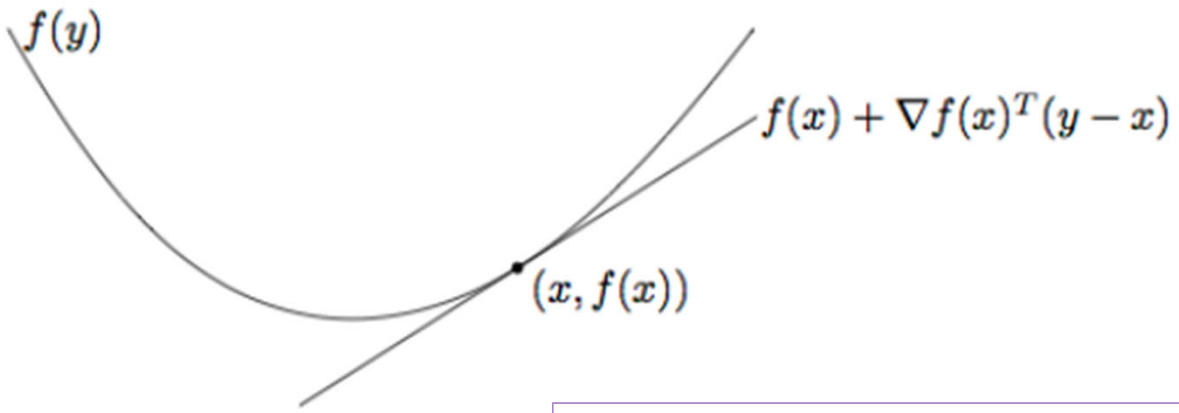
Convexity

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:

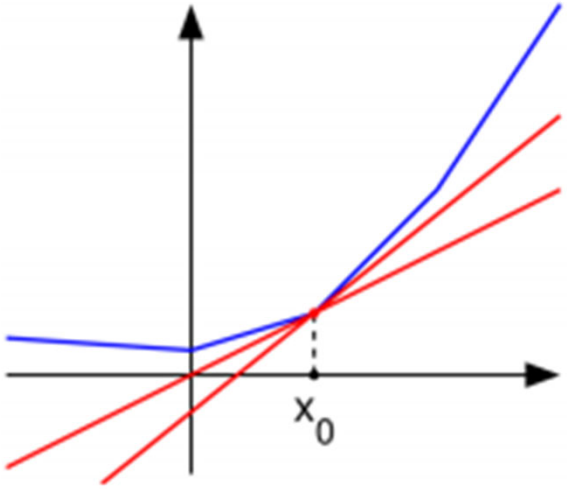


If $\nabla f(x) = 0$ then x is a global minimizer of f.

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,

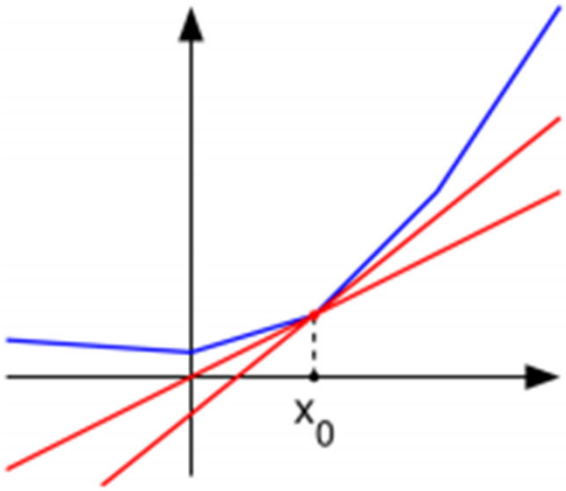
$$f(z) \geqslant f(x) + g^T(z - x).$$



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,

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

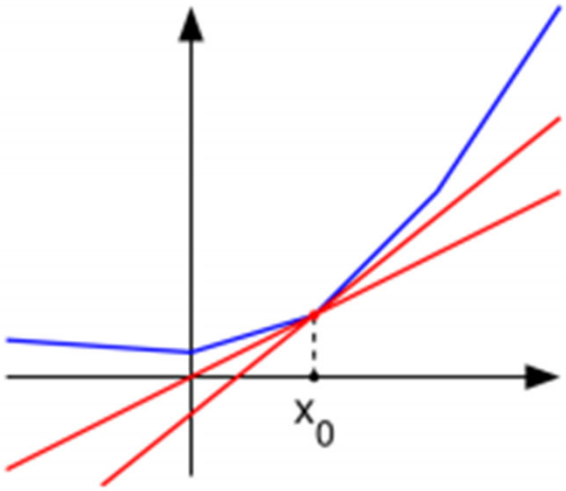


f is **subdifferentiable** at x if $\exists$ at least one subgradient at x.
The set of all subgradients at x is called the **subdifferential**: $\partial f(x)$

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,

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



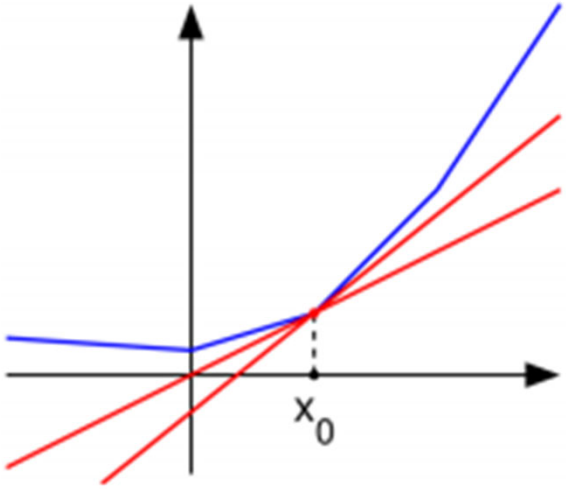
f is **subdifferentiable** at x if $\exists$ at least one subgradient at x.
The set of all subgradients at x is called the **subdifferential**: $\partial f(x)$

- 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,

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



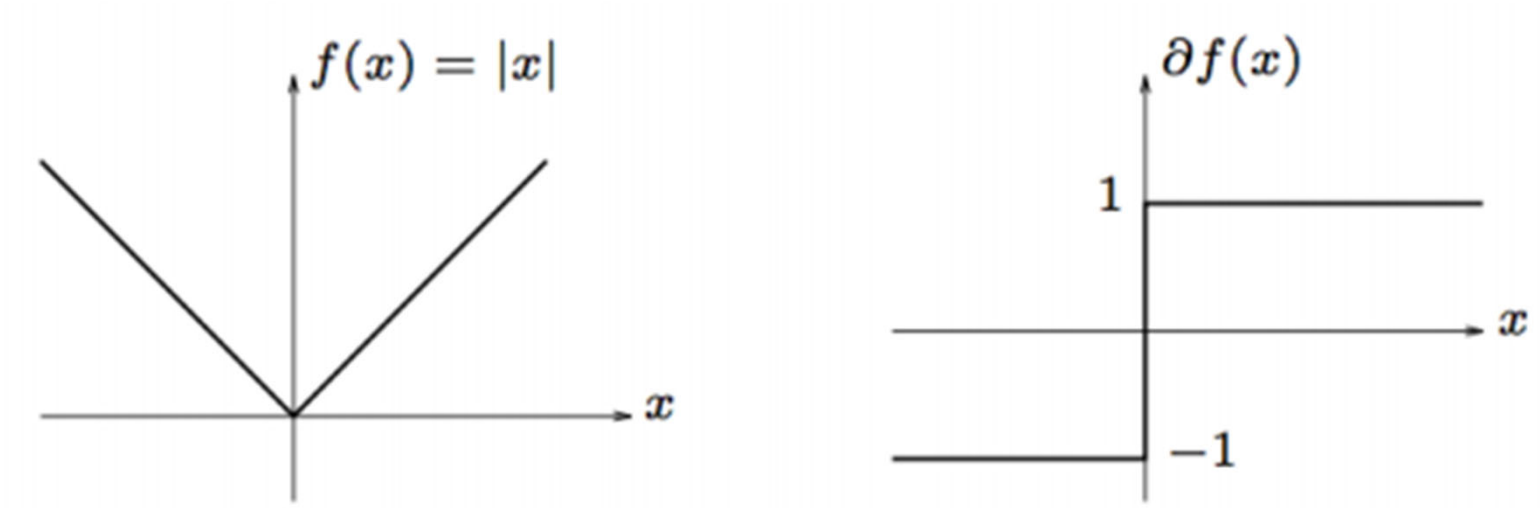
f is **subdifferentiable** at x if $\exists$ at least one subgradient at x.
The set of all subgradients at x is called the **subdifferential**: $\partial f(x)$

- 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.

What if

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

Subgradients



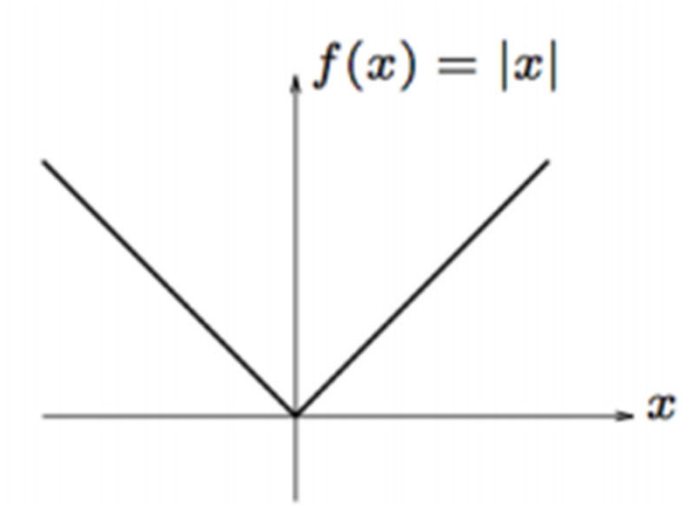
Subgradients

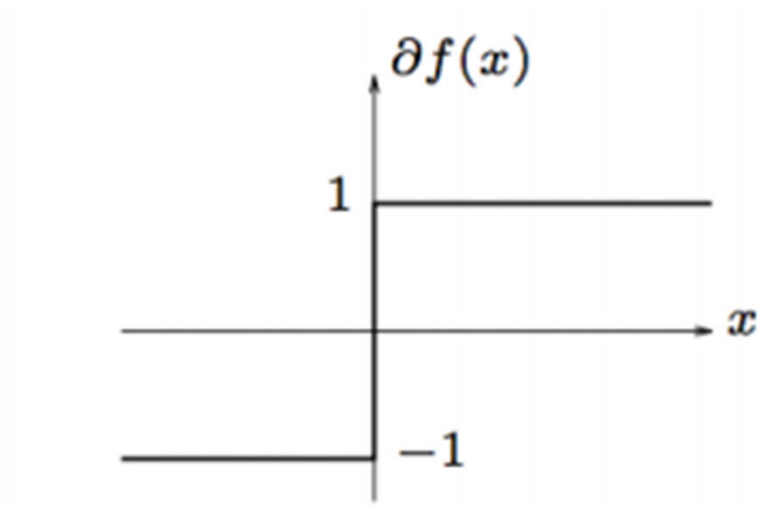
Let $\mathcal{X} = \{1, \ldots, 10\}$, let $\mathcal{Y} = \{1, \ldots, 10\}$, and let $A = \mathcal{Y}$. Suppose the data generating distribution, P, has marginal $X \sim \text{Unif}\{1, \ldots, 10\}$ and conditional distribution $Y|X = x \sim \text{Unif}\{1, \ldots, x\}$. For each loss function below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

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

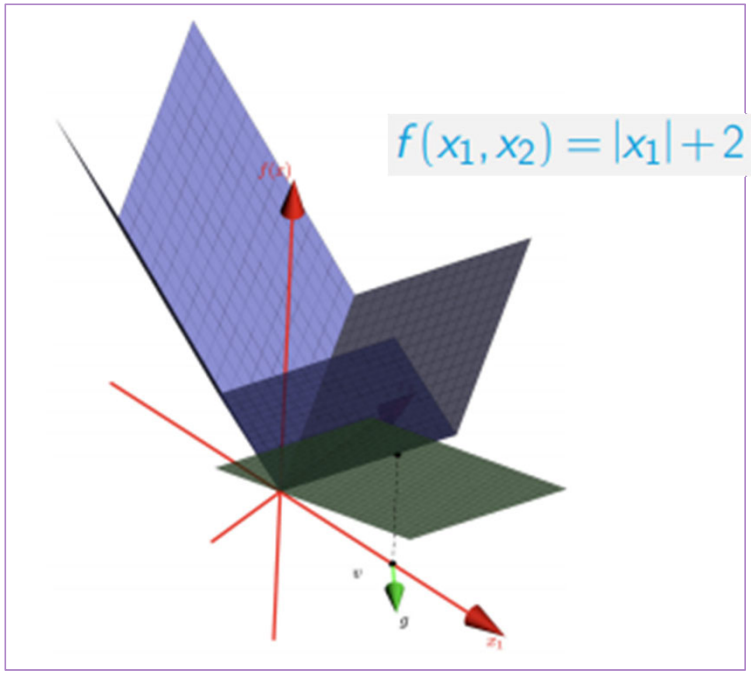
(b) $\ell(a, y) = |a - y|$,

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



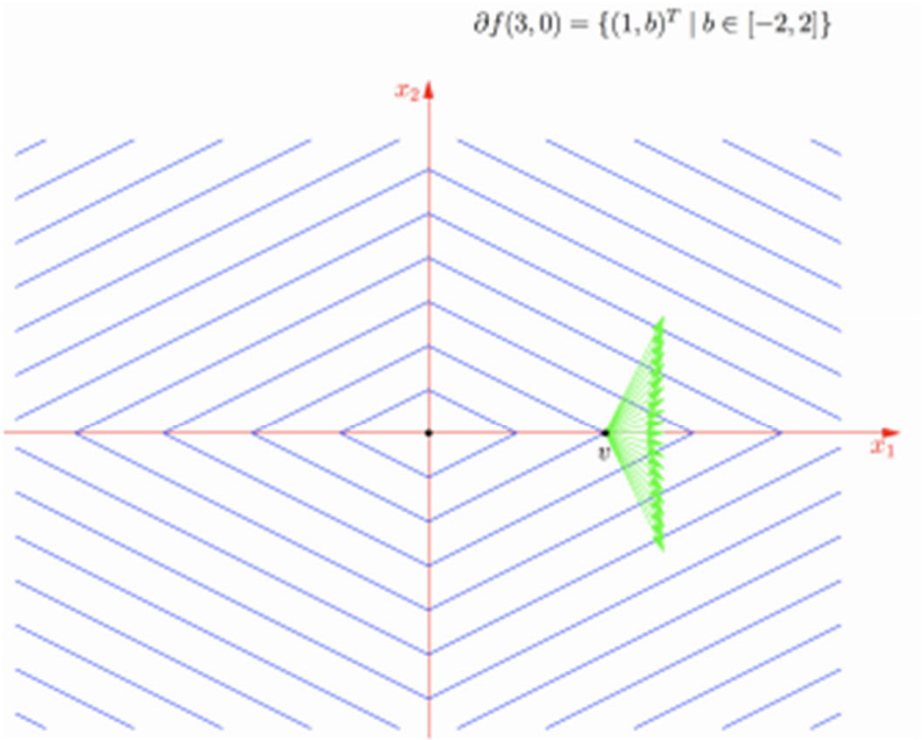


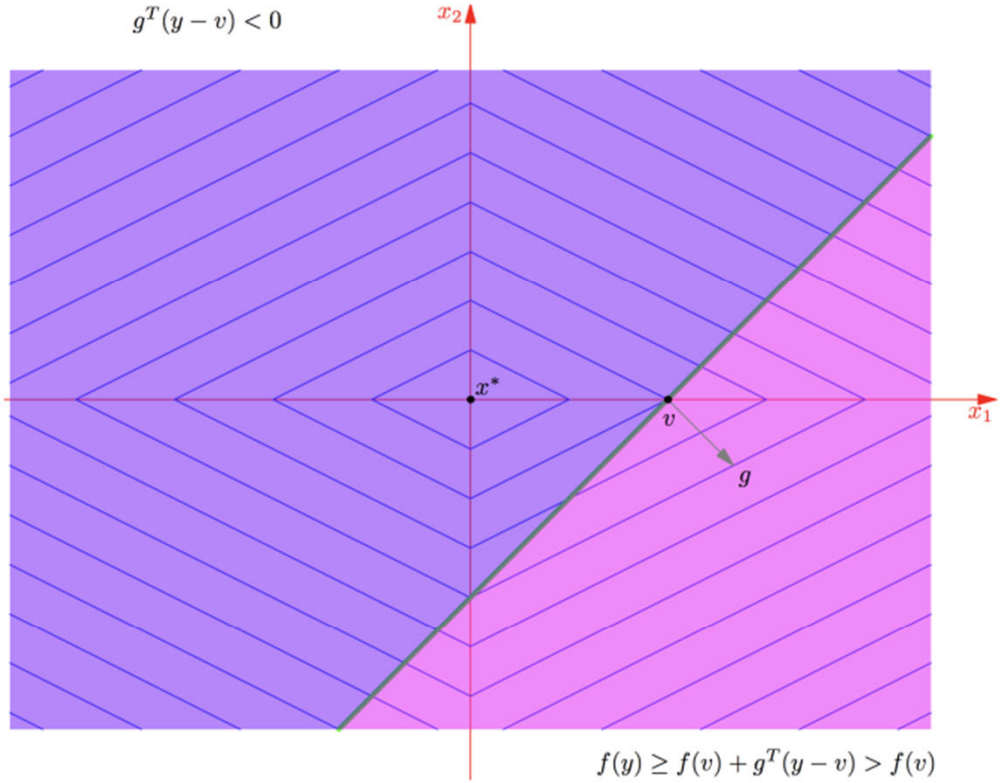
Subgradients



Subgradients

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





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 $t > 0$,*

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

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 $f(z) < f(x_0)$.

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

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

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

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

- 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

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

Summary

- Loss Functions
  - 0-1, Perceptron, Hinge, Log
  - Distance Based and Margin Based
- Absolute Value
  - Double the number of variables
- Subgradients
  - Any vector with properties of gradient
  - Subgradient descent