CSC 580 Principles of Machine Learning

06 Linear models and convexity

Chicheng Zhang

Department of Computer Science

The University of Arizona

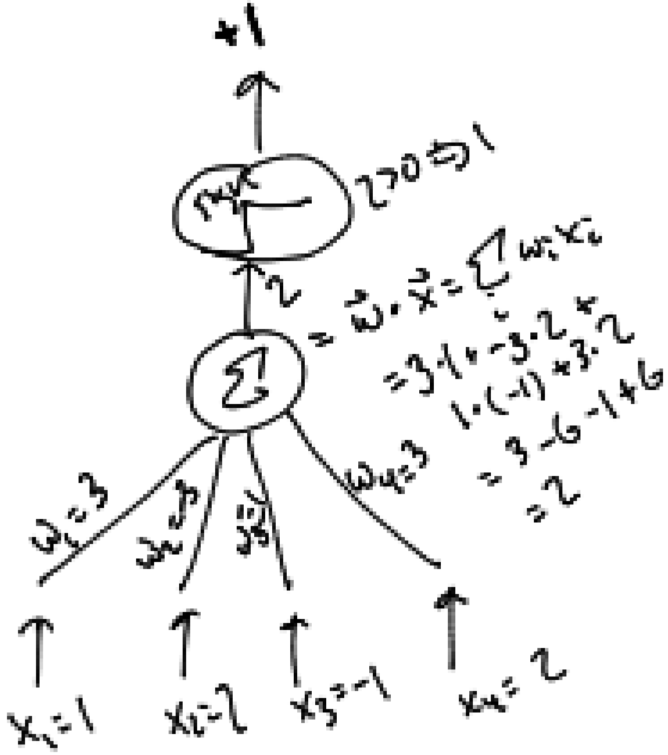
*slides credit: built upon CSC 580 Fall 2021 lecture slides by Kwang-Sung Jun

Overview

- Linearity – recall perceptron
  - $h(x) = \text{sign}(\langle w, x \rangle + b)$ – classification
  - $h(x) = \langle w, x \rangle + b$ - regression

- Why linear?
  - Simplicity
  - Interpretability
  - Computational efficiency

- First, linear regression (this lecture)

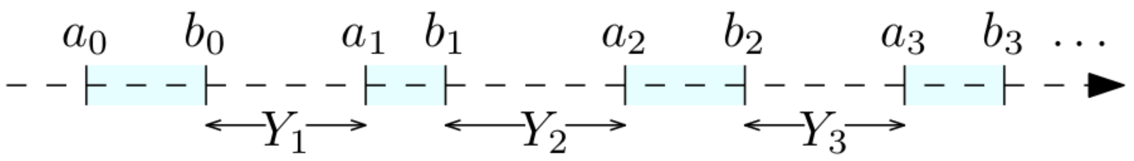


Regression example

Figure 2: Old Faithful geyser in Yellowstone National Park

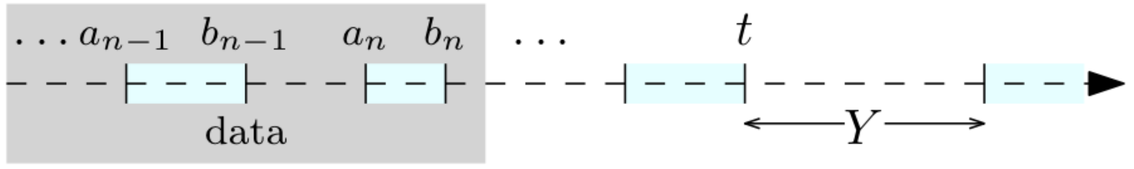
Eruption prediction

- Example: When will "Old Faithful" geyser erupt?
- Predict "time between eruptions"
- Old Faithful Geyser Data



$h(x) = b$ (no feature)

- Mean on past 136 observations: $\hat{\mu} = 70.7941$ minutes
  - So predict $\hat{y} = \hat{\mu} = 70.7941$



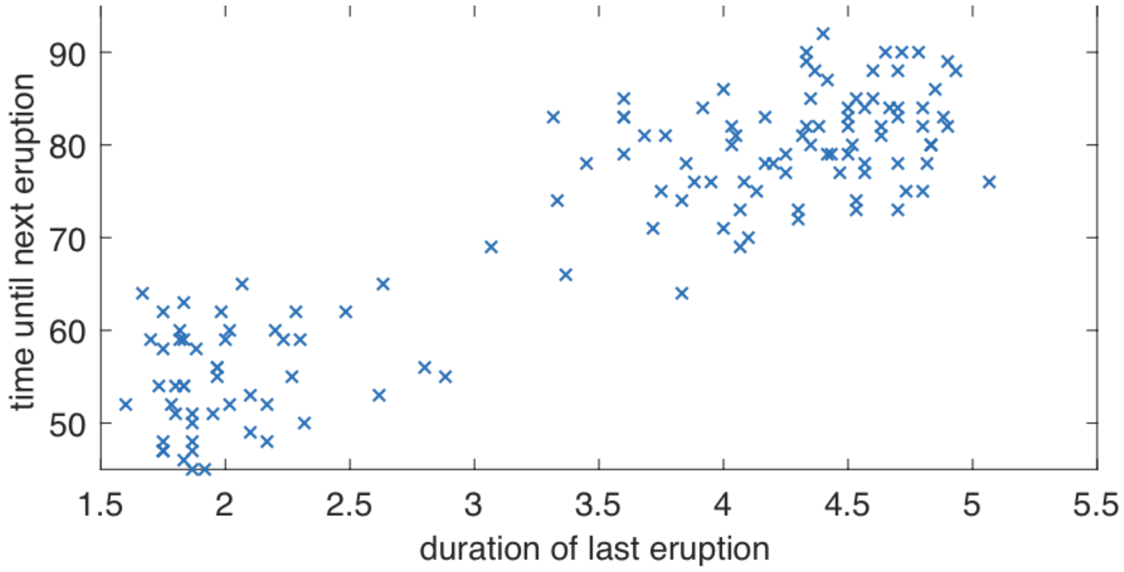
- Mean squared error on next 136 observations: 187.1894
  - Square root: 13.6817 minutes

mean squared error:

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$$

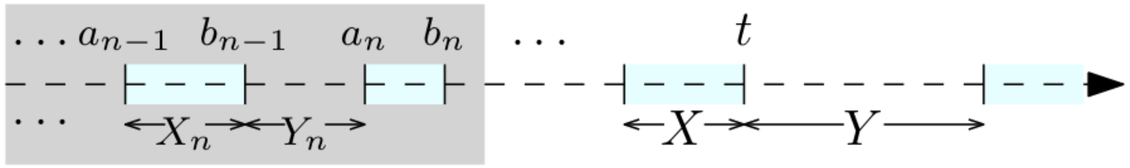
Eruption prediction

- Henry Woodward observed that "time between eruptions" seems related to "duration of latest eruption"

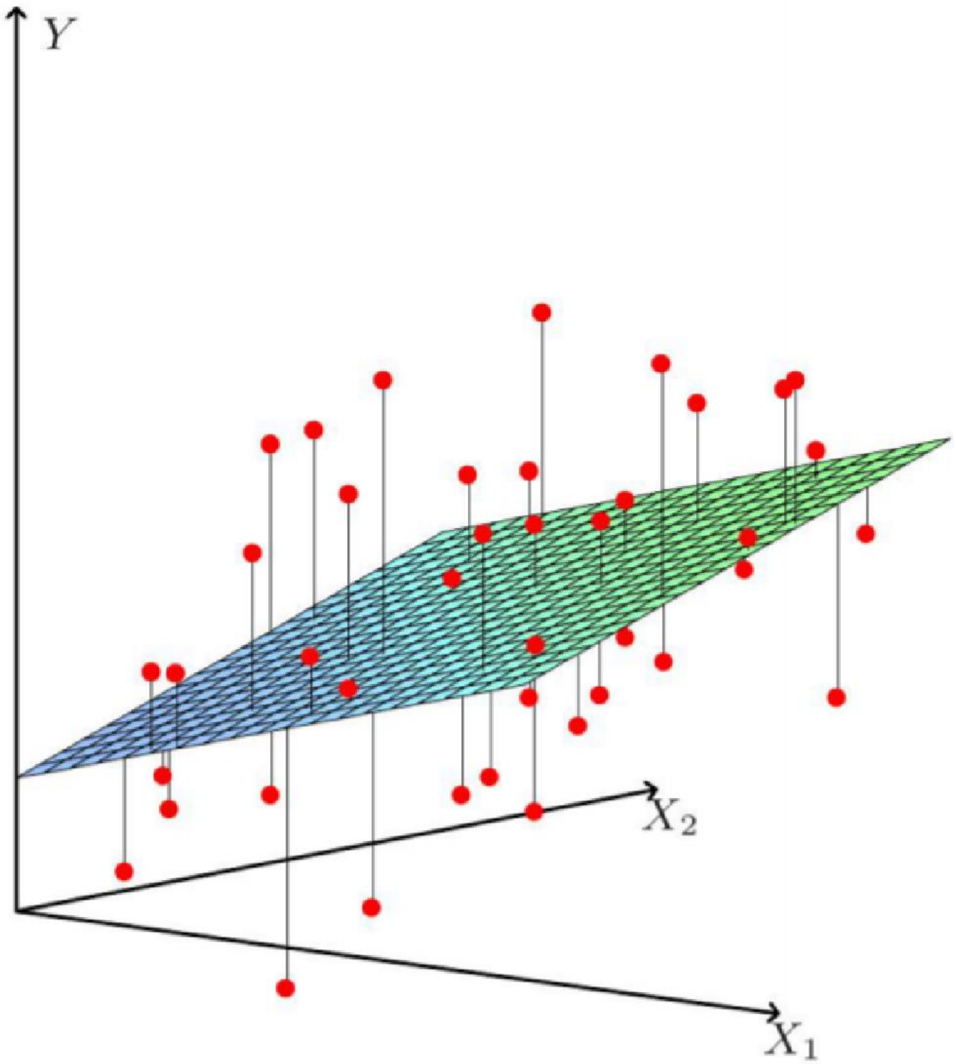


$$h(x) = w \cdot x + b$$

- Use "duration of latest eruption" as feature x
- Can use x to predict time until next eruption, y



Linear regression in dimension >= 2

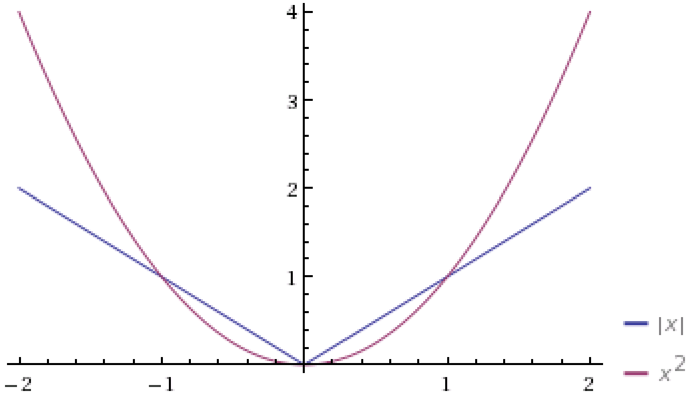


$$h(x) = w_1 \cdot x_1 + w_2 \cdot x_2 + b = \langle w, x \rangle + b$$

Formal intro to regression

- Recall classification: Y = 0 or 1; use 0/1 loss $\ell(y, \hat{y}) = I(y \neq \hat{y})$

- Regression: $Y \in \mathbb{R}$; which loss?
  - Square loss $\ell(y, \hat{y}) = (y - \hat{y})^2$
  - Absolute loss $\ell(y, \hat{y}) = |y - \hat{y}|$



- Terminology
  - expected loss (= risk) $R_D(h) = \mathbb{E}_D\left[\left(y - h(x)\right)^2\right]$ (cf. true error rate)
  - empirical loss (= emp. risk) $\hat{R}_n(h) = \mathbb{E}_S\left[\left(y - h(x)\right)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - h(x_i)\right)^2$ (cf. training error rate)
  - regression function $h^*(x) = \operatorname{argmin}_{\hat{y}} \mathbb{E}[(Y - \hat{y})^2 \mid X = x]$ (cf. Bayes classifier)
  - Bayes risk $R_D(h^*)$ (cf. Bayes error)

Linear regression

- The linear class of functions

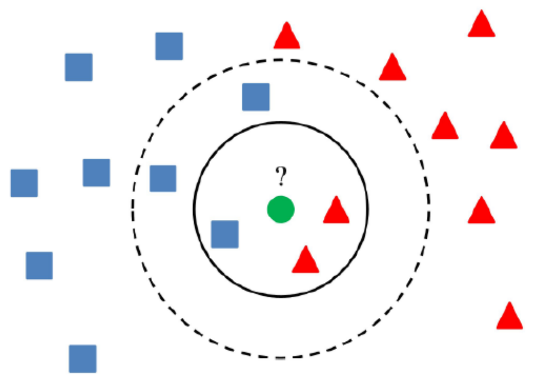
$\mathcal{H} = \{h: h(x) = \langle w, x\rangle + b,\ \text{for some } w \in \mathbb{R}^d, b \in \mathbb{R}\}$ (nonhomogeneous linear class)

$\mathcal{H} = \{h: h(x) = \langle w, x\rangle,\ \text{for some } w \in \mathbb{R}^d\}$ (homogeneous linear class)

- *Parametric* model class

- Cf. nonparametric models
  - it does not mean 'no parameters'
  - it means the number of parameters are not fixed before training

  - examples: decision trees, k-NN



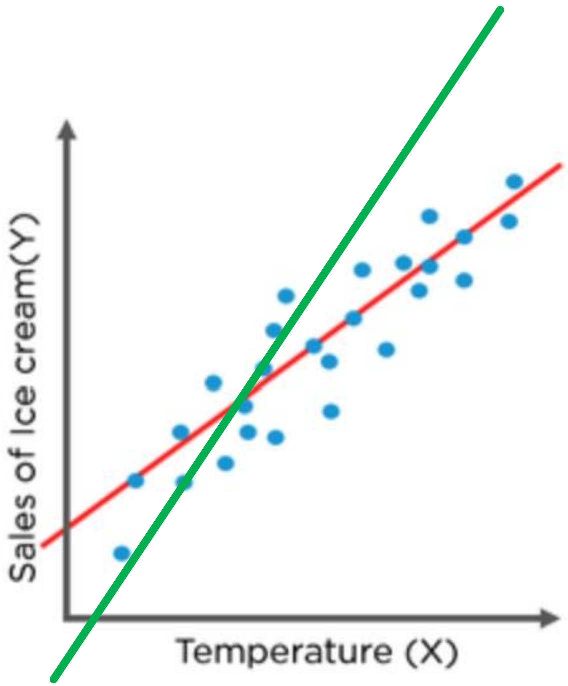
Training linear regression models

- The Empirical Risk Minimization (ERM) principle:

- The train data $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$

- $\hat{w} = \arg\min_{w \in \mathbb{R}^d} \left[\hat{R}_n(h_w) := \frac{1}{n} \sum_{i=1}^{n} (w^\top x_i - y_i)^2 \right]$
- An optimization problem

Objective function



- How to solve it?

Solving the optimization problem

$$\hat{w} = \arg\min_{w \in \mathbb{R}^d} \left[F(w) \coloneqq \sum_{i=1}^{n} (w^\top x_i - y_i)^2 \right]$$

- Optimality condition: $\hat{w}$ needs to satisfy $\nabla F(\hat{w}) = 0$,

  where $\nabla F(w) \coloneqq \left(\nabla_1 F(w), \ldots, \nabla_d F(w)\right) = \left(\frac{\partial F}{\partial w_1}, \frac{\partial F}{\partial w_2}, \ldots, \frac{\partial F}{\partial w_d}\right)$

$$w \longrightarrow (w^\top x - y) \longrightarrow (w^\top x - y)^2$$

- $\nabla_j (w^\top x - y)^2 = \frac{\partial (w^\top x - y)^2}{\partial w_j} = 2(w^\top x - y) \cdot \frac{\partial (w^\top x - y)}{\partial w_j} = 2(w^\top x - y) x_j \Longrightarrow \nabla (w^\top x - y)^2 = 2(w^\top x - y) x$
- $\nabla F(w) = \sum_{i=1}^{n} 2(w^\top x_i - y_i) x_i = 0$

  $\Longrightarrow \sum_{i=1}^{n} x_i x_i^\top w = \sum_{i=1}^{n} y_i x_i$

  $\Longrightarrow w = V^{-1} c$ where $c = \sum_{i=1}^{n} y_i x_i$, $V = \sum_{i=1}^{n} x_i x_i^\top$
- One issue? When does that happen?

Same derivation with matrix notations

$$\hat{w} = \arg\min_{w \in \mathbb{R}^d} F(w) \coloneqq \|Xw - y\|_2^2$$

- $F(w) = f\big(g(w)\big)$, where $g(w) = Xw - y$, $f(v) = \|v\|_2^2$
- Chain rule of differentiation:

$$w \xrightarrow{g} v \xrightarrow{f} F$$

$$\frac{\partial F}{\partial w} = \frac{\partial F}{\partial v} \cdot \frac{\partial v}{\partial w}, \text{ where } \frac{\partial u}{\partial z} = \frac{\partial}{\partial z}\begin{pmatrix} u_1 \\ \cdots \\ u_n \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial z_1}u_1 & \cdots & \frac{\partial}{\partial z_m}u_1 \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial z_1}u_n & \cdots & \frac{\partial}{\partial z_m}u_n \end{pmatrix}$$ is the Jacobian of u wrt z

- $\frac{\partial F}{\partial v} = 2v, \frac{\partial v}{\partial w} = X$

- $\nabla F(w)^\top = \frac{\partial F}{\partial w} = 2v \cdot X = 2(Xw - y)^\top X = 2(w^\top V - c^\top) = 2(V w - c)^\top$

The issue of inversion

- The inverse may not exist! when does it happen?
  - The instances $\{x_1, \ldots, x_n\}$ do not span the full $\mathbb{R}^d$ space
  - Guaranteed to happen if $n < d$

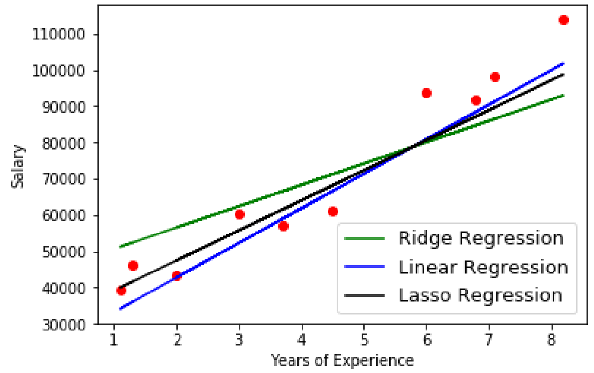
(Zico Kolter's linear algebra review p12; link in lec00 slides)

- In this case, turns out there are infinitely many w's that satisfies $X^\top X w = X^\top y$ (thus an optimal solution)
  - Among those w's, **<u>the one with the minimum norm</u>** can be found by replacing the inverse with *Penrose-Moore pseudo inverse* (function pinv() in numpy):

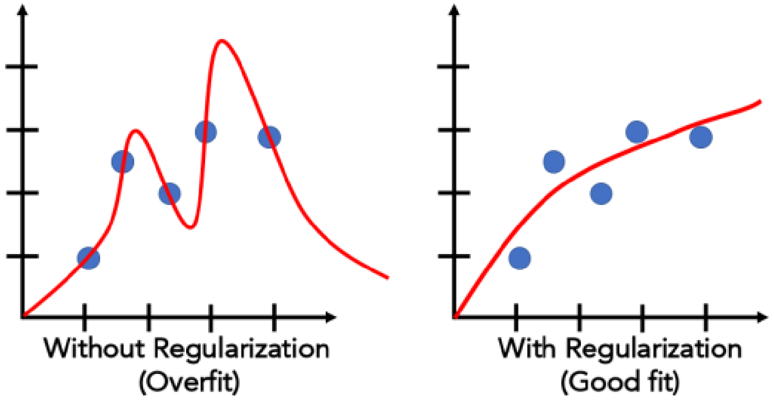
$$w = (X^\top X)^+ X^\top Y = X^+ Y$$

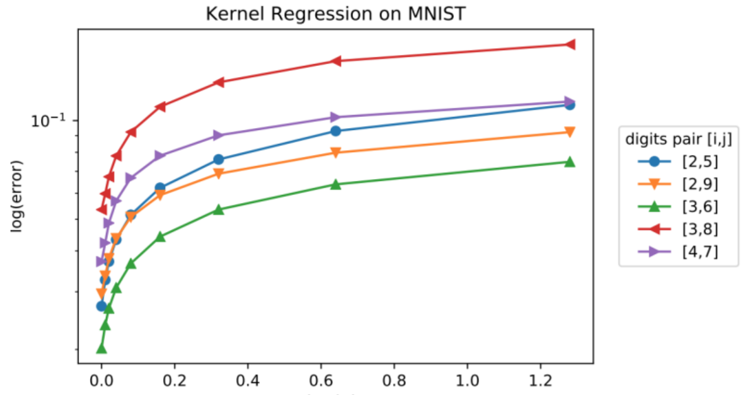
Regularized linear regression

- Ordinary least squares (**OLS**) vs Regularized least squares (**RLS, ridge regression**)
- $\underset{w}{\arg\min}\ \|Xw - y\|_2^2 + \lambda\|w\|_2^2$



- Why regularize?
  - Control the complexity of predictor
  - Avoid overfitting





- When does the regularization **not help**?
  - Regression function is in the class & there is no label noise

https://tyliang.github.io/Tengyuan.Liang/pdf/slides-Duke.pdf

Variations: LASSO

- LASSO: replaces $\lambda \|w\|_2^2$ with $\lambda \|w\|_1$
  - variable selection property => most coefficients are 0
  - Under some mathematical assumptions & the right λ value, researchers have shown that features with zero coefficients are truly irrelevant features.
  - Prediction error is almost as good as an "oracle" linear regression that is run with only those relevant features.
  - no more closed form => iterative methods

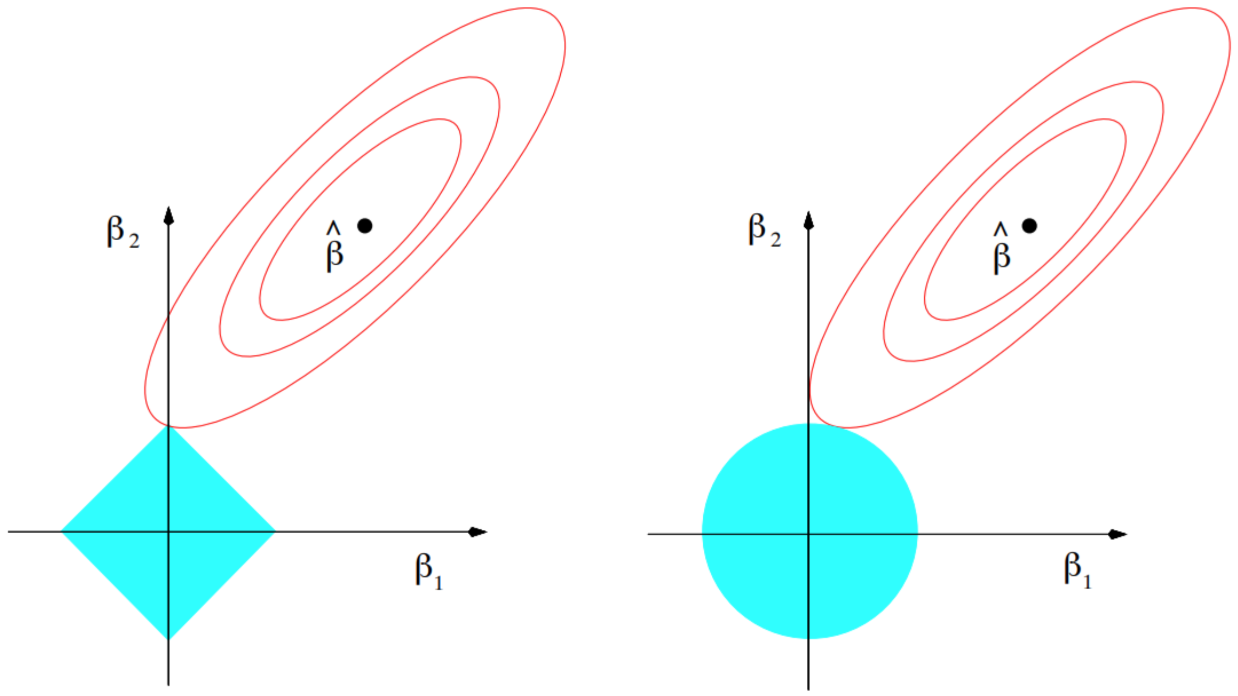
- **A big open problem in ML**: being able to throw in all the possible features in, but still perform as good as knowing the truly relevant features ahead of time (i.e., not affected by irrelevant features)
  - Recall irrelevant features can be harmful.

- LASSO is close, but it works under some assumptions only, and only for the linear model.

https://www.stat.cmu.edu/~ryantibs/statml/lectures/sparsity.pdf

LASSO prefers sparse solutions: intuition

- $\arg\min_{w} \|Xw - y\|_2^2 + \lambda\|w\|_1$
- Constrained optimization form: $\arg\min_{w:\|w\|_1 \le R_\lambda} \|Xw - y\|_2^2$ for some R_λ



How LASSO are often used in practice

- Treat λ as a hyperparameter

- Let $\Lambda = \{10^{-3}, 10^{-2}, \dots\}$
- For $\lambda \in \Lambda$:
  - Run LASSO(λ) on $S \Longrightarrow$ obtain w'
  - $B_\lambda \leftarrow \{i : w'_i \neq 0\}$
  - Train OLS on S but only use features in B_λ, obtain $\hat{w}_\lambda$
- Use validation set to choose $\hat{w} \in \{\hat{w}_\lambda :\ \lambda \in \Lambda\}$

Probabilistic point of view

- So far, we motivated OLS from the ERM principle.
- Statisticians would have described it differently!
- Probabilistic model on data: $X \sim \mathcal{D}_{\mathcal{X}}$ $\quad X \in \mathbb{R}^d$

$$Y \mid X \sim N(X^\top w^*, \sigma^2)$$

maximum likelihood estimation (MLE)

$$\hat{w} = \arg\max_w \prod_{i=1}^n p_w(X = x_i, Y = y_i)$$

$$= \arg\max_w \prod_{i=1}^n p_w(Y = y_i \mid X = x_i) \cdot \underbrace{\cancel{\prod_i p(X = x_i)}}_{\text{independent of } w.}$$

$$= \arg\max_w \sum_{i=1}^n \log p_w(Y = y_i \mid X = x_i).$$

$\circledast$ pdf of $Z \sim N(\mu, \sigma^2)$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right)$$

log pdf

$$-\frac{(z-\mu)^2}{2\sigma^2} + \frac{1}{2}\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)$$

$$\begin{cases} z \leftarrow y_i \\ \mu \leftarrow w^\top x_i \end{cases}$$

$$= \arg\max_w -\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - w^\top x_i)^2 + \underbrace{\sum_i \frac{1}{2}\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)}_{\text{does not have } w.}$$

$$= \arg\max_w -\sum_{i=1}^n (y_i - w^\top x_i)^2$$

$$= \arg\min_w \sum_{i=1}^n (y_i - w^\top x_i)^2 \Rightarrow \text{ERM!!}$$

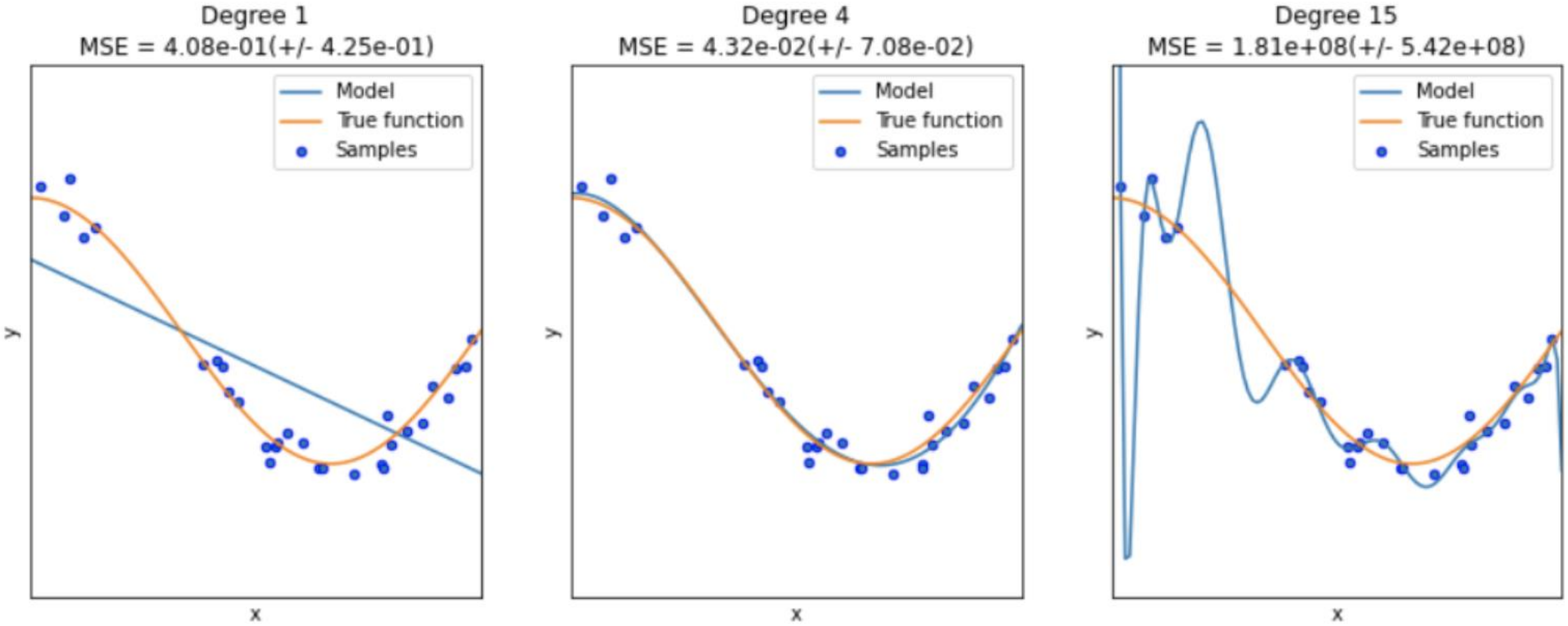
Beyond linearity

- Introduce nonlinear mapping with basis functions $\phi: \mathbb{R}^d \to \mathbb{R}^{d'}$:
  - $\phi(x) = (x^2, x, 1)$: 2nd order polynomial
  - $\phi(x) = (x^d, x^{d-1}, \dots, 1)$: d-th order polynomial (= degree d)

- Higher order => strictly larger class of predictors

$$\mathcal{F} = \{h\colon\ h(x) = \langle w, \phi(x) \rangle, \text{for some } w \in \mathbb{R}^{d'}\}$$

Feature embedding trick



- overfitting vs underfitting
- bias-variance tradeoff.

$$\mathrm{err}(\hat{f}) \;=\; [\mathrm{err}(\hat{f}) - \min_{f^* \in \mathcal{F}} \mathrm{err}(f^*)] \;+\; \min_{f^* \in \mathcal{F}} \mathrm{err}(f^*)$$

from https://datascience.foundation/sciencewhitepaper/underfitting-and-overfitting-in-machine-learning

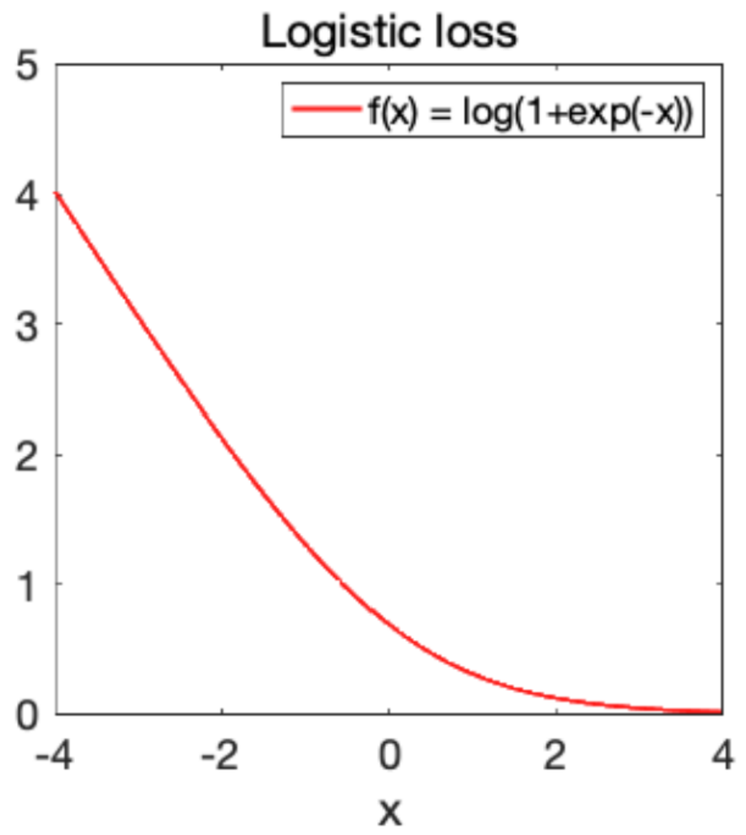
Convexity

This is why setting the gradient = 0 gives optimal solutions

Motivation

- What if the loss function is not quadratic?

- E.g., classification: $x \in \mathbb{R}^d, y \in \{-1,1\}$
- logistic loss: $\ell(w; x, y) = \log(1 + e^{-y \cdot w^\top x})$



Convex sets

- [Def] A set C is convex if

  $\forall u, v \in C, \forall \alpha \in [0,1]$, we have $\alpha u + (1-\alpha)v \in C$　　　convex combination

(○)

(×).

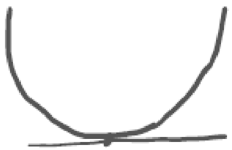
Convex function: intuition

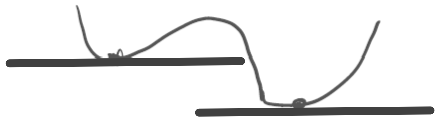
- **Informally**,
  - A convex function is one that looks “convex” from the bottom
  - A convex function has only one “valley”

Convex functions

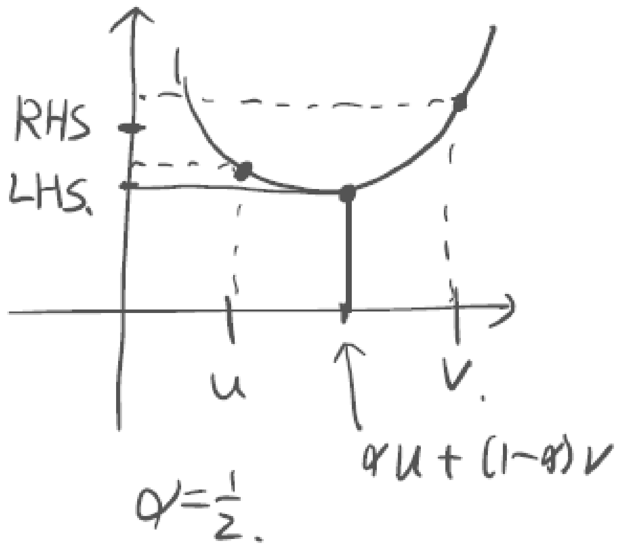
Nonconvex function

- Why setting $\nabla f(w) = 0$ for convex f yields a minimizer?

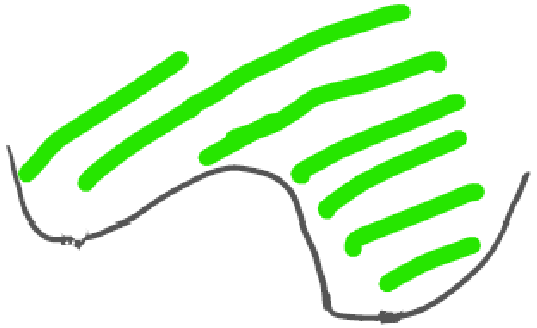




Convex function: definitions



- [Def] Let C be a convex set. A function $f: C \to \mathbb{R}$ is convex if $\forall u, v \in C$ and $\forall \alpha \in [0,1]$,
$$f(\alpha u + (1-\alpha)v) \leq \alpha f(u) + (1-\alpha)f(v)$$
- [Def] concave: change '$\leq$' to '$\geq$'
- (Thm) $f: C \to \mathbb{R}$ is convex if and only if its epigraph $\mathrm{epi}(f) = \{(x,t):\ f(x) \leq t\}$ is a convex set



- Convex functions are easy to optimize
  - Imagine "dropping a ball on the surface"

Exercise: show $h(x) = x^2$ is convex

- Goal: show $(\alpha v + (1-\alpha)u)^2 \le \alpha v^2 + (1-\alpha)u^2$ for all $\alpha \in [0,1]$

$$\Leftrightarrow \alpha^2 v^2 + 2(1-\alpha)\alpha u v + (1-\alpha)^2 u^2 - \alpha v^2 - (1-\alpha)u^2 \le 0$$

proof. $$\left((1-\alpha)^2 - (1-\alpha)\right)u^2 + 2(1-\alpha)\alpha u v + (\alpha^2 - \alpha)v^2$$

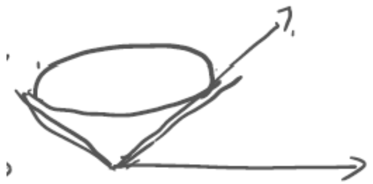
$$= (\alpha^2 - \alpha)u^2 + 2(1-\alpha)\alpha u v + (\alpha^2 - \alpha)v^2$$

$$= \alpha(1-\alpha)(-u^2 + 2uv - v^2)$$

$$= \alpha(1-\alpha)\cdot(-1)(u-v)^2 \le 0$$

Properties

- (a) -f is concave ⇔ f is convex
- (b) linear functions are both convex and concave
- (c) Norms are convex (norms: see Zico Kolter note 3.5)



- Let f, g be convex.
  - (d) max{f(x), g(x)} is convex

  - (e) f(x) + g(x) is convex
  - (f) if g is nondecreasing, then h(x) := g(f(x)) is convex => e.g., $h(w) = \|w\|^2$
- (g) f is concave, g is convex and nonincreasing, then h(x) := g(f(x)) is convex. e.g $h(x) = \frac{1}{\log(1+x)}, x \geq 0$
- (h) convexity is invariant under affine maps:
  if f is convex, then $f(Ax + b)$ is also convex where $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n$
  (this includes linear maps, of course)

(Thm) the OLS objective function is convex.

$$F(w) \coloneqq \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

- Is $f_i(w) = (w^\top x_i - y_i)^2$ convex?
- Yes, it is $h(g(w))$, a composition of $h(z) = z^2$ and affine mapping $g(w) = w^\top x_i - y_i$

- Is the RLS objective $F_\lambda(w) \coloneqq \sum_{i=1}^{n} (w^\top x_i - y_i)^2 + \lambda \|\mathrm{w}\|^2$ convex? What about the LASSO objective?

Check convexity: an oftentimes more convenient criterion

- (Prop) Let a function $f: \mathbb{R} \to \mathbb{R}$ be twice differentiable on a convex set $C \subseteq \mathbb{R}$

  Then, f is convex $\Leftrightarrow f''(x) \geq 0, \forall x \in C$

- [Def] $A \in \mathbb{R}^{d \times d}$ is positive semi-definite (PSD) $\Leftrightarrow x^\top A x \geq 0 \ \forall x \in \mathbb{R}^d$
  - notation: $A \succcurlyeq 0$
  - analogue of nonnegative coefficient in 1d.
  - (prop) Suppose A is symmetric. Then, A is PSD $\Leftrightarrow \mathrm{eigval}_i(A) \geq 0, \forall i$

$\underline{a}x^2 \longrightarrow x^\top A x.$

$a > 0.$ $\quad A \succeq 0.$

- (Prop) Let a function $f: \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable on a convex set $C \subseteq \mathbb{R}^d$.

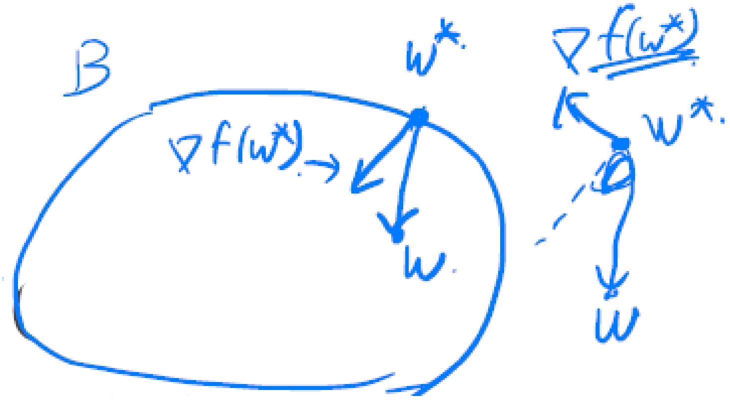
  Then, f is convex $\Leftrightarrow \nabla^2 f(x)$ is PSD, $\forall x \in C$

Showing $h(x) = x^2$ is convex: an alternative proof

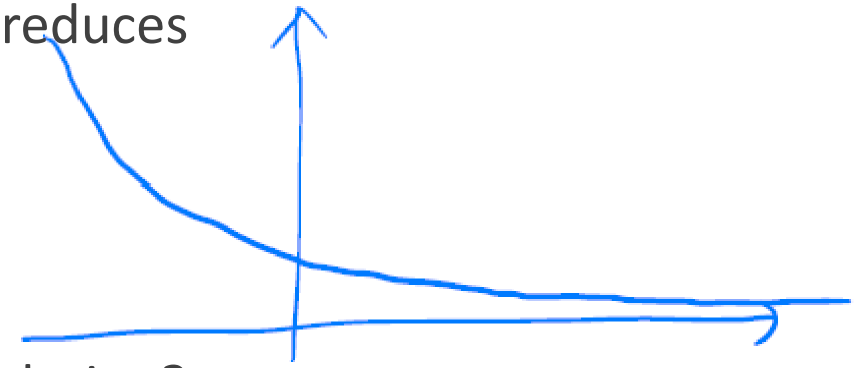
- $C = \mathbb{R}$
- For all $x \in C$:
- $h'(x) = 2x$
- $h''(x) = 2 \geq 0$

So we know it's convex. But why derivative = 0?

- (Thm) [Optimality condition]
  Let f be convex and differentiable, B be a convex set. Then,
  $w^* \in \arg\min_w f(w) \quad \text{s.t.} \quad \mathrm{w} \in B \quad \Leftrightarrow$
  $$\begin{cases} w^* \in B \\ \forall w \in B, \quad \nabla f(w^*)^\top (w - w^*) \geq 0 \end{cases}$$



- Furthermore, if $B = \mathbb{R}^d$ (unconstrained), then the RHS above reduces to $\nabla f(w^*) = 0$



- Q: does this tell us something about existence of an optimal solution?

Next lecture (9/21)

- Linear classification; regularized loss minimization formulations
- Support Vector Machines (SVMs)
- Assigned Reading: CIML Section 7.7