CPSC 340/540 Tutorial 4

Winter 2024 Term 1

T1A: Tuesday 16:00-17:00;
T1C: Thursday 10:00-11:00;
Office Hour: Wednesday 15:00-16:00

Slides can be found at Piazza and my personal page after T1C.

Yi (Joshua) Ren

https://joshua-ren.github.io/
renyi.joshua@gmail.com
PhD with Danica

Machine Learning:
Learning dynamics, LLM, Compositional Generalization

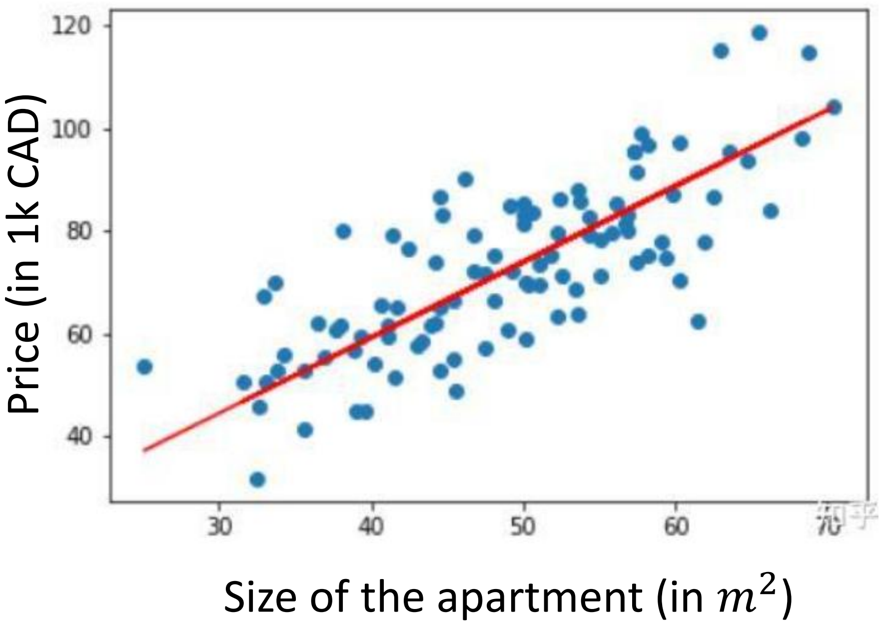
More helpful on theory
Less helpful on coding

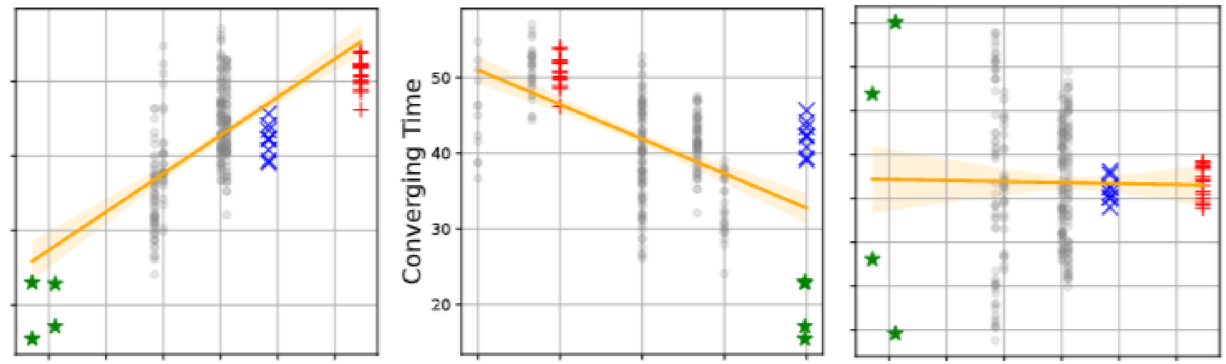
Slides Credit: To various pervious TA's of this course

- Linear Regression
- Some mid-term questions

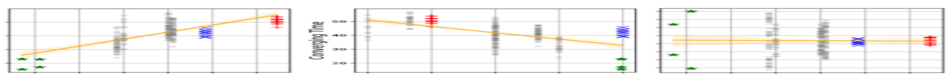
Regression: (fundamentals)

- Sutiable tasks: if we want a model to
  - ➢ Predict a numerical value given features
    - Here is an apartment with $50m^2$, can you estimate its price?
    - Tom bought an apartment with 80k CAD, can you guess how big it is?
  - ➢ Find linear correlation relationship between two variables
    - Is the price of an apartment is influenced by its size?
    - What about the initial letter of the apartment's owner?





Be careful about the scale of axis



➢ Correlation is not causality (switch X and Y, LR is similar)

Regression: (formulars, start from 1-d problem)

- The model, parameterized by w, makes prediction using: $\hat{y}_i = w\tilde{x}_i$
- How good each prediction is is estimated using L2-distance: $r_i = \hat{y}_i - y_i$
- The total residual for the training dataset:

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

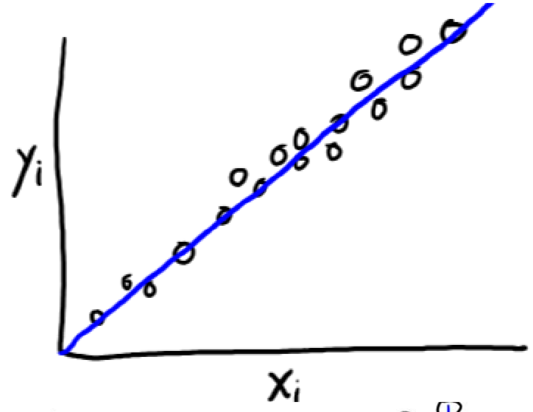
Sum up the squared differences over all training examples.

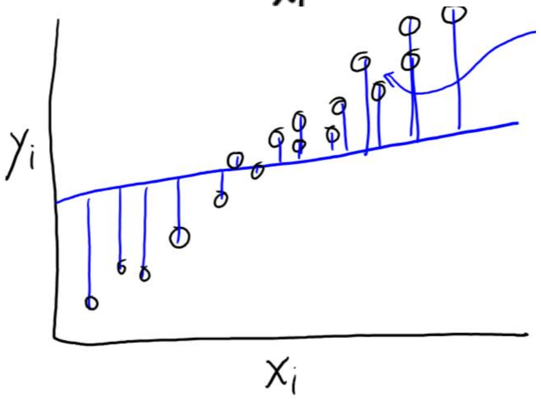
$w x_i$ → Our prediction $\hat{y}_i$

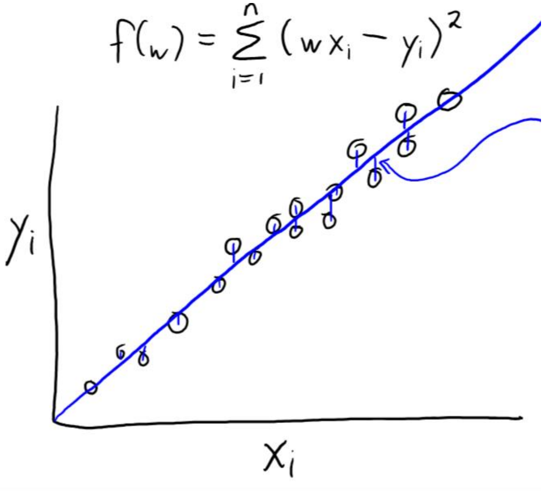
y_i → True value of y_i

$(w x_i - y_i)$ → Difference between prediction and true value for example i (residual)

- Our target is to find good w that makes residual for the test set small. To achieve this, minimize f(w) on training set.







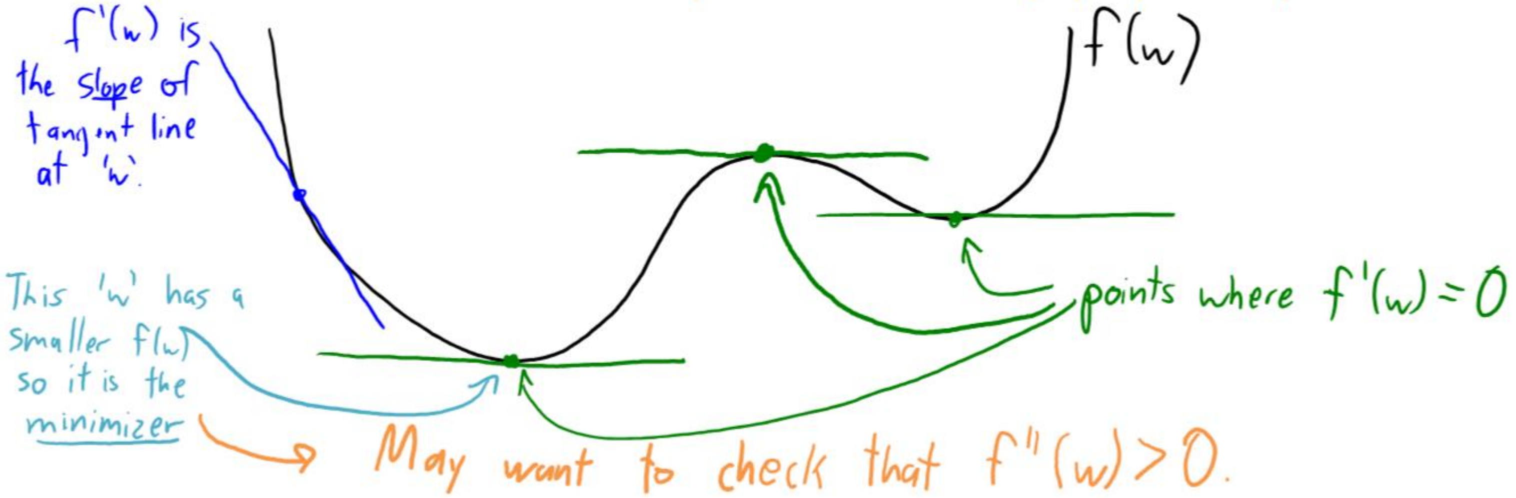
Regression: (solve it in closed-form, 1-dim)

- Training a regression model is equivalently solving the following optimization problem:

$$\min_{w} \frac{1}{2}\sum_{i=1}^{n}(wx_i - y_i)^2$$

- Recap how we find the optimum solution for 1-d case:

1. Take the derivative of ‘f’.
2. Find points ‘w’ where the derivative f’(w) is equal to 0.
3. Choose the smallest one (and check that f’’(w) is positive).



$$f(x) = \frac{1}{2}\sum_{i=1}^{n}(wx_i - y_i)^2$$

$$= \frac{w^2}{2}\sum_{i=1}^{n}x_i^2 - w\sum_{i=1}^{n}x_i y_i + \frac{1}{2}\sum_{i=1}^{n}y_i^2$$

$$= \frac{w^2}{2}a - wb + c$$

$$f'(w) = wa - b$$

$$\boldsymbol{w}^* = \frac{\boldsymbol{b}}{\boldsymbol{a}} = \frac{\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{y}_i}{\sum_{i=1}^{n}\boldsymbol{x}_i^2}$$

$f''(w) = a$, always ≥ 0

Regression: (high-d, matrix form)

- Usually, $\boldsymbol{x}$ is features rather than raw inputs, it might contains multiple dimensions:

$$\hat{y}_i = w_1 x_{i1} + w_2 x_{i2}$$

Value of feature 2 in example 'i'

"weight" on feature 2.

"weight" of feature 1

Value of feature 1 in example 'i'

$$\hat{y}_i = w_1 x_{i1} + w_2 x_{i2} + w_3 x_{i3} + \cdots + w_d x_{id}$$

- We can **design different features**, recall our polynomial regression problem:

$$f = w_0 + w_1 x + w_2 x^2 + \cdots + w_n x^n + k|W|_2^2 = W \begin{bmatrix} x^0 \\ \dots \\ x^{\boldsymbol{n}} \end{bmatrix} + \boldsymbol{k}|W|_2^2$$

- For notation conciseness, and also to better utilize math tools in linear algebra, we prefer matrix form

$$f(w_1, w_2, \dots, w_d) = \sum_{i=1}^{n} \left(\sum_{j=1}^{d} w_j x_{ij} - y_i \right)^2 \Rightarrow f(w) = \|Xw - y\|^2$$

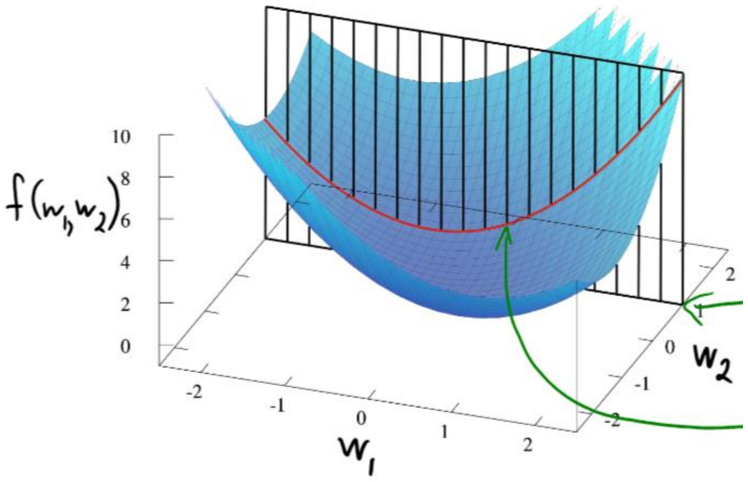
Regression: (high-d, matrix form)

$$f(w_1, w_2, \dots, w_d) = \frac{1}{2}\sum_{i=1}^{n}\Big(\underbrace{\sum_{j=1}^{d} w_j x_{ij}}_{\hat{y}_i} - y_i\Big)^2$$

- Then for a high-dimension case, we extend derivative to **gradients** (stacking of partial derivatives)

$$\underbrace{\nabla f(w)}_{\text{gradient vector}} = \begin{bmatrix} \frac{\partial f}{\partial w_1} \\ \frac{\partial f}{\partial w_2} \\ \vdots \\ \frac{\partial f}{\partial w_d} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n}\left(\sum_{j=1}^{d} w_j x_{ij} - y_i\right) x_{i1} \\ \sum_{i=1}^{n}\left(\sum_{j=1}^{d} w_j x_{ij} - y_i\right) x_{i2} \\ \vdots \\ \sum_{i=1}^{n}\left(\sum_{j=1}^{d} w_j x_{ij} - y_i\right) x_{id} \end{bmatrix}$$

$\frac{\partial f}{\partial w_2}$: partial derivative of f with respect to variable w_2



- Set this gradient to 0 vector:

$$\nabla f(w) = 0 \iff \begin{cases} \sum_{i=1}^{n}\left(\sum_{j=1}^{d} w_j x_{ij} - y_i\right) x_{i1} = 0 \\ \sum_{i=1}^{n}\left(\sum_{j=1}^{d} w_j x_{ij} - y_i\right) x_{i2} = 0 \\ \vdots \\ \sum_{i=1}^{n}\left(\sum_{j=1}^{d} w_j x_{ij} - y_i\right) x_{id} = 0 \end{cases}$$

Regression: (example: with L2 regularizer)

Express the following functions in terms of vectors, matrices, and norms (there should be no summations or maximums),

$$f(w) = \frac{1}{2}\sum_{i=1}^{n}(w^T x_i - y_i)^2 + \frac{\lambda}{2}\sum_{j=1}^{d} w_j^2$$

Recall, that all vectors are column-vectors,

w_j is the scalar parameter j.
y_i is the label of example i.
x_i is the column-vector of features for example i.
x_j^i is feature j in example i.

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad x_i = \begin{bmatrix} x_1^i \\ x_2^i \\ \vdots \\ x_d^i \end{bmatrix}$$

Regression: (matrix form and with L2 regularizer)

Let's first focus on the regularization term,

$$f(w) = \frac{1}{2}\sum_{i=1}^{n}(w^T x_i - y_i)^2 + \frac{\lambda}{2}\sum_{j=1}^{d} w_j^2$$

Recall the definition of inner product and L2-norm of vectors,

$$\|v\| = \sum_{j=1}^{d} v_j^2 \quad u^T v = \sum_{j=1}^{d} u_j v_j$$

Hence, we can write the regularizer in various forms using,

$$\|w\|^2 = \sum_{j=1}^{d} w_j^2 = \sum_{j=1}^{d} w_j w_j = w^T w$$

Regression: (matrix form and with L2 regularizer)

Let's next focus on the least squares term,

$$f(w) = \frac{1}{2}\sum_{i=1}^{n}(w^T x_i - y_i)^2 + \frac{\lambda}{2}\sum_{j=1}^{d} w_j^2$$

Let's define the residual vector r with elements

$$r_i = w^T x_i - y_i$$

We can write the least squares term as squared L2-norm of residual,

$$\sum_{i=1}^{n}(w^T x_i - y_i)^2 = \sum_{i=1}^{n} r_i^2 = r^T r = \|r\|^2$$

Regression: (matrix form and with L2 regularizer)

Let's next focus on the least squares term,

$$f(w) = \frac{1}{2}\|r\|^2 + \frac{\lambda}{2}\|w\|^2, \quad r_i = w^T x_i - y_i$$

X denotes the matrix containing the x_i (transposed) in the rows: $X = \begin{bmatrix} — (x_1)^T — \\ — (x_2)^T — \\ \vdots \\ — (x_n)^T — \end{bmatrix}$

Using $w^T x_i = (x_i)^T w$ and the definitions of r, y, and X:

$$r = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{bmatrix} = \begin{bmatrix} w^T x_1 - y_1 \\ w^T x_2 - y_2 \\ \vdots \\ w^T x_n - y_n \end{bmatrix} = \begin{bmatrix} (x_1)^T w \\ (x_2)^T w \\ \vdots \\ (x_n)^T w \end{bmatrix} - \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{y} = \underbrace{\begin{bmatrix} — (x_1)^T — \\ — (x_2)^T — \\ \vdots \\ — (x_n)^T — \end{bmatrix}}_{X} w - y = Xw - y$$

Therefore $f(w) = \frac{1}{2}\|Xw - y\|^2 + \frac{\lambda}{2}\|w\|^2$,

Regression: (matrix form and with L2 regularizer)

A quadratic function is a function of the form

$$f(w) = \frac{1}{2} w^T A w + b^T w + y,$$

for a square matrix A, vector b, and scalar y.

Write the minimizer of the following function as a system of linear equations, using vector/matrix notation.

$$f(w) = \frac{1}{2} \|Xw - y\|^2 + \frac{\lambda}{2} \|w\|^2,$$

minimize convex functions, it is sufficient to find w $s.t$

$$f(w) = 0.$$

Regression: (matrix form and with L2 regularizer)

Convert to vector/matrix form:

$$f(w) = \frac{1}{2}\|Xw - y\|^2 + \frac{\lambda}{2}\|w\|^2 = \frac{1}{2}(Xw - y)^T(Xw - y) + \frac{\lambda}{2}w^T w$$

$$\rightarrow f(w) = \frac{1}{2}w^T X^T X w - w^T X^T y + \frac{1}{2}y^T y + \frac{\lambda}{2}w^T w$$

Find w such that $f'(w) = 0$:

$$f'(w) = X^T X w - X^T y + \lambda w = 0 \rightarrow (X^T X + \lambda I)w = X^T y$$

Note $f(w)$ is a column vector with dimension $d \times 1$.

- $f(w) = a^\top w$
  - $\nabla_w f(w) = a$
  - $\nabla_w^2 f(w) = 0$
- $f(w) = w^\top A w$
  - $\nabla_w f(w) = (A^\top + A)w$
  - $\nabla_w^2 f(w) = (A^\top + A)$ (=

Regression: (matrix form and with L2 regularizer)

- When $\lambda = 0$, compare the following two forms:

$$f(x) = \frac{1}{2}\sum_{i=1}^{n}(wx_i - y_i)^2 \qquad f(x) = \frac{1}{2}\|Xw - y\|_2^2$$

$$w^* = \left(\sum_{i=1}^{n} x_i^2\right)^{-1}\left(\sum_{i=1}^{n} x_i y_i\right) \qquad w^* = (X^T X)^{-1} X^T y$$

More convinent if you know how to compute matrix derivatives.

The Matrix Cookbook

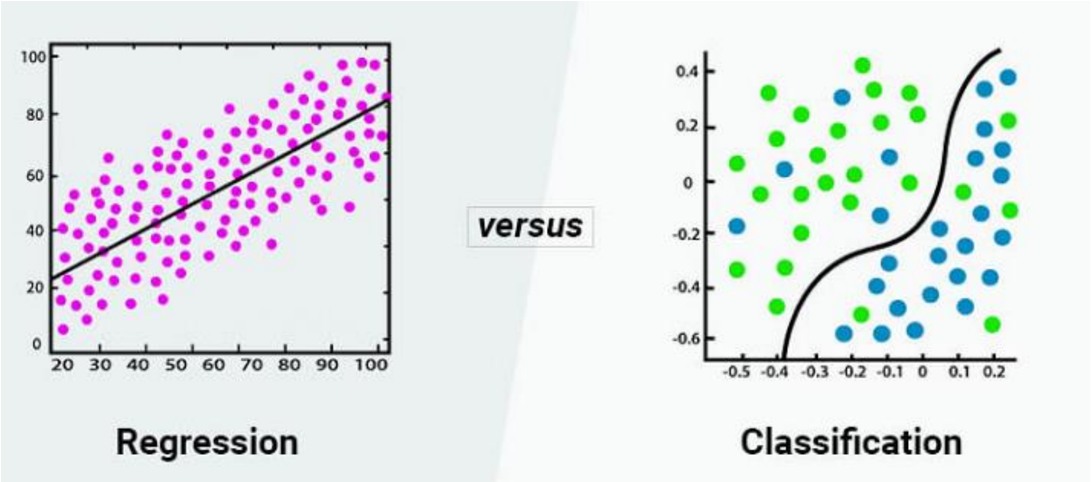
[http://matrixcookbook.com]

Kaare Brandt Petersen
Michael Syskind Pedersen

Version: November 15, 2012

https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

Regression: (v.s. Classification)



- Similarities:
  - Both are supervised learning: x is dataset, y is label, model $y = h_w(x)$, parameterized by w
  - Almost identical expression for linear model: $y = Xw$
  - L2 can be used as a default loss function: $\mathcal{L}(h_w(x), \bar{y}) = \frac{1}{2}\left|\left|Xw - \bar{y}\right|\right|_2^2 + r(w)$
- Differences – Output part:
  - Regression: $\bar{y} \in \mathbb{R}$, $y_1 > y_2$ means sth., $y_1 + y_2$ means sth.
  - Classification: $\bar{y} \in [K]$, $y_1 > y_2$ or $y_1 + y_2$ means NOTHING
  - Regression: $h_w \in \mathbb{R}$, usually the same space as $\bar{y}$, then L2 loss is a reasonable measurement
  - Classification: output can be a distribution $h_w = p(y|x) \in [0,1]^K$, L2 loss works, but not the best
    usually not the same space as $\bar{y} \in [K]$, one-hot encoding is usually applied

Regression: (v.s. Classification)

- **Interchangable:**

  - A regression task can also be solved using a classification framework:
    - **Discretize**, e.g., age → {age<20, 20<age<30, age>30}
    - Can **introduce non-convexity**, e.g., age → {age<20 or age>30, 20<age<30}

  - A classification can also be solved using a regression framework (L2-loss):
    - Use **one-hot encoding** to convert label to a distribution
    - Directly use L2 loss
    - Use **argmax** when making predictions

  - Usually the default setting for the last layer of a DNN

Regression: (different regularizers)

Unit ball, i.e., $\|\mathrm{x}\|_p = 1$

- Recap of different norms

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

  - L0-norm: non-zero elements in a vector
  - L1-norm: usually use to introduce sparsity (vertex at axis)
  - L2-norm: Gaussian, Euclidian distance, most common
  - L∞-norm: select the maximum value

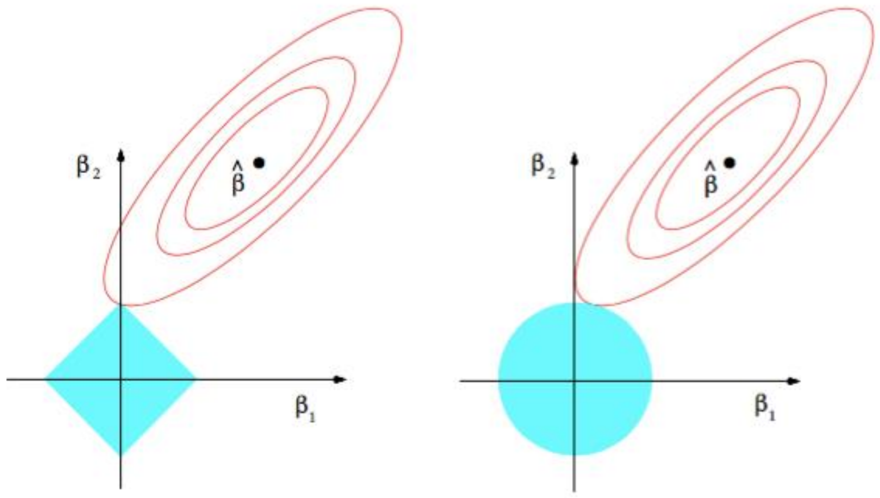
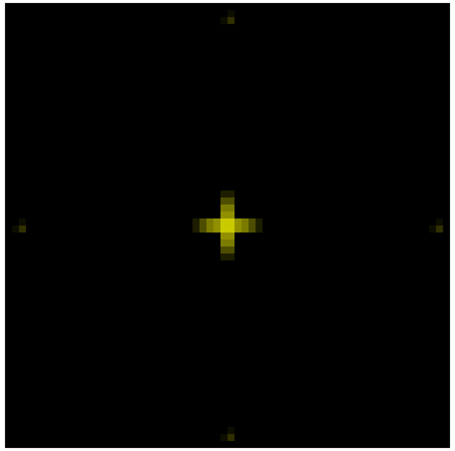
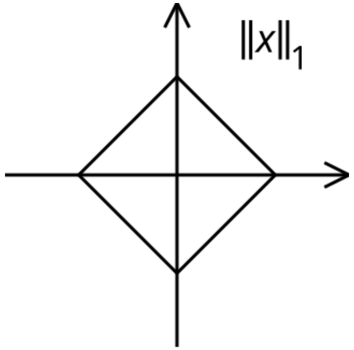
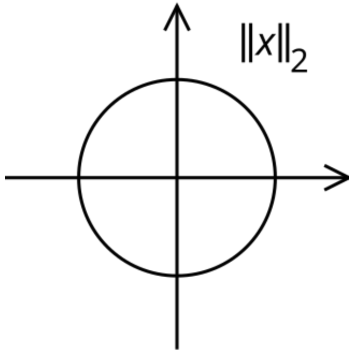


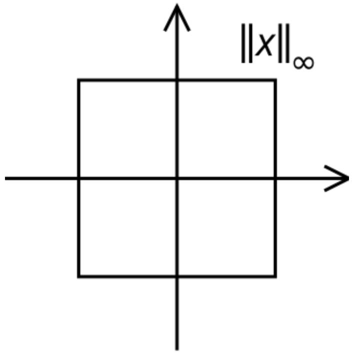
FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function.*

Unit ball, p=0 to 2









Thanks for your time!
Questions?