Machine Learning
Regularization and Binary Data

Rajesh Ranganath

- What happens if there are too many features?
- What if the outcome variable is binary?

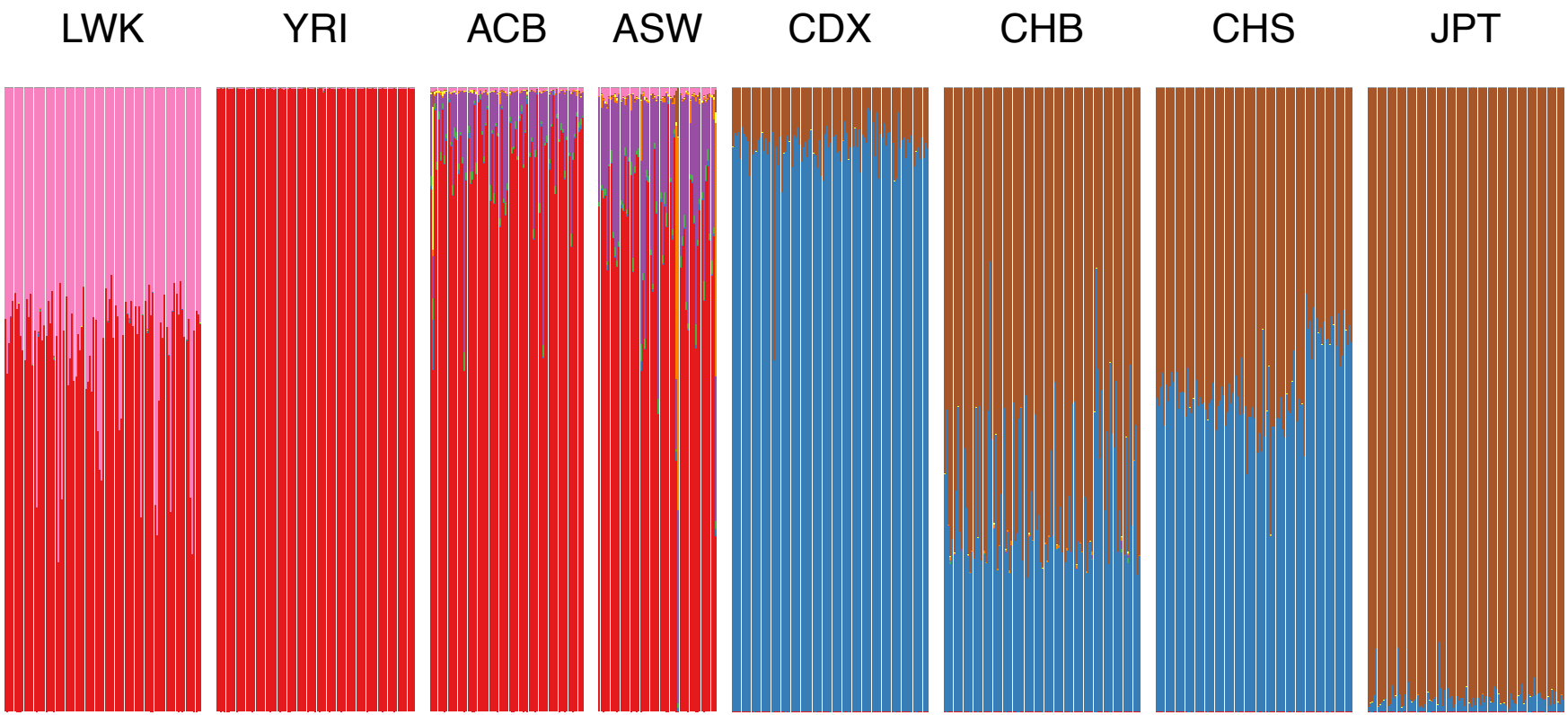
Crohn's Disease

Goal: Understand Role of Genetics in Severity of Crohn's Disease

How to answer this question?

[Kerras+ 2017]

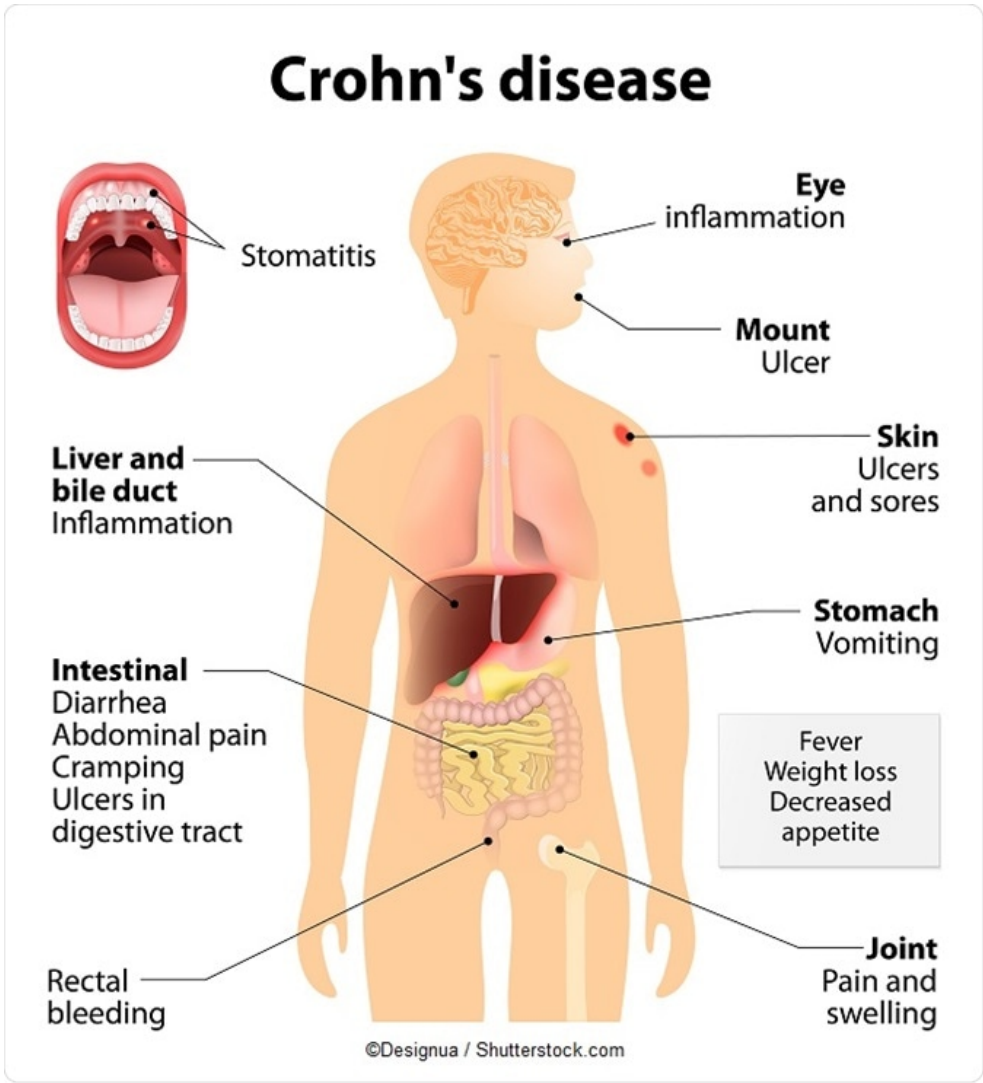
How to answer this question?



[Gopalan+ 2016]

How to answer this question?

Measure their Crohn’s disease severity risk



How to answer this question?

- How many people?
- A high estimate of prevalence is .2% of the US population
- Approximately 600K people in the US
- 10K people would be a broad study

How to answer this question?

How many genetic variants?

> 10 million

How to answer this question?

How many genetic variants?

> 10 million

Math to find importance?

Use Linear Regression!

Math to find importance?

Use Linear Regression!

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{\theta}^\top \mathbf{x}_i - y_i)^2$$

Math to find importance?

Use Linear Regression!

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{\theta}^\top \mathbf{x}_i - y_i)^2$$

- How do we get feature importances?

Math to find importance?

Use Linear Regression!

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{\theta}^\top \mathbf{x}_i - y_i)^2$$

- How do we get feature importances?
- Just $\boldsymbol{\theta}_i$ doesn't work. *Why?*

Math to find importance?

Use Linear Regression!

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{\theta}^\top \mathbf{x}_i - y_i)^2$$

- How do we get feature importances?
- Just $\boldsymbol{\theta}_i$ doesn't work. *Why?*
- Need to standardize $\boldsymbol{\theta}_i$. *How?*

Math to find importance?

Use Linear Regression!

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{\theta}^\top \mathbf{x}_i - y_i)^2$$

- How do we get feature importances?
- Just $\boldsymbol{\theta}_i$ doesn't work. *Why?*
- Need to standardize $\boldsymbol{\theta}_i$. *How?*
- One way: subtract mean and divide by standard deviation

What happens when we do linear regression?

Number of features is greater than number of data points

What happens when we do linear regression?

Number of features is greater than number of data points

Lots of solutions! Given by

$$\mathbf{X}\boldsymbol{\theta} = \mathbf{y}$$

Solutions span $p - n$ subspace (under linear independence of $\mathbf{x}_i$)

Need to restrict model complexity in some way

How do we restrict linear regression?

Model Complexity in Linear Regression

- Throw away variables
- Combine variables
- Sequentially add variables one at a time
- Restrict the parameters of linear regression

Norms

The α-norm of a vector:

$$||\boldsymbol{\theta}||_\alpha = \left(\sum_{i=1}^{p} |\theta_i|^\alpha \right)^{\frac{1}{\alpha}}$$

Common norms included

- $\alpha = 2$: Euclidean norm
- $\alpha = 1$: Manhattan distance
- $\alpha = \infty$: Maximum norm

Norms of difference give you a type of distance

Regularization

Add a function to the minimization

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} (\boldsymbol{\theta}^\top \mathbf{x}_i - y_i)^2 + \lambda ||\boldsymbol{\theta}||_2^2$$

Penalizes the norm of the parameters

Called

- Ridge regression
- L2 regression

Solution

We take the derivative and set it equal to zero

$$\nabla \mathcal{L} = \sum_{i=1}^{n} 2(\boldsymbol{\theta}^\top \mathbf{x}_i - y_i)\mathbf{x}_i + 2\lambda\boldsymbol{\theta}$$

Solution

We take the derivative and set it equal to zero

$$\nabla \mathcal{L} = \sum_{i=1}^{n} 2(\boldsymbol{\theta}^\top \mathbf{x}_i - y_i)\mathbf{x}_i + 2\lambda\boldsymbol{\theta}$$

Setting it equal to zero

$$\boldsymbol{\theta} = \left(\left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \right) + \lambda I) \right)^{-1} \sum_{i=1}^{n} y_i \mathbf{x}_i$$

Or in matrix form

$$\boldsymbol{\theta} = \left(\mathbf{X}^\top \mathbf{X} + \lambda I \right)^{-1} \mathbf{X}^\top y$$

$$\boldsymbol{\theta} = \left(\left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \right) + \lambda I) \right)^{-1} \sum_{i=1}^{n} y_i \mathbf{x}_i$$

- When λ large

$$\left(\left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \right) + \lambda I) \right)^{-1}$$

  is small on diagonal

- When λ small

$$\left(\left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \right) + \lambda I) \right)^{-1}$$

  Looks close to least squares. Single solution!

- When n is large converges to least squares

Understanding l2-regression

Consider a general regularized minimization problem

$$\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) + \lambda g(\boldsymbol{\theta})$$

Understanding l2-regression

Consider a general regularized minimization problem

$$\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) + \lambda g(\boldsymbol{\theta})$$

Now consider an alternative formulation

$$\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$$
$$g(\boldsymbol{\theta}) \leq M$$

Directly limits $g(\boldsymbol{\theta})$ rather than penalizing

Understanding l2-regression

$$\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) + \lambda g(\boldsymbol{\theta})$$

Suppose for a fixed λ, the optimal $\boldsymbol{\theta}$ is $\boldsymbol{\theta}^*$

Understanding l2-regression

$$\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) + \lambda g(\boldsymbol{\theta})$$

Suppose for a fixed λ, the optimal $\boldsymbol{\theta}$ is $\boldsymbol{\theta}^*$

Then in

$$\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$$
$$g(\boldsymbol{\theta}) \leq M$$

set $M = g(\boldsymbol{\theta}^*)$

Understanding l2-regression

Now suppose that there $\hat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}^*$ is the minimum of

$$\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$$
$$g(\boldsymbol{\theta}) \leq M$$

with $M = g(\boldsymbol{\theta}^*)$. That is,

$$f(\hat{\boldsymbol{\theta}}) < f(\boldsymbol{\theta}^*)$$

Putting this together with the constraint gives

$$f(\hat{\boldsymbol{\theta}}) + \lambda g(\hat{\boldsymbol{\theta}}) < f(\boldsymbol{\theta}^*) + \lambda g(\boldsymbol{\theta}^*)$$

This contradicts the minimality of $\boldsymbol{\theta}^*$

Understanding l2-regression

Consider the ridge regression problem

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (\boldsymbol{\theta}^\top \mathbf{x}_i - y_i)^2 + \lambda ||\boldsymbol{\theta}||_2^2$$

Understanding l2-regression

Consider the ridge regression problem

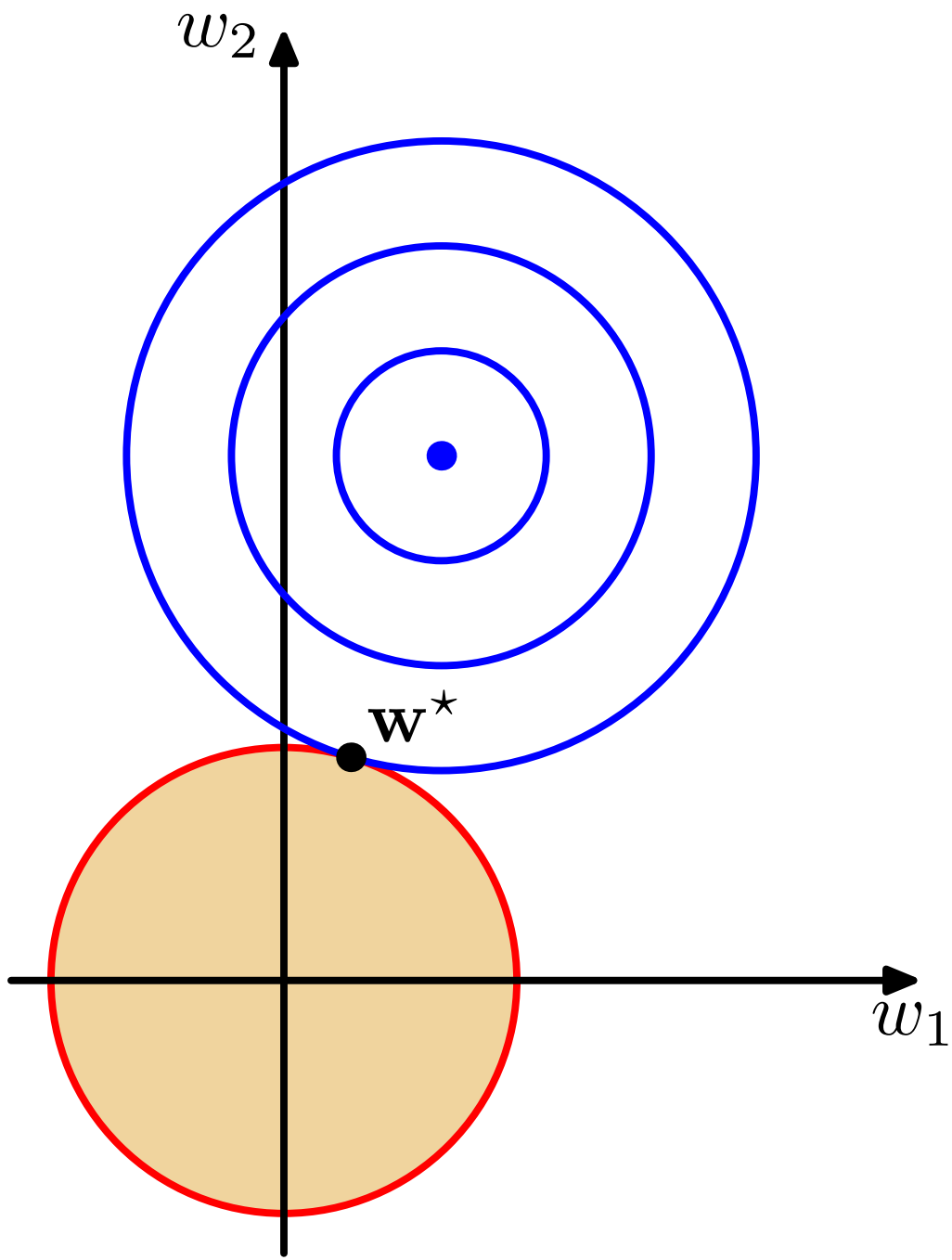
$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (\boldsymbol{\theta}^\top \mathbf{x}_i - y_i)^2 + \lambda ||\boldsymbol{\theta}||_2^2$$

for every λ is equivalent to

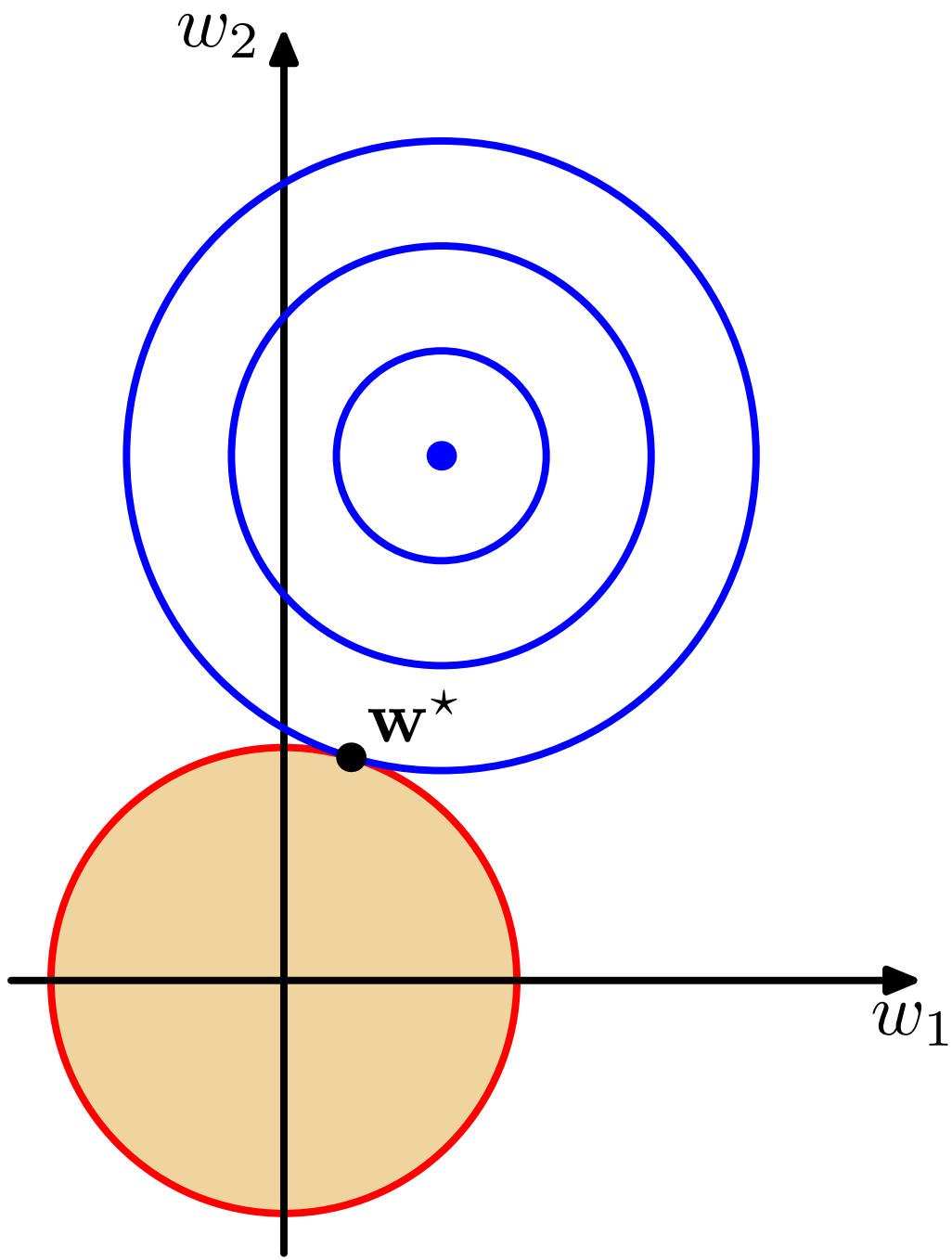
$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (\boldsymbol{\theta}^\top \mathbf{x}_i - y_i)^2$$
$$||\boldsymbol{\theta}||_2^2 \leq M$$

for some M.

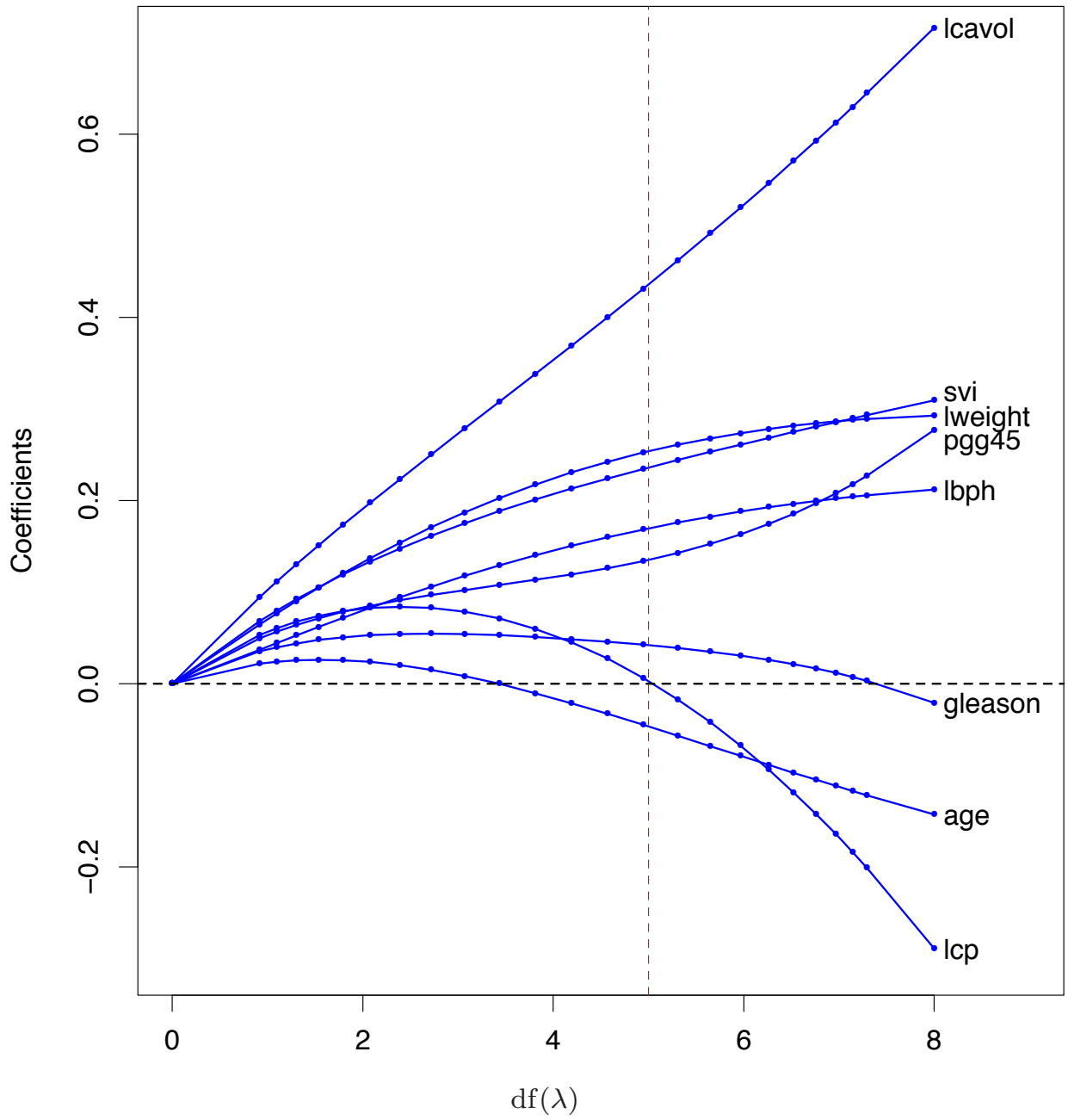
L2 regularized ridge regression constrains parameters to a ball



[Bishop 2006]



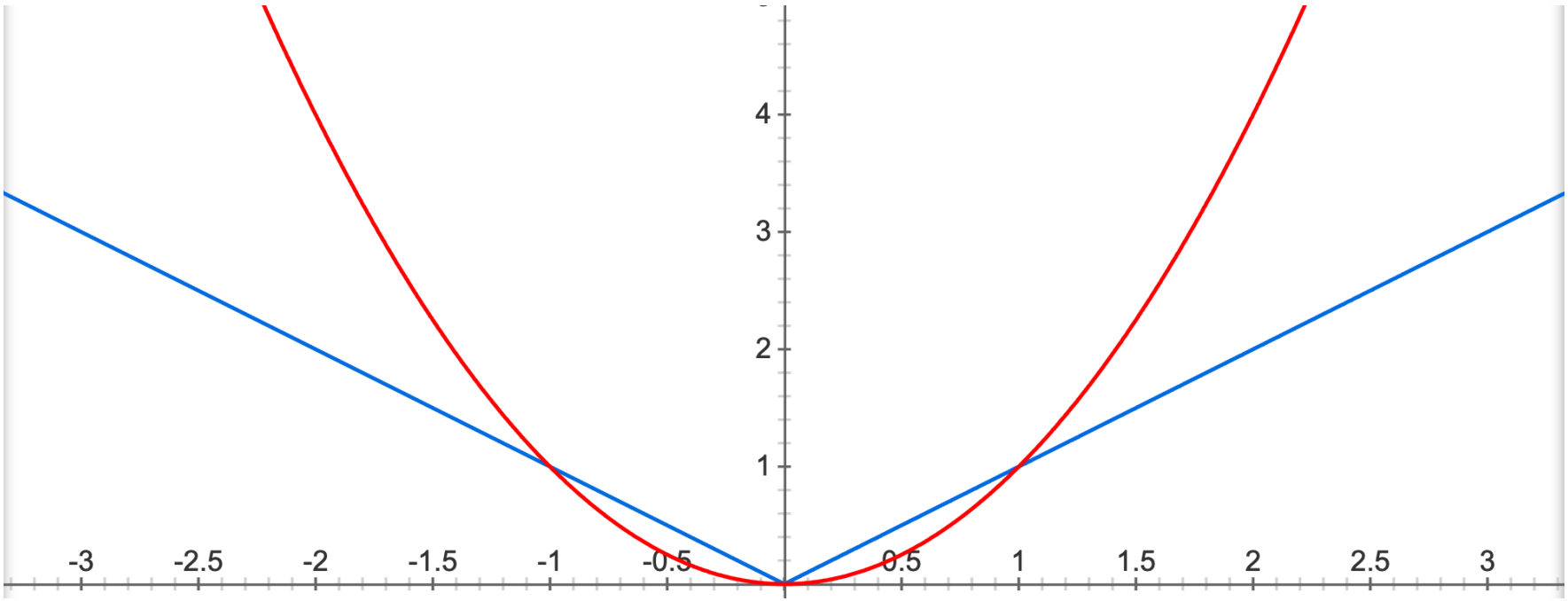
[Bishop 2006] Why do the level sets look like circles?



[Elements of Statistical Learning]

Why not penalize with other norms?

L1 vs L2 norm



Leads to sparsity because of bigger penalties near zero

Regularization

Add a function to the minimization

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} (\boldsymbol{\theta}^\top \mathbf{x}_i - y_i)^2 + \lambda ||\boldsymbol{\theta}||_1$$

Penalizes the norm of the parameters

Called

- Lasso regression
- L1 regression

L1 Regularization: How do we solve?

Add a function to the minimization

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} (\boldsymbol{\theta}^\top \mathbf{x}_i - y_i)^2 + \lambda ||\boldsymbol{\theta}||_1$$

The absolute value is not differentiable. Options

- Can use quadratic programming
- Custom solver?

An Alternative View

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (\boldsymbol{\theta}^\top \mathbf{x}_i - y_i)^2 + \lambda ||\boldsymbol{\theta}||_1$$

An Alternative View

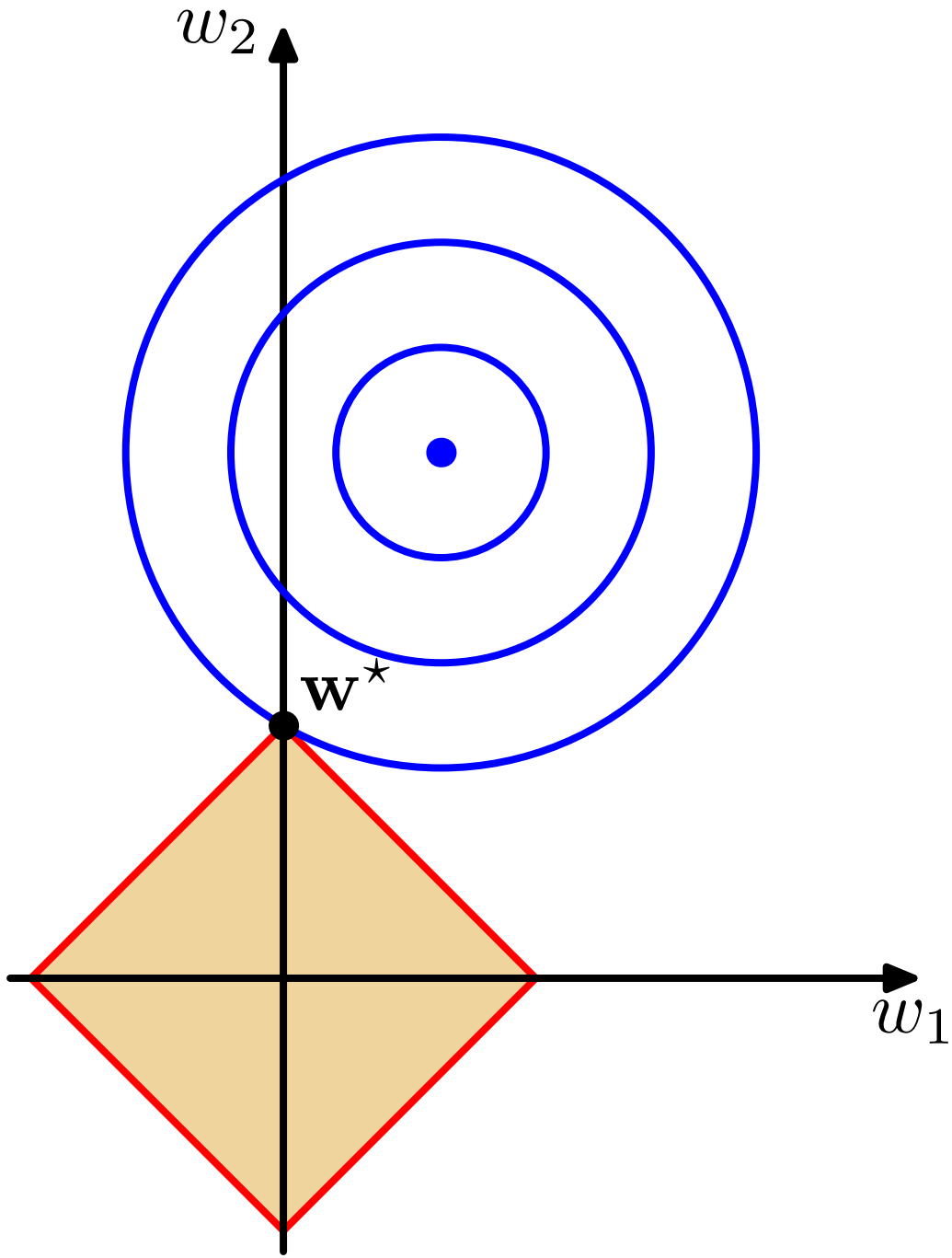
$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (\boldsymbol{\theta}^\top \mathbf{x}_i - y_i)^2 + \lambda ||\boldsymbol{\theta}||_1$$

for every λ is equivalent to

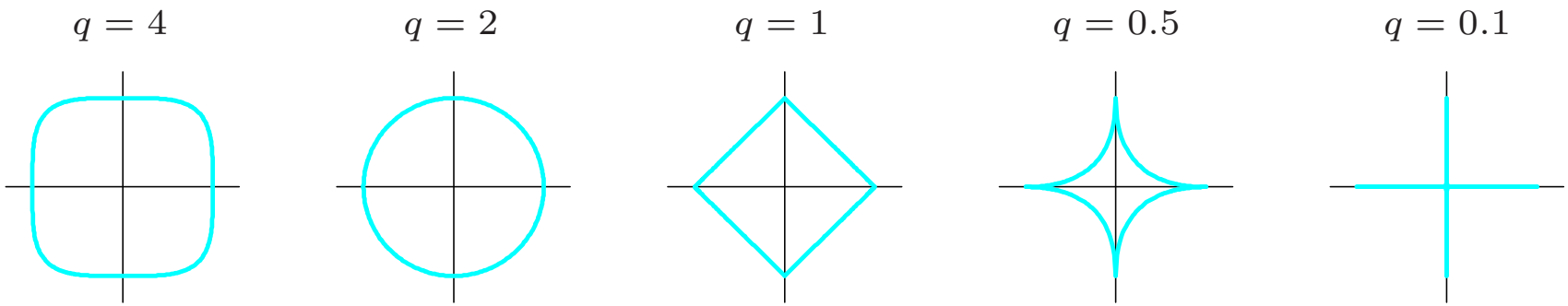
$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (\boldsymbol{\theta}^\top \mathbf{x}_i - y_i)^2$$
$$||\boldsymbol{\theta}||_1 \leq M$$

for some M.

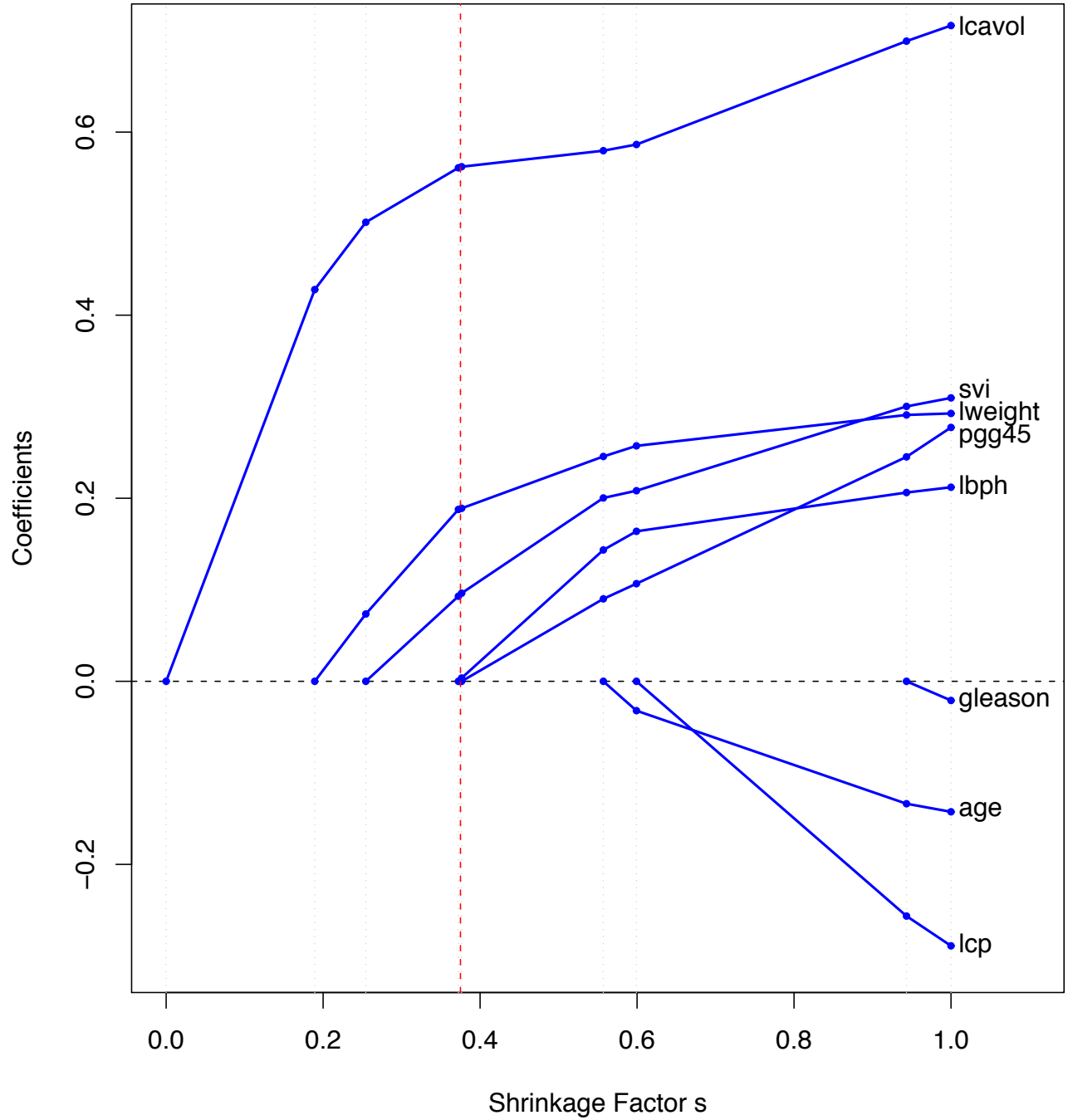
L1 regularization constrains parameters to a diamond



[Bishop 2006]



[Elements of Statistical Learning]



[Elements of Statistical Learning]

Understanding the L1 regularization

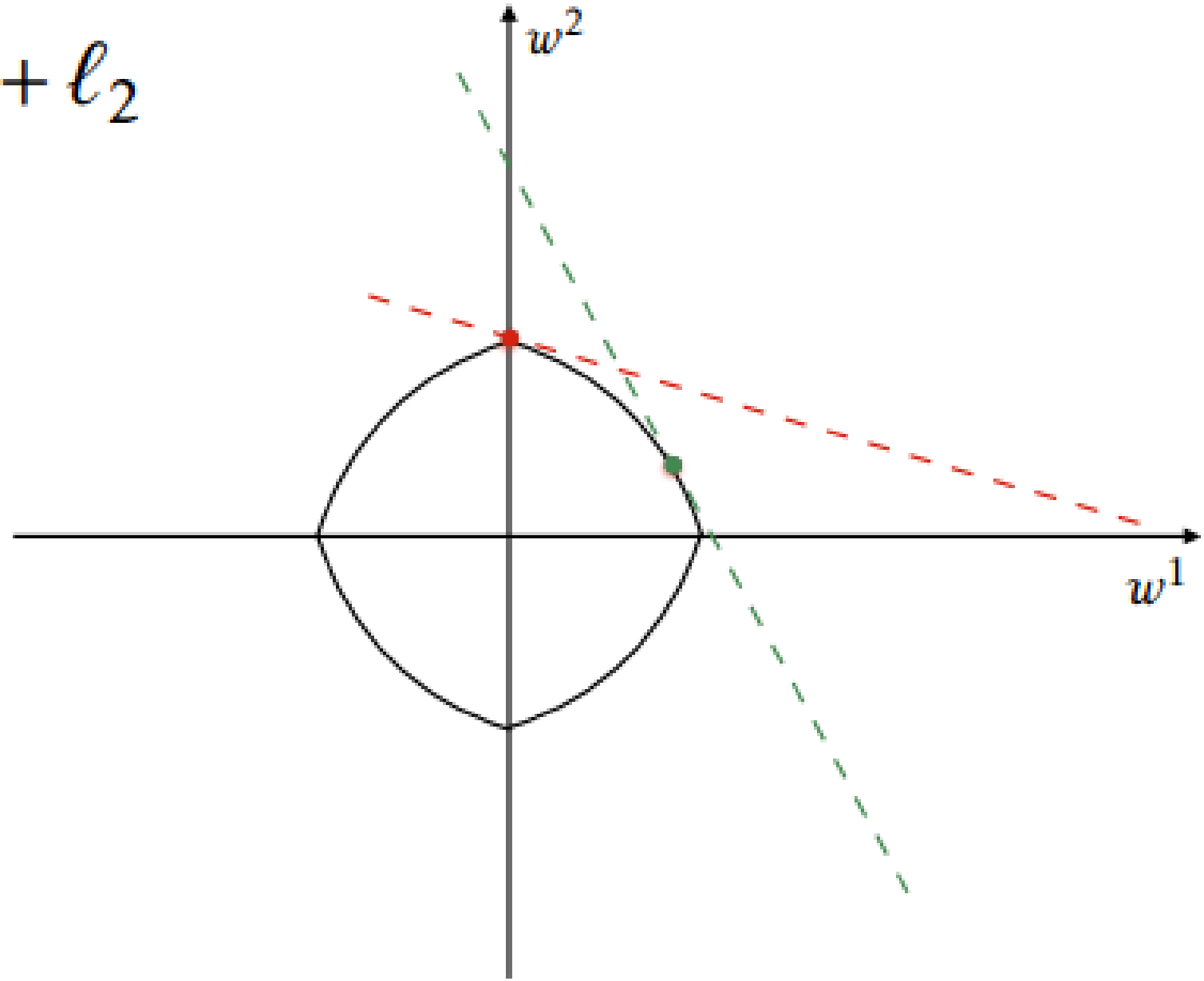
- Tries to pick individual features
- With a bunch of correlated features, what happens?

Understanding the L1 regularization

- Tries to pick individual features
- With a bunch of correlated features, what happens?
- Selects one
- This could harm prediction as a linear projection of the features might be more stable

Can we combine L1 and L2?

$\ell_1 + \ell_2$



[Wikipedia]

For $\alpha \in [0, 1]$

$$\mathscr{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} (\boldsymbol{\theta}^\top \mathbf{x}_i - y_i)^2 + \lambda(\alpha ||\boldsymbol{\theta}||_1 + (1-\alpha)||\boldsymbol{\theta}||_2^2)$$

Called the elastic net

- Used in many machine learning libraries
- Works well in practice

Recall

What if f is not linear

$$\min_f \mathbb{E}_{p(\mathbf{x},y)}[(y - f(\mathbf{x}))^2]$$

Recall

What if f is not linear

$$\begin{aligned} & \min_f \mathbb{E}_{p(\mathbf{x},y)}[(y - f(\mathbf{x}))^2] \\ = \quad & \min_f \mathbb{E}_{p(\mathbf{x})}\left[(\mathbb{E}[y \mid \mathbf{x}] - f(\mathbf{x}))^2\right] + \mathbb{E}_{p(\mathbf{x})}[\mathbb{V}\mathrm{ar}(y \mid \mathbf{x})] \end{aligned}$$

The optimal f is the conditional expectation

Where's overfitting? Where's the analysis loose?

Need to consider where the training data came from

$$\mathscr{D} = (x_1, y_1)...(x_n, y_n) \sim p$$

The training data set is a random variable!

The training data set is a random variable!

- Imagine wanting to predict heights from diet

The training data set is a random variable!

- Imagine wanting to predict heights from diet

- Step 1? Get Data

The training data set is a random variable!

- Imagine wanting to predict heights from diet
- Step 1? Get Data
- How? A person doing a random survey of people's heights and diet
  Gets $\mathscr{D}_1$
- Another doing a random survey of people's heights and diet
  Gets $\mathscr{D}_2$
- $\mathscr{D}_1$ and $\mathscr{D}_2$ are random because the full population was not surveyed
- $\mathscr{D}_1$ and $\mathscr{D}_2$ are drawn from the same distribution

Refine our Analysis

Include data randomness

$$\min_f \mathbb{E}_{p(\mathbf{x},y),\mathcal{D}\sim p}[(y - f(\mathbf{x};\mathcal{D}))^2]$$

Refine our Analysis

Include data randomness

$$
\begin{aligned}
& \min_f \mathbb{E}_{p(\mathbf{x},y),\mathscr{D}\sim p}[(y - f(\mathbf{x};\mathscr{D}))^2] \\
= \quad & \min_f \mathbb{E}_{p(\mathbf{x}),\mathscr{D}\sim p}\left[(\mathbb{E}[y\,|\,\mathbf{x}] - f(\mathbf{x};\mathscr{D}))^2\right] + \mathbb{E}_{p(\mathbf{x})}[\mathbb{V}\mathrm{ar}(y\,|\,\mathbf{x})]
\end{aligned}
$$

Refine our Analysis

Include data randomness

$$
\begin{aligned}
& \min_f \mathbb{E}_{p(\mathbf{x},y),\mathscr{D}\sim p}[(y - f(\mathbf{x};\mathscr{D}))^2] \\
= \; & \min_f \mathbb{E}_{p(\mathbf{x}),\mathscr{D}\sim p}\left[(\mathbb{E}[y\,|\,\mathbf{x}] - f(\mathbf{x};\mathscr{D}))^2\right] + \mathbb{E}_{p(\mathbf{x})}[\mathbb{V}\mathrm{ar}(y\,|\,\mathbf{x})] \\
= \; & \min_f \mathbb{E}_{p(\mathbf{x}),\mathscr{D}\sim p}\left[(\mathbb{E}[y\,|\,\mathbf{x}] - \mathbb{E}_{\mathscr{D}\sim p}[f(\mathbf{x};\mathscr{D})] + \mathbb{E}_{\mathscr{D}\sim p}[f(\mathbf{x};\mathscr{D})] - f(\mathbf{x};\mathscr{D}))^2\right] \\
& + \mathbb{E}_{p(\mathbf{x})}[\mathbb{V}\mathrm{ar}(y\,|\,\mathbf{x})]
\end{aligned}
$$

Refine our Analysis

Include data randomness

$$
\begin{aligned}
& \min_f \mathbb{E}_{p(\mathbf{x},y),\mathcal{D}\sim p}[(y - f(\mathbf{x};\mathcal{D}))^2] \\
= \; & \min_f \mathbb{E}_{p(\mathbf{x}),\mathcal{D}\sim p}\left[(\mathbb{E}[y\,|\,\mathbf{x}] - f(\mathbf{x};\mathcal{D}))^2\right] + \mathbb{E}_{p(\mathbf{x})}[\mathbb{V}\mathrm{ar}(y\,|\,\mathbf{x})] \\
= \; & \min_f \mathbb{E}_{p(\mathbf{x}),\mathcal{D}\sim p}\left[(\mathbb{E}[y\,|\,\mathbf{x}] - \mathbb{E}_{\mathcal{D}\sim p}[f(\mathbf{x};\mathcal{D})] + \mathbb{E}_{\mathcal{D}\sim p}[f(\mathbf{x};\mathcal{D})] - f(\mathbf{x};\mathcal{D}))^2\right] \\
& + \mathbb{E}_{p(\mathbf{x})}[\mathbb{V}\mathrm{ar}(y\,|\,\mathbf{x})] \\
= \; & \min_f \mathbb{E}_{p(\mathbf{x}),\mathcal{D}\sim p}\Big[(\mathbb{E}[y\,|\,\mathbf{x}] - \mathbb{E}_{\mathcal{D}\sim p}[f(\mathbf{x};\mathcal{D})])^2 \\
& + 2(\mathbb{E}[y\,|\,\mathbf{x}] - \mathbb{E}_{\mathcal{D}\sim p}[f(\mathbf{x};\mathcal{D})])(\mathbb{E}_{\mathcal{D}\sim p}[f(\mathbf{x};\mathcal{D})] - f(\mathbf{x};\mathcal{D})) \\
& + (\mathbb{E}_{\mathcal{D}\sim p}[f(\mathbf{x};\mathcal{D})] - f(\mathbf{x};\mathcal{D}))^2\Big] \\
& + \mathbb{E}_{p(\mathbf{x})}[\mathbb{V}\mathrm{ar}(y\,|\,\mathbf{x})]
\end{aligned}
$$

Refine our Analysis

Continuing

$$
\begin{aligned}
& \min_f \mathbb{E}_{p(\mathbf{x}),\mathscr{D}\sim p}[(y - f(\mathbf{x};\mathscr{D}))^2] \\
= \quad & \min_f \mathbb{E}_{p(\mathbf{x}),\mathscr{D}\sim p}\Big[(\mathbb{E}[y\,|\,\mathbf{x}] - \mathbb{E}_{\mathscr{D}\sim p}[f(\mathbf{x};\mathscr{D})])^2 \\
& + 2(\mathbb{E}[y\,|\,\mathbf{x}] - \mathbb{E}_{\mathscr{D}\sim p}[f(\mathbf{x};\mathscr{D})])(\mathbb{E}_{\mathscr{D}\sim p}[f(\mathbf{x};\mathscr{D})] - f(\mathbf{x};\mathscr{D})) \\
& + (\mathbb{E}_{\mathscr{D}\sim p}[f(\mathbf{x};\mathscr{D})] - f(\mathbf{x};\mathscr{D}))^2\Big] \\
& + \mathbb{E}_{p(\mathbf{x})}[\mathbb{V}\mathrm{ar}(y\,|\,\mathbf{x})]
\end{aligned}
$$

Refine our Analysis

Continuing

$$
\begin{aligned}
& \min_f \mathbb{E}_{p(\mathbf{x}),\mathscr{D}\sim p}[(y - f(\mathbf{x};\mathscr{D}))^2] \\
= \quad & \min_f \mathbb{E}_{p(\mathbf{x}),\mathscr{D}\sim p}\Big[(\mathbb{E}[y\,|\,\mathbf{x}] - \mathbb{E}_{\mathscr{D}\sim p}[f(\mathbf{x};\mathscr{D})])^2 \\
& +2(\mathbb{E}[y\,|\,\mathbf{x}] - \mathbb{E}_{\mathscr{D}\sim p}[f(\mathbf{x};\mathscr{D})])(\mathbb{E}_{\mathscr{D}\sim p}[f(\mathbf{x};\mathscr{D})] - f(\mathbf{x};\mathscr{D})) \\
& +(\mathbb{E}_{\mathscr{D}\sim p}[f(\mathbf{x};\mathscr{D})] - f(\mathbf{x};\mathscr{D}))^2\Big] \\
& +\mathbb{E}_{p(\mathbf{x})}[\mathbb{V}\mathrm{ar}(y\,|\,\mathbf{x})] \\
= \quad & \min_f \mathbb{E}_{p(\mathbf{x}),\mathscr{D}\sim p}\Big[(\mathbb{E}[y\,|\,\mathbf{x}] - \mathbb{E}_{\mathscr{D}\sim p}[f(\mathbf{x};\mathscr{D})])^2 \\
& +(\mathbb{E}_{\mathscr{D}\sim p}[f(\mathbf{x};\mathscr{D})] - f(\mathbf{x};\mathscr{D}))^2\Big] \\
& +\mathbb{E}_{p(\mathbf{x})}[\mathbb{V}\mathrm{ar}(y\,|\,\mathbf{x})]
\end{aligned}
$$

Refine our Analysis

$$\min_f \ \mathbb{E}_{p(\mathbf{x},y),\mathscr{D}\sim p}[(y - f(\mathbf{x};\mathscr{D}))^2]$$

$$= \min_f \ \mathbb{E}_{p(\mathbf{x}),\mathscr{D}\sim p}\Big[(\mathbb{E}[y\,|\,\mathbf{x}] - \mathbb{E}_{\mathscr{D}\sim p}[f(\mathbf{x};\mathscr{D})])^2 + (\mathbb{E}_{\mathscr{D}\sim p}[f(\mathbf{x};\mathscr{D})] - f(\mathbf{x};\mathscr{D}))^2\Big]$$

$$+ \mathbb{E}_{p(\mathbf{x},y)}[\mathbb{V}\mathrm{ar}(y\,|\,\mathbf{x})]$$

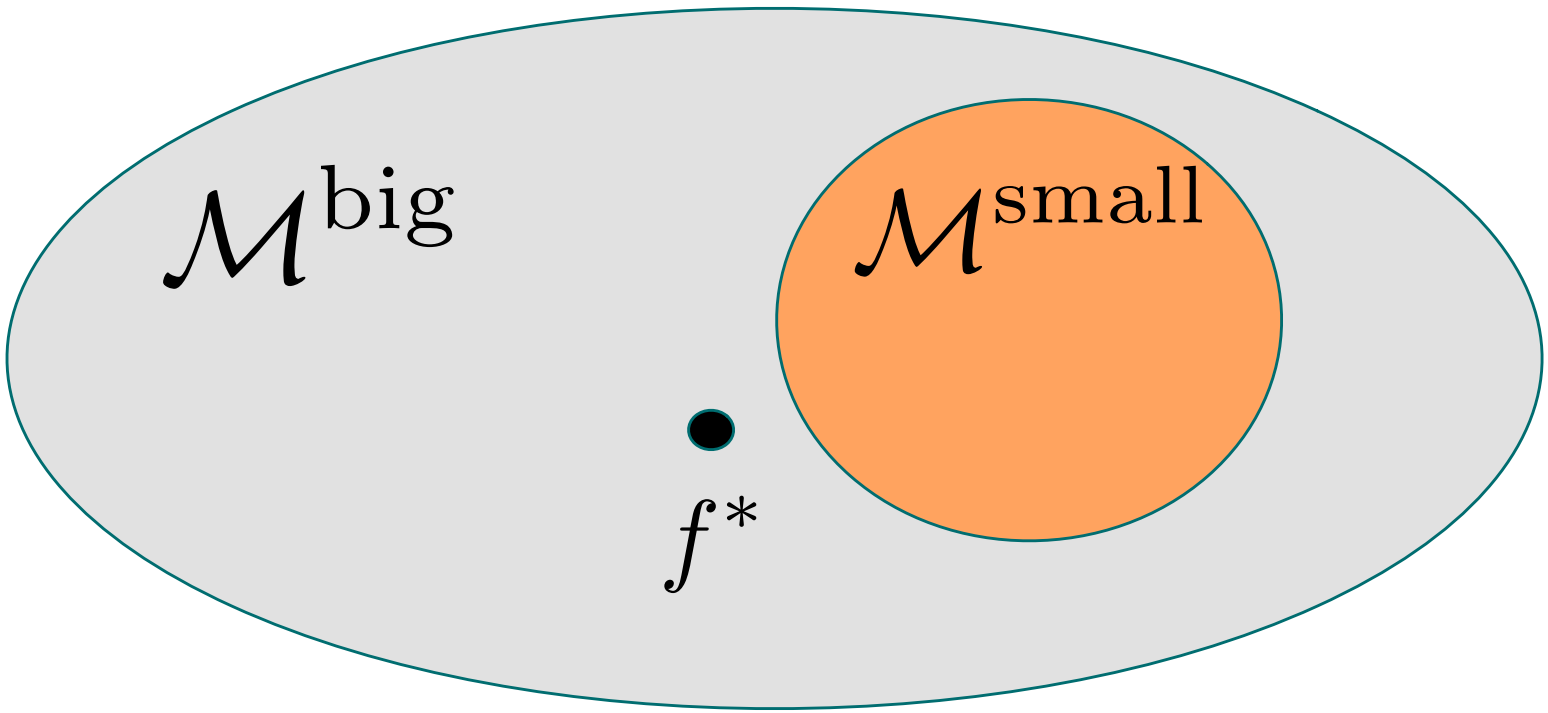
$$= \min_f \ \underbrace{\mathbb{E}_{p(\mathbf{x})}\Big[(\mathbb{E}[y\,|\,\mathbf{x}] - \mathbb{E}_{\mathscr{D}\sim p}[f(\mathbf{x};\mathscr{D})])^2\Big]}_{\text{Bias}}$$

$$+ \underbrace{\mathbb{E}_{\mathscr{D}\sim p}\Big[(\mathbb{E}_{\mathscr{D}\sim p}[f(\mathbf{x};\mathscr{D})] - f(\mathbf{x};\mathscr{D}))^2\Big]}_{\text{Variance}}$$

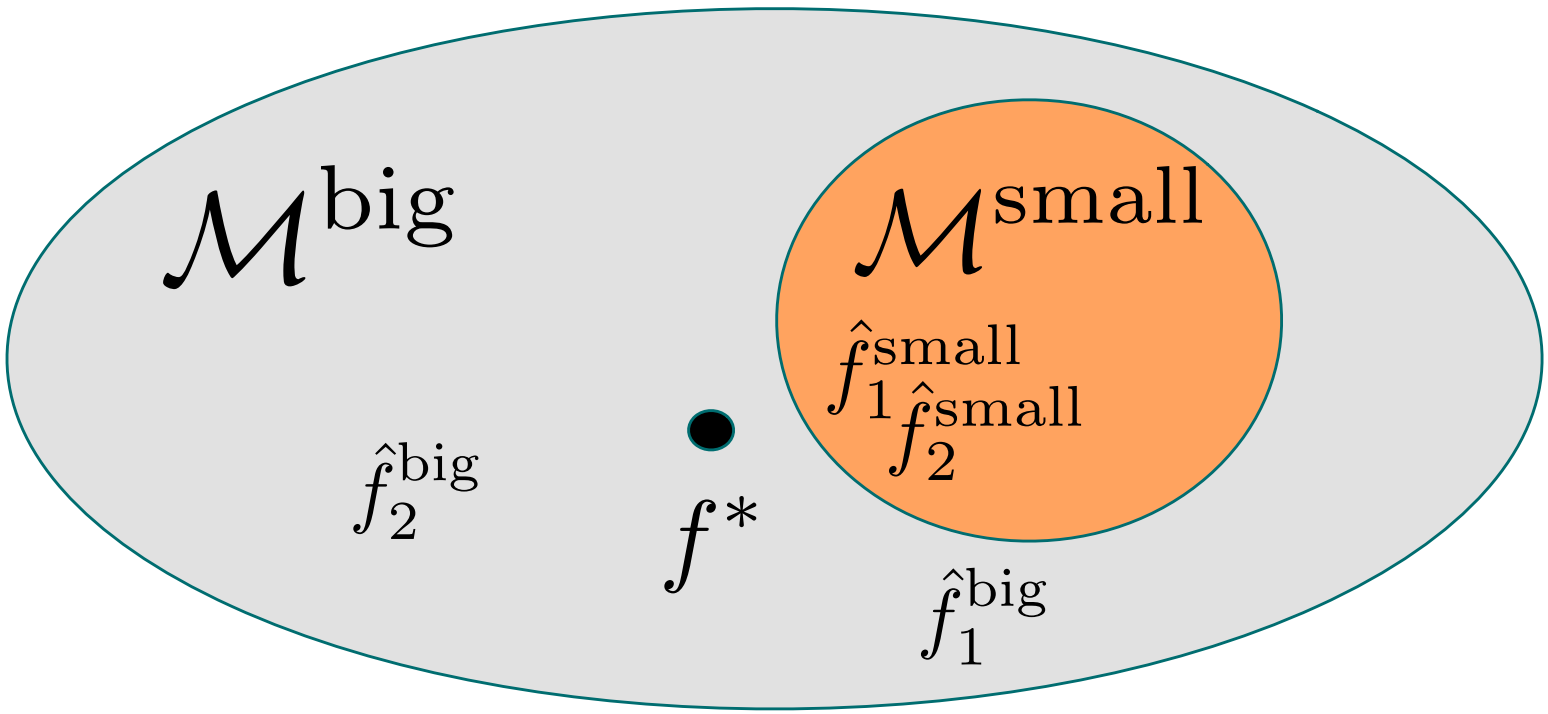
$$+ \underbrace{\mathbb{E}_{p(\mathbf{x})}[\mathbb{V}\mathrm{ar}(y\,|\,\mathbf{x})]}_{\text{Noise}}$$

$$\min_f \underbrace{\mathbb{E}_{p(\mathbf{x})}\Big[(\mathbb{E}[y\,|\,\mathbf{x}] - \mathbb{E}_{\mathscr{D}\sim p}[f(\mathbf{x};\mathscr{D}])^2\Big]}_{\text{Bias}}$$
$$+ \underbrace{\mathbb{E}_{\mathscr{D}\sim p}\Big[(\mathbb{E}_{\mathscr{D}\sim p}[f(\mathbf{x};\mathscr{D}] - f(\mathbf{x};\mathscr{D}))^2\Big]}_{\text{Variance}}$$
$$+ \underbrace{\mathbb{E}_{p(\mathbf{x})}[\mathbb{V}\mathrm{ar}(y\,|\,\mathbf{x})]}_{\text{Noise}}$$

- Bias accounts for model mismatch of average predictor
- Variance accounts for finite data
- Noise accounts for y not being a deterministic function of $\mathbf{x}$

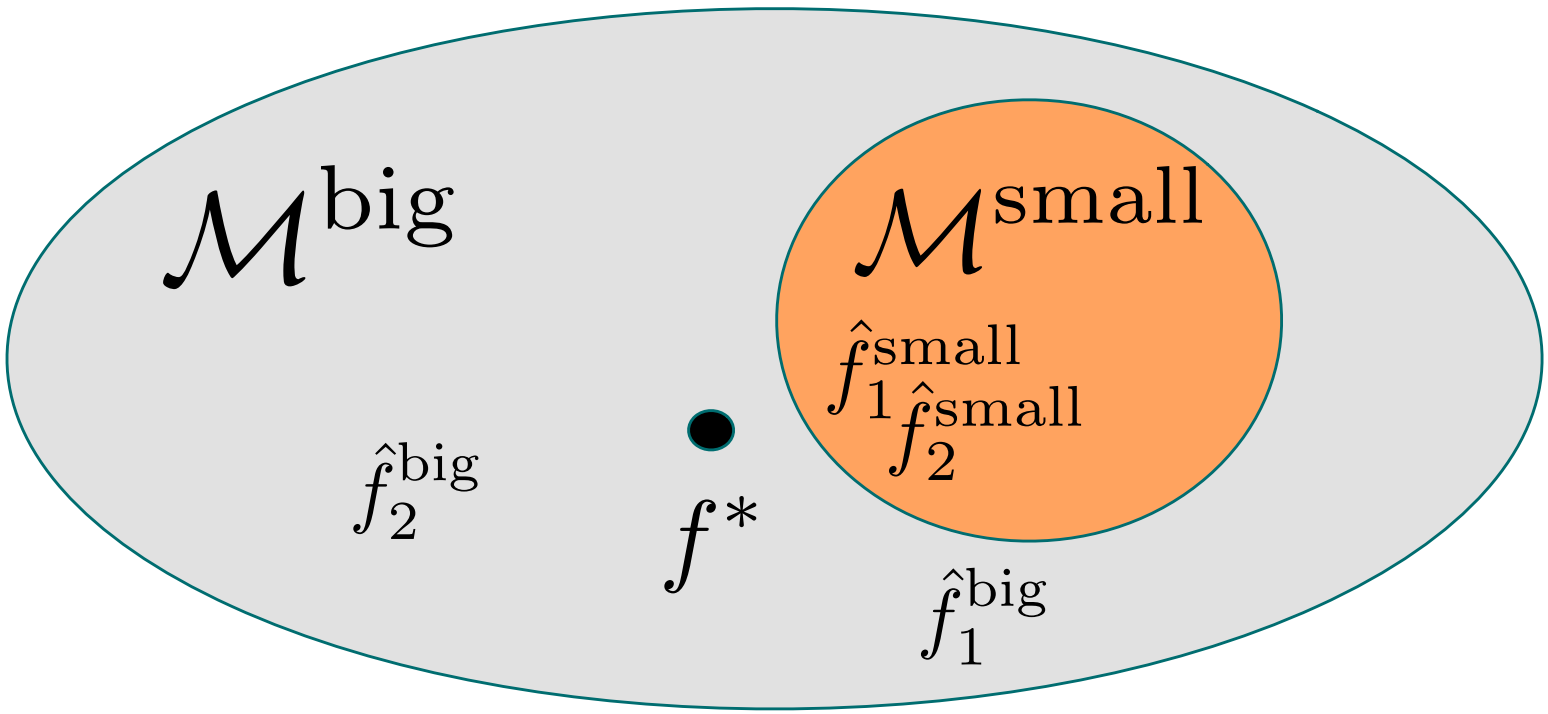

$\mathcal{M}^{\text{big}}$
$\mathcal{M}^{\text{small}}$
f^*

Get two house price data sets $\mathscr{D}^1$, $\mathscr{D}^2$



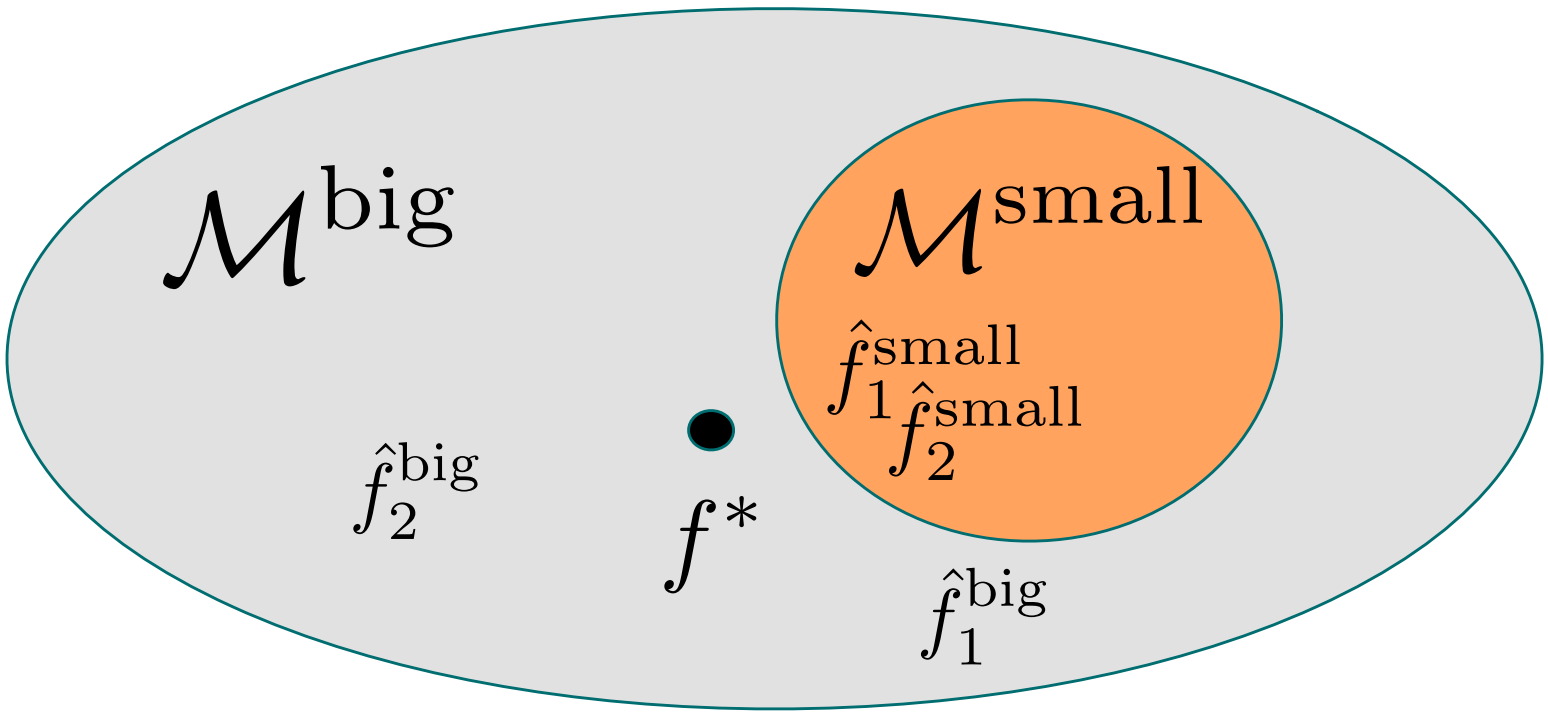
- Small models tend to have more bias
- Big models tend to have more variance

Get two house price data sets $\mathscr{D}^1$, $\mathscr{D}^2$



- Small models tend to underfit
- Big models tend to overfit

Get two house price data sets $\mathscr{D}^1$, $\mathscr{D}^2$



- Bias is like underfitting
- Variance is like overfitting

Changing Gears: A New Problem

Crohn's Disease

Goal: Understand Role of Genetics in Getting Crohn's Disease

Crohn's Disease

- Here the y for each person is $\{0, 1\}$
- 0 means they do not have Crohn's Disease
- 1 means they do have have Crohn's Disease

Called a *classification* problem

Maximum Likelihood

Principle of Maximum Likelihood

- Tries to maximize likelihood of the data
- Has some nice efficiency properties with growing data

Maximum Likelihood

Likeihood function measures “probability” of data

$$\ell(\mathbf{x}_i; \boldsymbol{\theta})$$

When data are independent, we get

$$\prod_i \ell(\mathbf{x}_i; \boldsymbol{\theta})$$

Generally work with logs for stability

$$\sum_i \log \ell(\mathbf{x}_i; \boldsymbol{\theta})$$

and stochastic optimization for speed

Maximum Likelihood

Suppose we flipped a biased coin many times and got samples

$$x_i \in 0, 1$$

Now we want to compute the maximum likelihood estimate for the probability of heads of the coin.

How do we start?

Maximum Likelihood

1. We need a distribution over coin flips?

Maximum Likelihood

1. We need a distribution over coin flips?

$$x_i \sim \text{Bernoulli}(p) = p^{x_i}(1-p)^{1-x_i}$$

Maximum Likelihood

1. We need a distribution over coin flips?

$$x_i \sim \text{Bernoulli}(p) = p^{x_i}(1-p)^{1-x_i}$$

   Why not another distribution?

Maximum Likelihood

1. We need a distribution over coin flips?

$$x_i \sim \text{Bernoulli}(p) = p^{x_i}(1-p)^{1-x_i}$$

   Why not another distribution?

2. Write down the log-likelihood of the observations

Maximum Likelihood

1. We need a distribution over coin flips?

$$x_i \sim \text{Bernoulli}(p) = p^{x_i}(1-p)^{1-x_i}$$

   Why not another distribution?

2. Write down the log-likelihood of the observations

$$\mathscr{L}(p) = \sum_{i=1}^{n} \log(p^{x_i}(1-p)^{1-x_i})$$

Maximum Likelihood

1. We need a distribution over coin flips?

$$x_i \sim \text{Bernoulli}(p) = p^{x_i}(1-p)^{1-x_i}$$

   Why not another distribution?

2. Write down the log-likelihood of the observations

$$\mathscr{L}(p) = \sum_{i=1}^{n} \log(p^{x_i}(1-p)^{1-x_i})$$

3. Maximize by taking derivatives and setting to zero

Maximum Likelihood

1. We need a distribution over coin flips?

$$x_i \sim \text{Bernoulli}(p) = p^{x_i}(1-p)^{1-x_i}$$

   Why not another distribution?

2. Write down the log-likelihood of the observations

$$\mathcal{L}(p) = \sum_{i=1}^{n} \log(p^{x_i}(1-p)^{1-x_i})$$

3. Maximize by taking derivatives and setting to zero

What happens when you only observe zeros?

Maximum Likelihood

1. Need a distribution over each observed data point
2. Write down the log-likelihood of the observations
3. Maximize the log-likelihood (by gradients)

ML for Binary Outcome Linear Regression

1. Need a distribution over each observed data point

ML for Binary Outcome Linear Regression

1. Need a distribution over each observed data point

$$p(y_i = 1 \,|\, \mathbf{x}_i; \boldsymbol{\theta}) = \frac{1}{1 + \exp(-\boldsymbol{\theta}^\top \mathbf{x}_i)} := \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)$$

   σ is the sigmoid or logistic function maps from $\mathbb{R}$ to $(0, 1)$

ML for Binary Outcome Linear Regression

1. Need a distribution over each observed data point

$$p(y_i = 1 \,|\, \mathbf{x}_i; \boldsymbol{\theta}) = \frac{1}{1 + \exp(-\boldsymbol{\theta}^\top \mathbf{x}_i)} := \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)$$

   σ is the sigmoid or logistic function maps from $\mathbb{R}$ to $(0, 1)$

2. Write down log-likelihood of the n observations

ML for Binary Outcome Linear Regression

1. Need a distribution over each observed data point

$$p(y_i = 1 \,|\, \mathbf{x}_i; \boldsymbol{\theta}) = \frac{1}{1 + \exp(-\boldsymbol{\theta}^\top \mathbf{x}_i)} := \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)$$

   σ is the sigmoid or logistic function maps from $\mathbb{R}$ to $(0, 1)$

2. Write down log-likelihood of the n observations

$$\sum_{i=1}^{n} \log(p(y_i \,|\, \mathbf{x}_i; \boldsymbol{\theta})) = \sum_{i=1}^{n} y_i \log \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i) + (1 - y_i) \log(1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i))$$

ML for Binary Outcome Linear Regression

1. Need a distribution over each observed data point

$$p(y_i = 1 \,|\, \mathbf{x}_i; \boldsymbol{\theta}) = \frac{1}{1 + \exp(-\boldsymbol{\theta}^\top \mathbf{x}_i)} := \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)$$

   σ is the sigmoid or logistic function maps from $\mathbb{R}$ to $(0, 1)$

2. Write down log-likelihood of the n observations

$$\sum_{i=1}^{n} \log(p(y_i \,|\, \mathbf{x}_i; \boldsymbol{\theta})) = \sum_{i=1}^{n} y_i \log \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i) + (1 - y_i) \log(1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i))$$

3. Maximize with respect to $\boldsymbol{\theta}$

Called logistic regression

Maximize

$$\nabla_{\boldsymbol{\theta}} \mathcal{L} = \nabla_{\boldsymbol{\theta}} \sum_{i=1}^{n} y_i \log \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i) + (1 - y_i) \log(1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i))$$

$$= \sum_{i=1}^{n} y_i \frac{\sigma'(\boldsymbol{\theta}^\top \mathbf{x}_i)\mathbf{x}_i}{\sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)} + (y_i - 1) \frac{\sigma'(\boldsymbol{\theta}^\top \mathbf{x}_i)\mathbf{x}_i}{1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)}$$

Maximize

$$\nabla_{\boldsymbol{\theta}} \mathscr{L} = \nabla_{\boldsymbol{\theta}} \sum_{i=1}^{n} y_i \log \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i) + (1 - y_i) \log(1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i))$$

$$= \sum_{i=1}^{n} y_i \frac{\sigma'(\boldsymbol{\theta}^\top \mathbf{x}_i)\mathbf{x}_i}{\sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)} + (y_i - 1) \frac{\sigma'(\boldsymbol{\theta}^\top \mathbf{x}_i)\mathbf{x}_i}{1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)}$$

Use

$$\sigma'(a) = \sigma(a)(1 - \sigma(a))$$

Maximize

$$\nabla_{\boldsymbol{\theta}} \mathscr{L} = \nabla_{\boldsymbol{\theta}} \sum_{i=1}^{n} y_i \log \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i) + (1 - y_i) \log(1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i))$$

$$= \sum_{i=1}^{n} y_i \frac{\sigma'(\boldsymbol{\theta}^\top \mathbf{x}_i)\mathbf{x}_i}{\sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)} + (y_i - 1) \frac{\sigma'(\boldsymbol{\theta}^\top \mathbf{x}_i)\mathbf{x}_i}{1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)}$$

Use

$$\sigma'(a) = \sigma(a)(1 - \sigma(a))$$

Then

$$\nabla_{\boldsymbol{\theta}} \mathscr{L} = \sum_{i=1}^{n} \mathbf{x}_i (y_i - y_i \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i) + (y_i - 1)\sigma(\boldsymbol{\theta}^\top \mathbf{x}_i))$$

$$= \sum_{i=1}^{n} \mathbf{x}_i (y_i - \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i))$$

Maximize

$$\nabla_{\boldsymbol{\theta}} \mathscr{L} = \nabla_{\boldsymbol{\theta}} \sum_{i=1}^{n} y_i \log \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i) + (1 - y_i) \log(1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i))$$

$$= \sum_{i=1}^{n} y_i \frac{\sigma'(\boldsymbol{\theta}^\top \mathbf{x}_i)\mathbf{x}_i}{\sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)} + (y_i - 1) \frac{\sigma'(\boldsymbol{\theta}^\top \mathbf{x}_i)\mathbf{x}_i}{1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)}$$

Use

$$\sigma'(a) = \sigma(a)(1 - \sigma(a))$$

Then

$$\nabla_{\boldsymbol{\theta}} \mathscr{L} = \sum_{i=1}^{n} \mathbf{x}_i (y_i - y_i \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i) + (y_i - 1)\sigma(\boldsymbol{\theta}^\top \mathbf{x}_i))$$

$$= \sum_{i=1}^{n} \mathbf{x}_i (y_i - \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i))$$

Can we set it equal to zero?

$$\nabla_{\boldsymbol{\theta}} \mathcal{L} = \sum_{i=1}^{n} \mathbf{x}_i (y_i - y_i \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i) + (y_i - 1)\sigma(\boldsymbol{\theta}^\top \mathbf{x}_i))$$

$$= \sum_{i=1}^{n} \mathbf{x}_i (y_i - \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i))$$

Use gradient or stochastic gradient ascent to maximize.

Is there a more intuitive form?

$$\nabla_{\boldsymbol{\theta}} \mathcal{L} = \sum_{i=1}^{n} \mathbf{x}_i (y_i - y_i \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i) + (y_i - 1)\sigma(\boldsymbol{\theta}^\top \mathbf{x}_i))$$
$$= \sum_{i=1}^{n} \mathbf{x}_i (y_i - \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i))$$

Use gradient or stochastic gradient ascent to maximize.

Is there a more intuitive form?

$$\nabla_{\boldsymbol{\theta}} \mathcal{L} = \sum_{i=1}^{n} \mathbf{x}_i (y_i - E_{\boldsymbol{\theta}}[y_i \,|\, \mathbf{x}_i])$$

Most gradients can seen as balancing various terms

Understanding Logistic Regression

We could have used other functions to map real $\mathbf{x}^\top \boldsymbol{\theta}$

- CDF of Gaussian maps from reals to (0, 1)
- CDF of Student-T maps similarly

Why use logistic regression?

- All machine learning methods use assumptions to generalize
- Important to understand those assumptions

Log odds. Start with logit function

$$logit(p) = \log\left(\frac{p}{1-p}\right)$$

Compute logit of $p(y = 1 \mid \mathbf{x})$

$$logit(p(y = 1 \mid \mathbf{x})) = \log\left(\frac{\frac{1}{1+\exp(-\boldsymbol{\theta}^\top \mathbf{x}_i)}}{\frac{\exp(-\boldsymbol{\theta}^\top \mathbf{x}_i)}{1+\exp(-\boldsymbol{\theta}^\top \mathbf{x}_i)}}\right) = -\log(\exp(-\boldsymbol{\theta}^\top \mathbf{x})) = \boldsymbol{\theta}^\top \mathbf{x}$$

Logistic regression is linear in the logit of probabilites

Measure under which importance should be reported

Probabilistic Assumptions

Probabilistic Assumptions

Start with the logistic distribution

$$p(a) = \frac{1}{4}\text{sech}^2\left(\frac{x}{2}\right) = \frac{1}{(\exp(x/2) + \exp(-x/2))^2}$$

Logistic regression has a noise model

$$\epsilon_i \sim \text{logistic}$$
$$y_i = \boldsymbol{\theta}^\top \mathbf{x}_i + \epsilon_i > 0$$

- Noise comes from missing features under correct model
- Noise is logistic

Probabilistic Assumptions

Start with the logistic distribution

$$p(a) = \frac{1}{4}\operatorname{sech}^2\left(\frac{x}{2}\right) = \frac{1}{(\exp(x/2) + \exp(-x/2))^2}$$

Logistic regression has a noise model

$$\epsilon_i \sim \text{logistic}$$
$$y_i = \boldsymbol{\theta}^\top \mathbf{x}_i + \epsilon_i > 0$$

- Can use lots of distributions for noise
- Normal noise yields *probit model*. Popular in economics

Probabilistic Assumptions

Start with the logistic distribution

$$p(a) = \frac{1}{4}\operatorname{sech}^2\left(\frac{x}{2}\right) = \frac{1}{(\exp(x/2) + \exp(-x/2))^2}$$

Logistic regression has a noise model

$$\epsilon_i \sim \text{logistic}$$
$$y_i = \boldsymbol{\theta}^\top \mathbf{x}_i + \epsilon_i > 0$$

- Has heaver tails than the normal distribution
- Fits “outliers” less

Regularization

If $p > n$, can overfit like linear regression. Can regularize!

Regularization

If $p > n$, can overfit like linear regression. Can regularize!

$$\sum_{i=1}^{n} y_i \log \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i) + (1 - y_i) \log(1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)) - \lambda ||\boldsymbol{\theta}||_2^2$$

Regularization

If $p > n$, can overfit like linear regression. Can regularize!

$$\sum_{i=1}^{n} y_i \log \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i) + (1 - y_i) \log(1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)) - \lambda ||\boldsymbol{\theta}||_1$$

Does L1 give sparsity here?

Crohn's Disease

Goal: Understand Role of Genetics in the Number of Crohn's Related Doctor Visits

- $\mathbf{x}_i$ are features
- y_i are now counts

Do we have to derive a new model again for this?

- $\mathbf{x}_i$ are features
- y_i are now counts

Do we have to derive a new model again for this?

We need a distribution for $p(y \mid \mathbf{x})$. Can use *exponential families*

Exponential Families

Exponential families contain many popular distributions

- Bernoulli: Binary values
- Normal: Real values
- Gamma: Positive values
- Categorical: Multiple types

Exponential Families

Exponential family distribution

$$p(\mathbf{x}) = h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x}) - a(\boldsymbol{\eta}))$$

- h: Base measure
- $\boldsymbol{\eta}$: Parameters
- $t(\mathbf{x})$: Sufficient statistics
- $a(\boldsymbol{\eta})$: log-normalizer

$$a(\boldsymbol{\eta}) = \log \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}$$

Important property of exponential family

$$\nabla a(\boldsymbol{\eta}) \quad = \quad \nabla \log \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}$$

Important property of exponential family

$$\begin{aligned}\nabla a(\boldsymbol{\eta}) &= \nabla \log \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x} \\ &= \frac{\nabla_{\boldsymbol{\eta}} \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}}{\int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}}\end{aligned}$$

Important property of exponential family

$$
\begin{aligned}
\nabla a(\boldsymbol{\eta}) &= \nabla \log \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x} \\
&= \frac{\nabla_{\boldsymbol{\eta}} \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}}{\int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}} \\
&= \frac{\int h(\mathbf{x}) \nabla_{\boldsymbol{\eta}} \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}}{\int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}}
\end{aligned}
$$

Important property of exponential family

$$
\begin{aligned}
\nabla a(\boldsymbol{\eta}) &= \nabla \log \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x} \\
&= \frac{\nabla_{\boldsymbol{\eta}} \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}}{\int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}} \\
&= \frac{\int h(\mathbf{x}) \nabla_{\boldsymbol{\eta}} \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}}{\int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}} \\
&= \frac{\int h(\mathbf{x}) t(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}}{\exp(\log(\int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}))}
\end{aligned}
$$

Important property of exponential family

$$
\begin{aligned}
\nabla a(\boldsymbol{\eta}) &= \nabla \log \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x} \\
&= \frac{\nabla_{\boldsymbol{\eta}} \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}}{\int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}} \\
&= \frac{\int h(\mathbf{x}) \nabla_{\boldsymbol{\eta}} \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}}{\int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}} \\
&= \frac{\int h(\mathbf{x}) t(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}}{\exp(\log(\int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}))} \\
&= \frac{\int h(\mathbf{x}) t(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}}{\exp(a(\boldsymbol{\eta}))}
\end{aligned}
$$

Important property of exponential family

$$
\begin{aligned}
\nabla a(\boldsymbol{\eta}) &= \nabla \log \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x} \\
&= \frac{\nabla_{\boldsymbol{\eta}} \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}}{\int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}} \\
&= \frac{\int h(\mathbf{x}) \nabla_{\boldsymbol{\eta}} \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}}{\int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}} \\
&= \frac{\int h(\mathbf{x}) t(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}}{\exp(\log(\int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}))} \\
&= \frac{\int h(\mathbf{x}) t(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}}{\exp(a(\boldsymbol{\eta}))} \\
&= \int h(\mathbf{x}) t(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x}) - a(\boldsymbol{\eta})) d\mathbf{x}
\end{aligned}
$$

Important property of exponential family

$$
\begin{aligned}
\nabla a(\boldsymbol{\eta}) &= \nabla \log \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x} \\
&= \frac{\nabla_{\boldsymbol{\eta}} \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}}{\int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}} \\
&= \frac{\int h(\mathbf{x}) \nabla_{\boldsymbol{\eta}} \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}}{\int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}} \\
&= \frac{\int h(\mathbf{x}) t(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}}{\exp(\log(\int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}))} \\
&= \frac{\int h(\mathbf{x}) t(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x})) d\mathbf{x}}{\exp(a(\boldsymbol{\eta}))} \\
&= \int h(\mathbf{x}) t(\mathbf{x}) \exp(\boldsymbol{\eta}^\top t(\mathbf{x}) - a(\boldsymbol{\eta})) d\mathbf{x} \\
&= \mathbb{E}[t(\mathbf{x})]
\end{aligned}
$$

Build regression models with exponential families

Build regression models with exponential families

- Use exponential families for conditional distribution

$$p(y_i \mid \mathbf{x}_i) \sim \text{Expfam}$$

- Maximum likelihood to estimate parameters

More basic. How do you design conditional distributions?

$$p(y \mid \mathbf{x}) \sim \text{Normal}(\mu, \sigma)$$

More basic. How do you design conditional distributions?

$$p(y \mid \mathbf{x}) \sim \text{Normal}(\mu, \sigma)$$

Mean and variance functions of the conditioning variable

$$\mu(\mathbf{x}_i) = f_{\boldsymbol{\theta}}(\mathbf{x}_i)$$
$$\sigma(\mathbf{x}_i) = \log(1 + \exp(g_{\boldsymbol{\theta}}(\mathbf{x}_i)))$$

Generalized Linear Models

$$p(y_i \mid \mathbf{x}_i) \sim \text{Expfam}$$

Generalized Linear Models

$$p(y_i \mid \mathbf{x}_i) \sim \text{Expfam}$$

The parameter is the natural parameter. One option

$$\boldsymbol{\eta}_i = \boldsymbol{\theta}^\top \mathbf{x}_i$$

Generalized Linear Models

$$p(y_i \mid \mathbf{x}_i) \sim \text{Expfam}$$

The parameter is the natural parameter. One option

$$\boldsymbol{\eta}_i = \boldsymbol{\theta}^\top \mathbf{x}_i$$

Learn parameters by maximizing likelihood

$$\mathscr{L}(\boldsymbol{\theta}) = \sum_i \log(h(y_i) \exp(\boldsymbol{\eta}_i^\top t(y_i) - a(\boldsymbol{\eta}_i)))$$

Generalized Linear Models

$$p(y_i \mid \mathbf{x}_i) \sim \text{Expfam}$$

The parameter is the natural parameter. One option

$$\boldsymbol{\eta}_i = \boldsymbol{\theta}^\top \mathbf{x}_i$$

Learn parameters by maximizing likelihood

$$\mathscr{L}(\boldsymbol{\theta}) = \sum_i \log(h(y_i) \exp(\boldsymbol{\eta}_i^\top t(y_i) - a(\boldsymbol{\eta}_i)))$$

Parameters can be generalized with link functions

What happens with a Normal with fixed variance $\sigma = 1$?

What happens with a Normal with fixed variance $\sigma = 1$?

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

- h_σ

$$\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right)$$

- $t(x) = x$
- $a(\mu) = \frac{\mu^2}{2}$
- $\eta = \mu$

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

- $\eta = \mu$

Use $\eta_i = \theta^\top \mathbf{x}_i$. What is this?

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

- $\eta = \mu$

Use $\eta_i = \theta^\top \mathbf{x}_i$. What is this?

Generalized linear model with Normal is linear regression

What happens with a Bernoulli?

What happens with a Bernoulli?

$$p(x) = \exp(\eta x - \log(1 + \exp(\eta)))$$

What happens with a Bernoulli?

Set $\boldsymbol{\eta}_i = \boldsymbol{\theta}^\top \mathbf{x}_i$

$$p(y_i \,|\, \mathbf{x}_i) = \exp(\boldsymbol{\theta}^\top \mathbf{x}_i y_i - \log(1 + \exp(\boldsymbol{\theta}^\top \mathbf{x}_i)))$$

What happens with a Bernoulli?

Set $\boldsymbol{\eta}_i = \boldsymbol{\theta}^\top \mathbf{x}_i$

$$p(y_i \,|\, \mathbf{x}_i) = \exp(\boldsymbol{\theta}^\top \mathbf{x}_i y_i - \log(1 + \exp(\boldsymbol{\theta}^\top \mathbf{x}_i)))$$

Compute $\mathbb{E}[y_i \,|\, \mathbf{x}_i] = p(y \,|\, \mathbf{x}_i)$

$$\nabla_{\boldsymbol{\eta}} \log(1 + \exp(\boldsymbol{\theta}^\top \mathbf{x}_i)) = \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)$$

It's logistic regression

Regularization

Learn parameters by maximizing likelihood

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_i \log(h(y_i) \exp(\boldsymbol{\eta}_i^\top t(y_i) - a(\boldsymbol{\eta}_i)))$$

- Can regularize generalized linear models
- Use $L2$, $L1$, or elastic net

Complex models can be regularized by parameter size

L1 regularization finds sparse solutions

Logistic regression is for binary outcomes

Linear and logistic regression are examples of GLMs

GLMs plus regularization allow regression on all kinds of data