Lecture #2 | Part 4 | Regularization, Ridge Regression

Regularization helps avoid Overfitting!

Readings: Murphy 11.3-4

code: ridge_and_lasso.ipynb

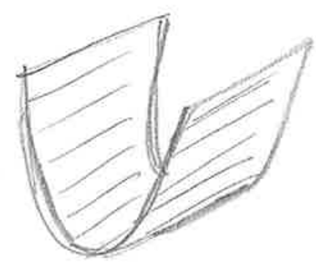
→ Regularization in Linear Regression

- Recall Least Squares: $\hat{\omega}_{LS} = \arg\min_{\omega} \sum_{i=1}^{n} (y_i - x_i^T \omega)^2$

$$= \arg\min_{\omega} (y - X\omega)^T (y - X\omega)$$

$$= (X^T X)^{-1} X^T y \quad \left(\text{when } (X^T X)^{-1} \text{ exists!}\right)$$

- What if $x_i \in \mathbb{R}^d$ and $d > n$?
  - The objective function is flat in some directions
  - Implies optimal solution is not unique and unstable due to curvature:
    - Small changes in training data result in large changes in solution
    - often the magnitudes of $\boxed{\omega}$ are "very large"



= Regularization imposes "simpler" solutions by a "complexity" penalty

→ Sensitivity increases overfitting

- For a linear model:

$$y \approx b + w_1 x_1 + w_2 x_2 + \dots + w_d x_d$$

<u>if</u> $|w_j|$ is large, then the prediction is $\boxed{\text{sensitive}}$ to small changes in $\boxed{x_j}$.

- Large sensitivity leads to overfitting and poor generalization, and equivalently, models that overfit tend to have large weights.
- Note: $\boxed{b}$ is a constant and hence there is no sensitivity for the offset b. Never regularize b

- In **Ridge Regression**, we use regularizer $\|w\|_2^2$ to measure and control the $\boxed{\text{sensitivity}}$ of the predictor.
  - And, optimize for <u>small loss</u> and <u>small sensitivity</u>, by adding a <u>regularizer</u> in the objective (assume no offset for now)

$$\boxed{\hat{w}_{ridge} = \arg\min_{w} \Big\{ \sum_{i=1}^{n} (y_i - x_i^T w)^2 + \underbrace{\lambda}_{\text{regularization coefficient}} \|w\|_2^2 \Big\}}$$

Compared to old Least Squares objective:

$$\boxed{\hat{w}_{LS} = \arg\min_{w} \sum_{i=1}^{n} (y_i - x_i^T w)^2}$$

→ Minimizing the Ridge Regression Objective

$X \in n \times d$
$w \in d \times 1$

$$\hat{\omega}_{ridge} = \arg\min_{w} \sum_{i=1}^{n} (y_i - x_i^T w)^2 + \lambda \|w\|_2^2 \quad L2 \text{ norm}$$

$$\|w\|_p \triangleq (|w_1|^p + \dots + |w_d|^p)^{\frac{1}{p}}$$

$$= \|y - Xw\|^2 + \lambda \|w\|^2$$

$$= (Xw - y)^T (Xw - y) + \lambda w^T w$$

$$\nabla_w f = \cancel{2} X^T (Xw - y) + \cancel{2} \lambda w = 0$$

$$(X^T X + \lambda I) w = X^T y$$

Similar to how we did in simple Linear Regression

$$\boxed{\hat{\omega}_{RIDGE} = (X^T X + \lambda I)^{-1} \cdot X^T \cdot y}$$

$$I^d = \begin{bmatrix} 1 & 0 & & \\ 0 & 1 & & \\ \vdots & & \ddots & \\ 0 & & & 1 \end{bmatrix} \in d \times d$$

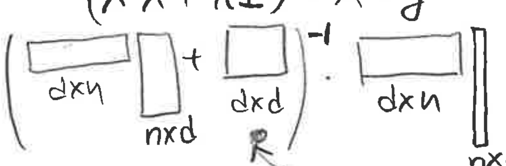
Under

Scalar Derivative		Vector Derivative	
$f(x) \rightarrow$	$\frac{df}{dx}$	$f(x) \rightarrow$	$\frac{df}{dx}$
$bx \rightarrow$	b	$x^T B \rightarrow$	B
$bx \rightarrow$	b	$x^T b \rightarrow$	b
$x^2 \rightarrow$	$2x$	$x^T x \rightarrow$	$2x$
$bx^2 \rightarrow$	$2bx$	$x^T B x \rightarrow$	$2Bx$

→ Shrinkage Properties

$$\hat{\omega}_{ridge} = \arg\min_{w} \sum_{i=1}^{n} (y_i - x_i^T w)^2 + \lambda \|w\|_2^2$$

$$= (X^T X + \lambda I)^{-1} \cdot X^T \cdot y$$

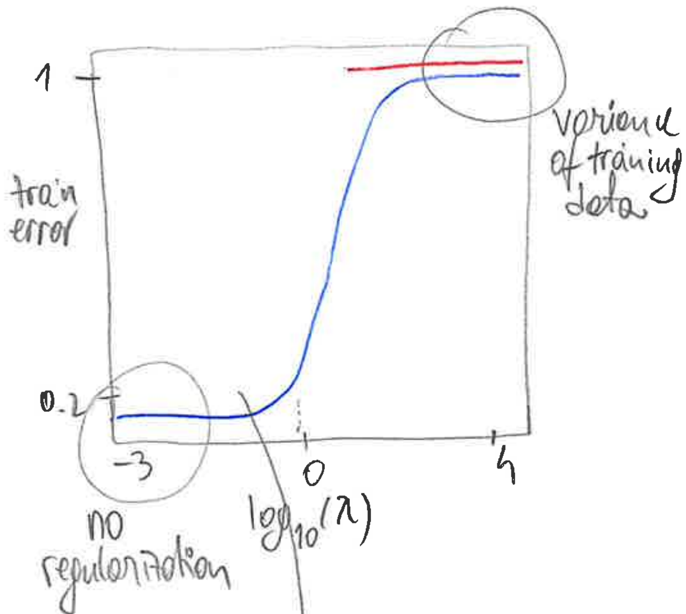


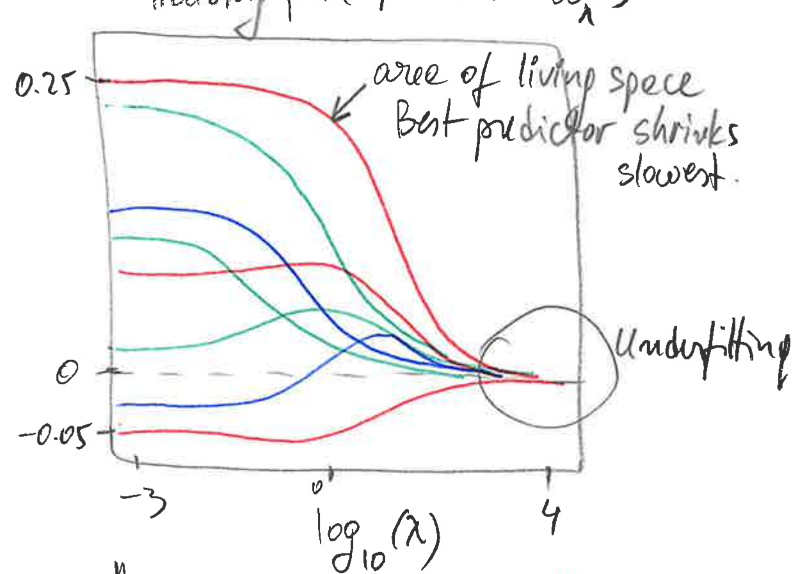
Larger λ "divides by" larger term! (makes w smaller)

- when $\lambda = 0 \Rightarrow$ least squares model
- this defines a family of models parameterized by $\boxed{\lambda}$
- Large λ means more regularization and simpler model
- Small λ means less regularization and more complex model.

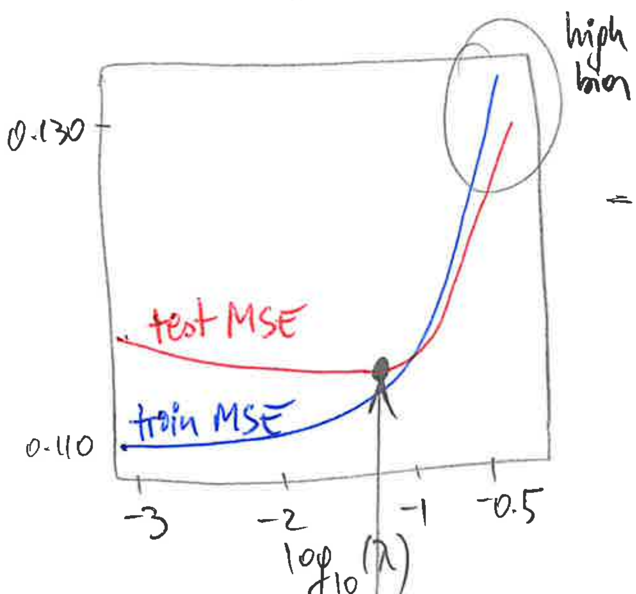
Ridge Regression : minimize $\sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|_2^2$

Training MSE $\frac{1}{n}\sum_{i=1}^{n} (y_i - x_i^T \hat{w}^{(\lambda)}_{ridge})^2$





"Regularization path" plot
(See Murphy pp. 383)



- Gain in test MSE comes from shrinking w's to get a less sensitive predictor;
- which in turn reduces the variance
- This is the role of regularizer: reduce sensitivity/variance!

Bias/Variance trade-off

We are not changing the model class, we are just changing $\boxed{\lambda}$.

Bias-Variance Properties

- $\hat{\omega}_{ridge} = (X^T X + \lambda I)^{-1} \cdot X^T \cdot y$
- To analyze bias-variance tradeoff, we need to assume probabilistic generative model:

$$x_i \sim P_X, \quad y = X\omega + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \tau^2 I)$$

for some ground truth model parameter $\boxed{\omega}$

- The true error at a sample with feature x is:

$$\mathbb{E}_{y, \mathcal{D}_{train} | x}\big[(y - \underbrace{x^T}_{\text{test sample}} \underbrace{\hat{\omega}_{ridge}}_{\text{fit to a given set } X, y})^2 \mid x\big] =$$

$\eta(x)$ same as before

$$= \underbrace{\mathbb{E}_{y|x}\big[(y - \mathbb{E}[y|x])^2 \mid x\big]}_{\text{irreducible error}} + \underbrace{\mathbb{E}_{\mathcal{D}_{train}}\big[(\mathbb{E}[y|x] - x^T \hat{\omega}_{ridge})^2 \mid x\big]}_{\text{Learning error}}$$

$$= \mathbb{E}_{y|x}\big[(y - x^T\omega)^2 \mid x\big] + \mathbb{E}_{\mathcal{D}_{train}}\big[(x^T \underbrace{\omega}_{\text{ground truth}} - x^T \hat{\omega}_{ridge})^2 \mid x\big]$$

$$= \underbrace{\tau^2}_{\text{irreducible error}} + \underbrace{\left(x^T\omega - \mathbb{E}_{\mathcal{D}_{train}}[x^T\hat{\omega}_{ridge} | x]\right)^2}_{\text{Bias-squared}} + \underbrace{\mathbb{E}_{\mathcal{D}_{train}}\left[\left(\mathbb{E}_{\mathcal{D}_{train}}[x^T\hat{\omega}_{ridge} | x] - x^T\hat{\omega}_{ridge}\right)^2 | x\right]}_{\text{Variance}}$$

= ...

Sample data from Independent Gaussians

Suppose $X^T X = nI$, then $\hat{\omega}_{ridge} = (X^T X + \lambda I)^{-1} \cdot X^T (X\omega + \varepsilon)$

$$= \frac{n}{n+\lambda} \underbrace{\omega}_{\text{ground truth}} + \frac{1}{n+\lambda} X^T \varepsilon$$

λ trades off weight vs. noise.

HW assignment to verify

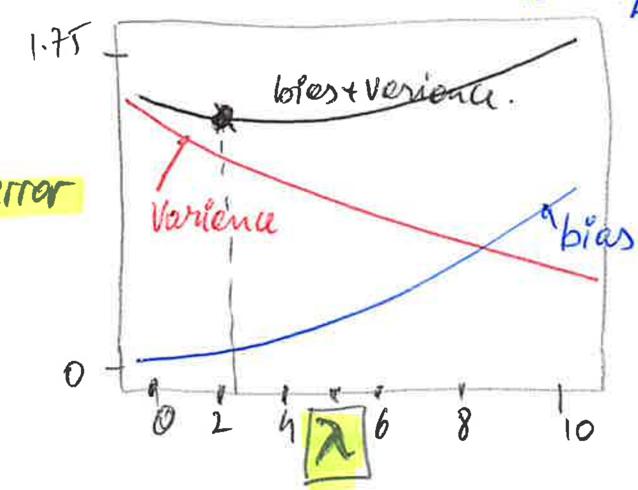
$$\sigma^2 + \frac{\lambda^2}{(n+\lambda)^2}(w^T x)^2 + \frac{\sigma^2 n}{(n+\lambda)^2}\|x\|_2^2$$

irreducible error — Bias-squared — Variance

- λ trades off Bias vs. Variance.
- Larger λ => smaller variance. Larger Bias.

True Error:

$$E_{y,\mathcal{D}_{train}|x}\left[(y - x^T\hat{w}_{ridge})^2 \mid x\right] = \sigma^2 + \underbrace{\frac{\lambda^2}{(n+\lambda)^2}(w^T x)^2}_{\text{Bias-squared}} + \underbrace{\frac{\sigma^2 n}{(n+\lambda)^2}\|x\|_2^2}_{\text{Variance}}$$



$\lambda \to 0$

$\hat{w}_{ridge} \to \hat{w}_{LS}$

$\lambda \to \infty$

$\hat{w}_{ridge} \to 0$

Takeaways:

- Regularization: penalizes complex models towards preferred, simpler models.
- Ridge Regression:
  - L2 penalized least-square regression.
  - Regularization parameter trades off model complexity w/ training error.
  - Never regularize the offset!
- We learned to measure $\boxed{\text{sensitivity}}$ by the size of weights $\|w\|_2^2$

$$\hat{w}_{LS} = \arg\min_{w} \sum_{i=1}^{n} (y_i - x_i^T w)^2$$

Sparsity & the LASSO

– how to make the model compact and interpretable.

$$\boxed{\hat{\omega}_{LS} = \arg\min_{\omega} \sum_{i=1}^{n} (y_i - x_i^T \omega)^2}$$

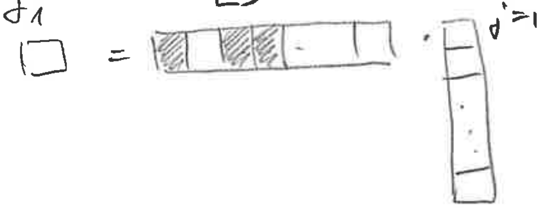
– vector $\boxed{\omega}$ is $\boxed{\text{sparse}}$ if many entries are ϕ

= A vector ω is said to be $\boxed{\text{k-sparse}}$ if at most k entries are non-zero

– We are interested in k-sparse ω, with $k \ll d$, because:
- computationally more efficient
- get rid of redundant/spurious features
- explanability/interpretability

Efficiency
- if size(ω) = 100 billion, each prediction $\omega^T x$ is expensive.
- if ω sparse, prediction computation depends only on non-zeros in ω.

$$\hat{y}_i = \hat{\omega}_{LS}^T \cdot x_i = \sum_{j=1}^{d} \hat{\omega}_{LS}[j] \times x_i[j] =$$



$$= \sum_{j:\hat{\omega}_{LS}[j] \neq 0}^{d} \hat{\omega}_{LS}[j] \times x_i[j]$$

– Computational complexity – decreases from $2d$ to $2k$ for k-sparse $\hat{\omega}_{LS}$

interpretability
– what are the relevant features to make a prediction?
– How do we find the "best" subset of features useful for predicting the price, among all combinations?

Lot size
Single family
Year Built
Last sold price
Finished sqft
Finished basement
Parking type
Cooling
Heating
Exterior materials
Roof type
Structure type
...

Finding Best subset:

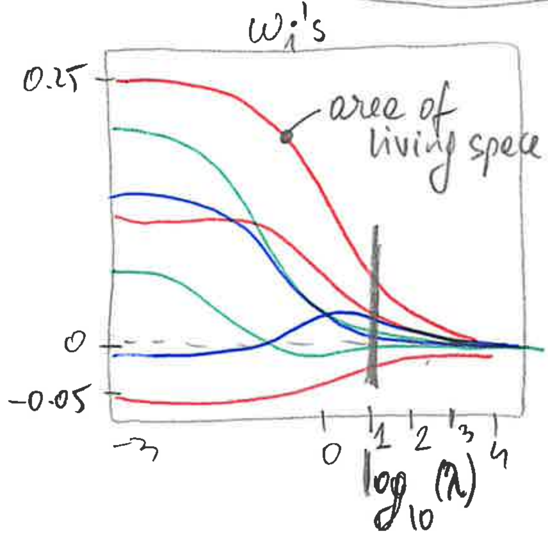
→ **Exhaustive** – time consuming

→ **Greedy** – Forward stepwise – start from simple model; add iteratively features most useful to fit

– Backward stepwise – start w/ full model; iteratively remove features least usefull to fit.

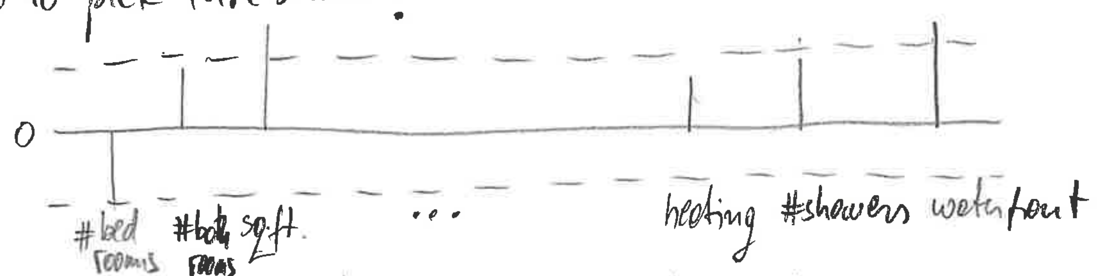
→ **Regularize** – Recall that Ridge Regression makes coefficients small

$$\boxed{\hat{\omega}_{ridge} = \arg\min_{\omega} \sum_{i=1}^{n} (y_i - x_i^T \omega)^2 + \lambda \|\omega\|_2^2}$$



Threshold Ridge Regression: – just set small ridge coefficient to ϕ

– How to pick threshold?



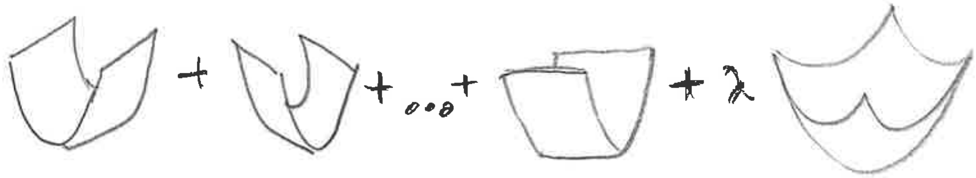
– Challenge: related features (bathrooms, showers)
  - if did not include showers, weight on bathrooms increases!
  - we want a feature selection scheme to select one of the two automatically!

– Can we do this automatic selection w/ another regularizer?

Ridge vs. Lasso Regression

- **Ridge** Regression objective:

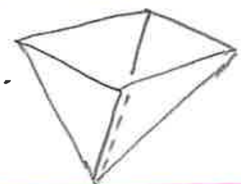
$$\hat{\omega}_{ridge} = \arg\min_{\omega} \sum_{i=1}^{n} (y_i - x_i^T \omega)^2 + \lambda \|\omega\|_2^2$$



- **Lasso** objective:

$$\hat{\omega}_{Lasso} = \arg\min_{\omega} \sum_{i=1}^{n} (y_i - x_i^T \omega)^2 + \lambda \|\omega\|_1$$

—//— —//— + $\lambda \cdot$



LASSO ≡ Least Absolute Shrinkage and Selection Operator

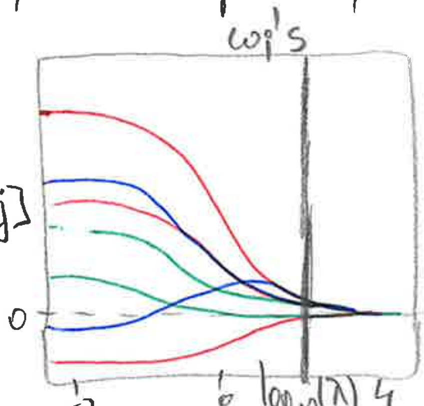
- Sensitivity of a model ω is measured in L1-norm:

$$\|\omega\|_1 = \sum_{i=1}^{d} \underbrace{|\omega[j]|}_{\text{absolute value}}.$$

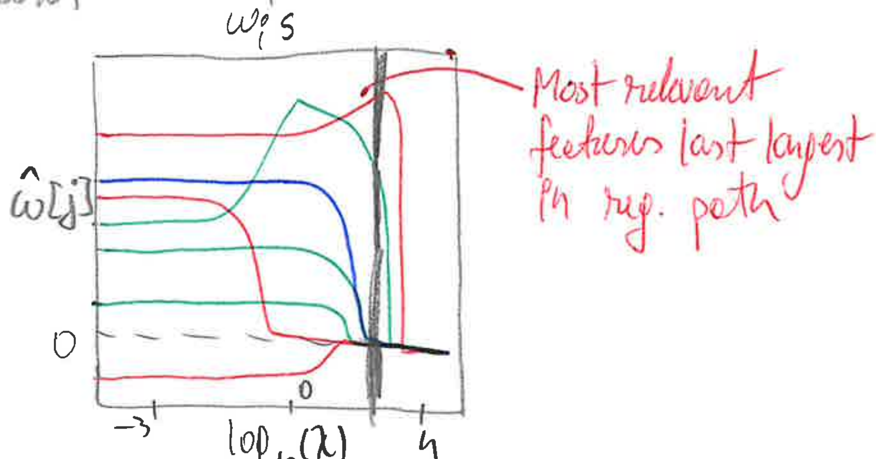
(ℓ_p-norm of a vector $\omega \in \mathbb{R}^d$ is:

$$\|\omega\|_p \triangleq \left(\sum_{j=1}^{d} |\omega[j]|^p \right)^{1/p}$$)

Example: house price w/ 16 features



Ridge regression



Lasso regression

- Lasso regression - naturally gives sparse features.

1 > model selection - choose λ based on (cross) validation error

2 > feature selection - keep only features with non-zero (or not-too-small) parameters in ω at optimal λ

3 > retrain with the sparse model and $\lambda = 0$.

- Regularized Least Squares

Optimization:
$$\boxed{\hat{\omega}_r = \arg\min_{\omega} \sum_{i=1}^{n} (y_i - x_i^T \omega)^2 + \lambda \cdot r(\omega)}$$

Ridge : $r(\omega) = \|\omega\|_2^2$

Lasso : $r(\omega) = \|\omega\|_1$

Theorem:

For any $\lambda^* \geq 0$ for which $\hat{\omega}_r$ achieves the minimum, there exists a $\mu^* \geq 0$ such that the solution of the constrained optimization, $\hat{\omega}_c$, is the same as the solution of the regularized optimization, $\hat{\omega}_r$, where:

$$\boxed{\hat{\omega}_c = \arg\min_{\omega} \sum_{i=1}^{n} (y_i - x_i^T \omega)^2} \quad \text{subject to } \boxed{r(\omega) \leq \mu^*}$$

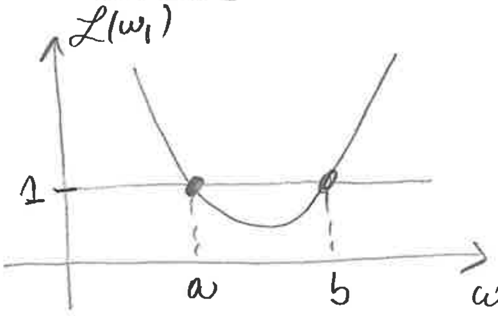
So, there are pairs (λ, μ) whose optimal solution $\hat{\omega}_r$ are the same for the regularized optimization and constrained optimization!

Why does Lasso give sparse solutions?

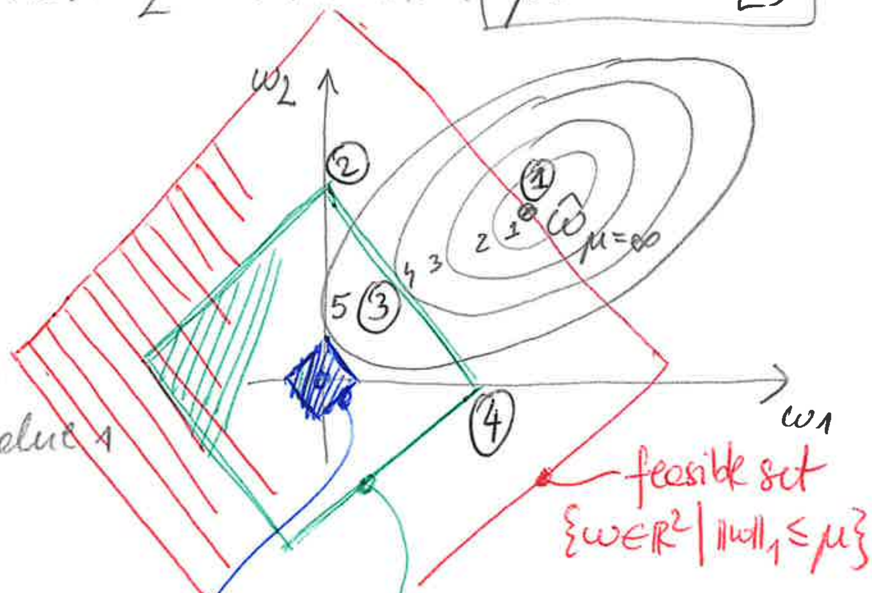
$$\underset{\omega}{\text{minimize}} \sum_{i=1}^{n} (\omega^T x_i - y_i)^2$$

$$\text{subject to: } \|\omega\|_1 \le \mu$$

- "Level Set" of a function $\mathcal{L}(\omega_1, \omega_2)$ is defined as set of points (ω_1, ω_2) that have the same function value.
- Level set of a quadratic function is an oval.
- Center of oval is the Least Squares solution $\boxed{\hat{\omega}_{\mu=\infty} = \hat{\omega}_{LS}}$
- Examples:



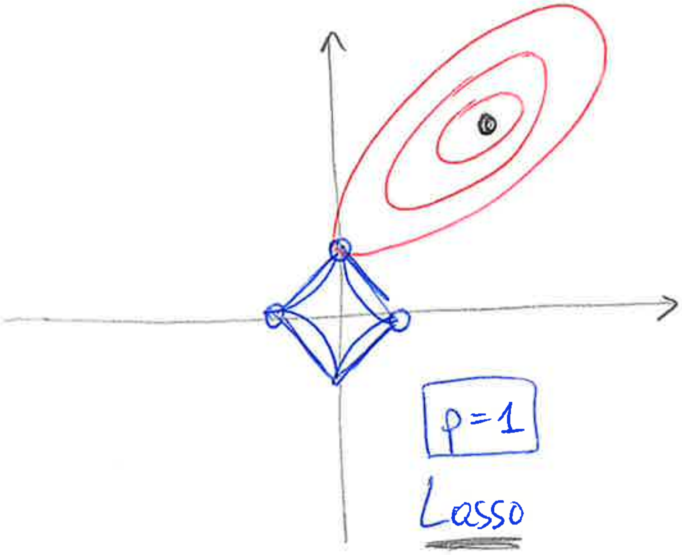
Level Set of $\mathcal{L}(\omega_1)$ at value 1 is $\{a, b\}$

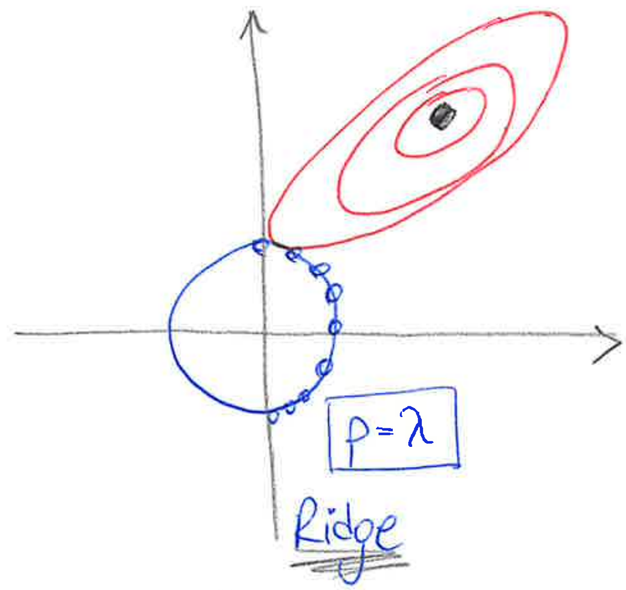


- As we decrease μ from ∞, the feasible set becomes smaller.
- The shape of the feasible set is known as L1 ball, which is a diamond
- in 2-dimensions, diamond is: $\{(\omega_1, \omega_2) \mid |\omega_1| + |\omega_2| \le \mu\}$
- when μ is large enough such that $\|\hat{\omega}_{\mu=\infty}\|_1 < \mu$, then, the optimal solution does not change as the feasible set includes the un-regularized optimal solution
- As μ decreases (which is equivalent to increasing regularization λ) the feasible set (green diamond in figure above) shrinks.
- For small enough μ, the optimal solution becomes $\boxed{\text{sparse}}$; because the L1-ball is "pointy" (i.e., has sharp edges aligned with the axes)

Constrained Least Squares

- Lasso regression finds sparse solutions, as L_1-ball is "pointy"
- Ridge regression finds dense solutions, as L_2-ball is "smooth".





$$\begin{cases} \underset{w}{\text{minimize}} \; \sum_{i=1}^{n} (w^T x_i - y_i)^2 \\ \text{subject to: } \|w\|_1 \leq \mu \end{cases}$$

$$\begin{cases} \underset{w}{\text{minimize}} \; \sum_{i=1}^{n} (w^T x_i - y_i)^2 \\ \text{subject to: } \|w\|_2^2 \leq \mu. \end{cases}$$

L1-Ball in Higher Dimensions

L1-ball in 3 dimensions