DISCUSSION2
2023.10.19

邵元明

Vector Caculus

Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. The *gradient* of f is the function $\nabla f\colon \mathbb{R}^n \to \mathbb{R}^n$ defined by

$$\nabla f(\vec{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\vec{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\vec{x}) \end{bmatrix}.$$

Consider the function $f(\vec{x}) = \vec{a}^\top \vec{x}$. Then

$$\begin{aligned} \frac{\partial f}{\partial x_i}(\vec{x}) &= \frac{\partial}{\partial x_i}\vec{a}^\top \vec{x} \\ &= \frac{\partial}{\partial x_i}\sum_{j=1}^{n} a_j x_j \\ &= \frac{\partial}{\partial x_i}(a_1 x_1 + \cdots + a_n x_n) \\ &= a_i. \end{aligned}$$

$$f(\vec{x}) = \vec{a}^\top \vec{x},$$

$$\begin{aligned} \nabla f(\vec{x}) &= \begin{bmatrix} \frac{\partial f}{\partial x_1}(\vec{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\vec{x}) \end{bmatrix} \\ &= \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \\ &= \vec{a}. \end{aligned}$$

$$f(\tilde{x}) = \tilde{x}^T A \tilde{x} = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{1n} \\ & & \\ a_{n1} & & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \\ x_n \end{bmatrix}$$

$$= \sum_i \sum_j x_i a_{ij} x_j$$

$$\frac{\partial f}{\partial x_i} = \sum_j (a_{ij} + a_{ji}) x_j$$

$$x_i : x_i a_{ij} x_j + x_j a_{ij} x_i + x_i^2 a_{ii}$$

$$\frac{\partial f}{\partial x_i} = a_{ij} x_j + x_j a_{ij} + 2 x_i a_{ii}$$

$$\nabla f(x) = (A + A^T) \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Vector caculus

$$f(\vec{x}) = \vec{x}^\top A\vec{x}$$

$$\begin{aligned}
\nabla f(\vec{x}) &= \begin{bmatrix} \frac{\partial f}{\partial x_1}(\vec{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\vec{x}) \end{bmatrix} \\
&= \begin{bmatrix} ((A + A^\top)\vec{x})_1 \\ \vdots \\ ((A + A^\top)\vec{x})_n \end{bmatrix} \\
&= (A + A^\top)\vec{x}.
\end{aligned}$$

$$f(\vec{x}) = \langle A\vec{x} - \vec{b},\ A\vec{x} - \vec{b} \rangle$$

$$= (A\vec{x} - \vec{b})^T (A\vec{x} - \vec{b})$$

$$= \vec{x}^T A^T A \vec{x} + \vec{b}^T \vec{b} - (A\vec{x})^T \vec{b} - \vec{b}^T (A\vec{x})$$

$$= \vec{x}^T A^T A \vec{x} + \vec{b}^T \vec{b} - 2\vec{b}^T A \vec{x}$$

$$\nabla_x f(\vec{x}) = 2A^T A \vec{x} + 0 - 2A^T \vec{b}$$

$$= 2A^T (A\vec{x} - \vec{b})$$

$$\nabla f(\vec{x}) = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2} \cdots \frac{\partial f}{\partial x_n} \right]^T$$

$f(x)$ $R \to R$ Taylor's Theorem

Derivative $\quad x_0 \in R$ fixed point

$\frac{\partial f}{\partial x}$

$$f(x_0 + \Delta x) = f(x_0) + \frac{\partial f}{\partial x}\Big|_{x=x_0} (\Delta x) + \frac{1}{2!} \frac{\partial^2 f}{\partial x^2} (\Delta x)^2 + \cdots$$

Taylor's Thm for Vectors

$f(\vec{x})$ $R^n \to R$

$$f(\vec{x}_0 + \Delta \vec{x}) = f(\vec{x}_0) + \underbrace{\nabla f\Big|_{\vec{x}=\vec{x}_0}^T}_{\text{row vector}} \Delta \vec{x} + (\Delta \vec{x})^T \underbrace{\nabla f\Big|_{\vec{x}=\vec{x}_0}}_{\text{Hessian}} (\Delta \vec{x})$$

$$f(\vec{x}) = \|A\vec{x} - \vec{b}\|_2^2 = g(h(\vec{x}))$$

$$g(\vec{x}) = \|X\|_2^2 \qquad \nabla_x g(\vec{x}) = 2\vec{x} \qquad \frac{dg(x)}{d\vec{x}} = 2\vec{x}^T$$

$$h(\vec{x}) = A\vec{x} - \vec{b} \qquad \nabla_x h(\vec{x}) = A^T \qquad \frac{dh(x)}{d\vec{x}} = A$$

$$\therefore \nabla f(\vec{x}) = 2A^T(A\vec{x} - \vec{b})$$

Bias-Variance Decomposition

Local Models in High Dimensions

$$E_{\mathcal{T}}(\hat{y}_0 - E_{\mathcal{T}}(\hat{y}_0))(E_{\mathcal{T}}(\hat{y}_0) - f(x_0)) = 0$$

$$\begin{aligned}
\text{MSE}(x_0) &= E_{\mathcal{T}}[f(x_0) - \hat{y}_0]^2 \\
&= E_{\mathcal{T}}[\hat{y}_0 - E_{\mathcal{T}}(\hat{y}_0) + E_{\mathcal{T}}(\hat{y}_0) - f(x_0)]^2 \\
&= E_{\mathcal{T}}\left[(\hat{y}_0 - E_{\mathcal{T}}(\hat{y}_0))^2 + 2\underline{(\hat{y}_0 - E_{\mathcal{T}}(\hat{y}_0))(E_{\mathcal{T}}(\hat{y}_0) - f(x_0))} + (E_{\mathcal{T}}(\hat{y}_0) - \underbrace{f(x_0)}_{\text{Constant}})^2\right] \\
&= E_{\mathcal{T}}\left[(\hat{y}_0 - E_{\mathcal{T}}(\hat{y}_0))^2\right] + (E_{\mathcal{T}}(\hat{y}_0) - f(x_0))^2 \\
&= \text{Var}_{\mathcal{T}}(\hat{y}_0) + \text{Bias}^2(\hat{y}_0)
\end{aligned}$$

This is known as the bias-variance decomposition.

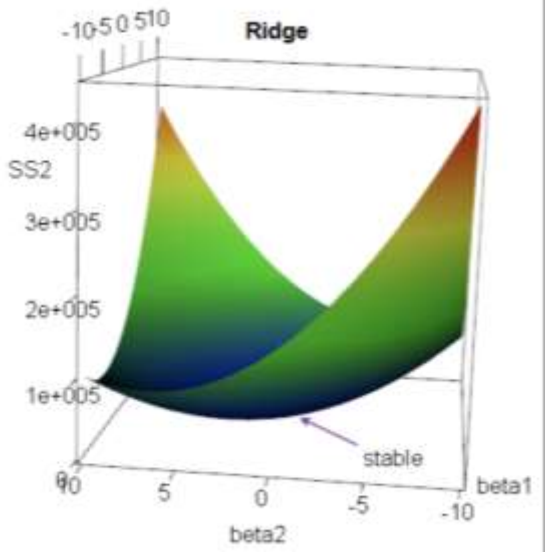
Ridge Regression

Let $A \in \mathbb{R}^{m \times n}$, $\vec{y} \in \mathbb{R}^m$, and $\lambda > 0$. The unique solution to the *ridge regression* problem

$$\min_{\vec{x} \in \mathbb{R}^n} \left\{ \|A\vec{x} - \vec{y}\|_2^2 + \lambda \|\vec{x}\|_2^2 \right\}$$

is given by

$$\vec{x}^\star = (A^\top A + \lambda I)^{-1} A^\top \vec{y}.$$



Proof. Let $f(\vec{x}) \doteq \|A\vec{x} - \vec{y}\|_2^2 + \lambda \|\vec{x}\|_2^2$. By taking gradients, we get

$$\begin{aligned}
\nabla_{\vec{x}} f(\vec{x}) &= \nabla_{\vec{x}} \left\{ \|A\vec{x} - \vec{y}\|_2^2 + \lambda \|\vec{x}\|_2^2 \right\} \\
&= \nabla_{\vec{x}} \{\vec{x}^\top A^\top A\vec{x} - 2\vec{y}^\top A\vec{x} + \vec{y}^\top \vec{y} + \lambda \vec{x}^\top \vec{x}\} \\
&= 2A^\top A\vec{x} - 2A^\top \vec{y} + 2\lambda\vec{x} \\
&= 2(A^\top A + \lambda I)\vec{x} - 2A^\top \vec{y}.
\end{aligned}$$

Thus we get that the optimal point is determined by solving the linear system

$$(A^\top A + \lambda I)\vec{x} = A^\top \vec{y}.$$

Since $A^\top A$ is PSD and $\lambda > 0$, we have $A^\top A + \lambda I$ is PD and thus invertible. Therefore

$$\vec{x}^\star = (A^\top A + \lambda I)^{-1} A^\top \vec{y}$$

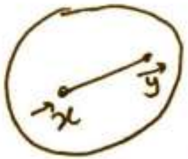
is the unique solution to the above linear system and therefore the unique solution to the optimization problem.

Def^n: Convex Combination:

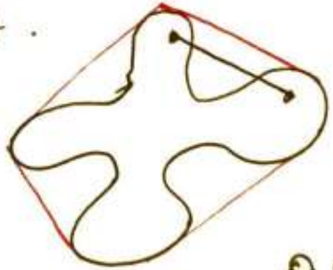
$\sum_{i=1}^{n} \lambda_i \cdot \vec{x}_i$ if $\sum_{i=1}^{n} \lambda_i = 1$ $\quad \lambda \geq 0$.

Def: Convex set:

A set C is convex if the line joining any two points in set is contained in the set.



Convex



Not convex.

$\vec{x}_1 \in C, \ \vec{x}_2 \in C.$ $\quad \theta \in [0, 1].$

$\theta \cdot \vec{x}_1 + (1-\theta) \cdot \vec{x}_2 \in C$

e.g. $C = \{\vec{x} \mid \vec{a}^T \vec{x} = b\}$. $\vec{a}^T(\vec{x}-\vec{x}_0) = 0$

Hyperplane.

$\vec{x_1} \in C$, $\vec{x_2} \in C$.

Consider: $\vec{x_3} = \theta \cdot \vec{x_1} + (1-\theta) \cdot \vec{x_2}$

$$\vec{a}^T \cdot \vec{x_3} = \theta \cdot \vec{a}^T \cdot \vec{x_1} + (1-\theta) \cdot \vec{a}^T \cdot \vec{x_2}$$

$$= \theta \cdot b + (1-\theta) \cdot b$$

$$= b .$$

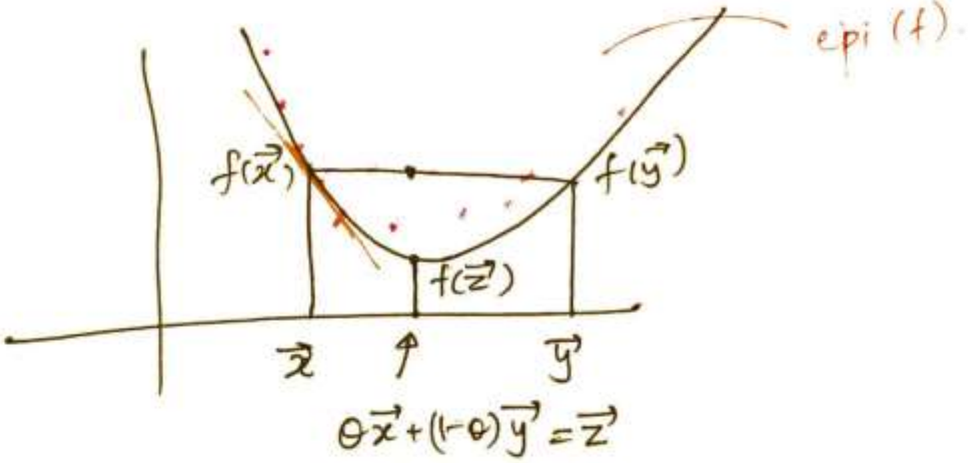
$\Rightarrow \vec{x_3} \in C$. $\therefore$ C is convex.

Convex functions:

$f: \mathbb{R}^n \rightarrow \mathbb{R}$ is convex if

domain f is a convex set.

$$f(\theta \vec{x} + (1-\theta)\vec{y}) \leq \theta \cdot f(\vec{x}) + (1-\theta) f(\vec{y})$$ Jensen's inequality



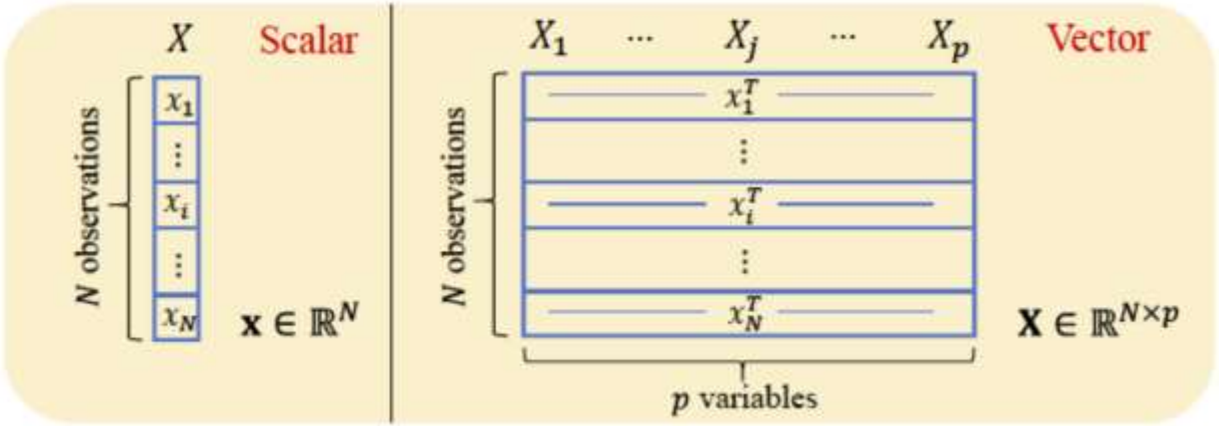
Variable Types and Terminology

Input: a variable X. If X is a vector, its j-th element is X_j

an observation x_i
(scalar or vector)

Typically, we use i to denote the index of observations, while use j to denote the index of variables.

Model



Shrinkage Methods – Ridge Regression

- Shrink the regression coefficients
  - impose a penalty on the size

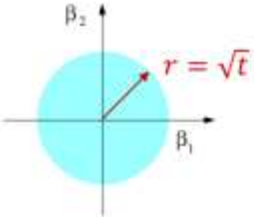
P1
$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2 + \lambda \boxed{\sum_{j=1}^{p} \beta_j^2} \right\}$$

  - the larger the value of λ, the greater the amount of shrinkage
  - the coefficients are shrunk toward zero
- An equivalent expression

P2
$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2$$
$$\text{subject to } \boxed{\sum_{j=1}^{p} \beta_j^2} \le t,$$

  - One-to-one correspondence between λ and t

- Squared ℓ_2-norm on β
$$\|\beta\|_2^2 = \beta^T\beta = \sum_{j=1}^{p} \beta_j^2$$
- Other possible constraints?



$$\text{problem } \mathcal{P}: \quad p^\star = \min_{\vec{x} \in \mathbb{R}^n} \quad f_0(\vec{x})$$
$$\text{s.t.} \quad f_i(\vec{x}) \leq 0, \qquad \forall i \in \{1, \ldots, m\}$$
$$h_j(\vec{x}) = 0, \qquad \forall j \in \{1, \ldots, p\}.$$

Let us denote its feasible set by

$$\Omega \doteq \left\{ \vec{x} \in \mathbb{R}^n \;\middle|\; \begin{array}{ll} f_i(\vec{x}) \leq 0, & \forall i \in \{1, \ldots, m\} \\ h_j(\vec{x}) = 0, & \forall j \in \{1, \ldots, p\} \end{array} \right\} \quad \text{so that} \quad p^\star = \min_{\vec{x} \in \Omega} f_0(\vec{x}).$$

For this : we define the Lagrangian

$$L(\vec{x}, \vec{\lambda}, \vec{\nu}) = f_0(\vec{x}) + \sum_{i=1}^{m} \lambda_i f_i(\vec{x}) + \sum_{i=1}^{p} \nu_i h_i(\vec{x}).$$

when $\lambda_i \geq 0$

$\vec{\lambda}, \vec{\nu}$ are called Lagrange multipliers.

dual variables.

$$\min_{\vec{x}} \; L(\vec{x}, \vec{\lambda}, \vec{\nu}) := \underbrace{g(\vec{\lambda}, \vec{\nu})}_{\text{function } \vec{\lambda}, \vec{\nu}}$$

Lagrange Dual Problem !

$$d^* = \max_{\substack{\vec{\lambda} \geq 0 \\ \vec{\nu}}} \; g(\vec{\lambda}, \vec{\nu}) \Bigg\} \text{ CONVEX PROGRAM.}$$

Shrinkage Methods – Ridge Regression *

- Equivalence between P1 and P2

P1: $\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2$

P2: $\tilde{\beta} = \underset{\beta}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\beta\|_2^2, \text{ s.t.} \|\beta\|_2^2 \leq t$

- Goal: $\forall\lambda, \exists t \geq 0$: $\hat{\beta} = \tilde{\beta}$ (Step 1)
- $\forall t, \exists \lambda \geq 0$: $\hat{\beta} = \tilde{\beta}$ (Step 2)

Proof:

- Step 1: assume that P1 is solved

$$\boxed{-\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\beta}) + \lambda\hat{\beta} = 0}$$

- Lagrange form of P2

$$L(\beta, \mu) = \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \mu(\|\beta\|_2^2 - t)$$

- KKT conditions
  1. $\nabla_\beta L(\tilde{\beta}, \tilde{\mu}) = 0 \implies \boxed{-\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\tilde{\beta}) + \tilde{\mu}\tilde{\beta} = 0}$
  2. $\tilde{\mu}\left(\|\tilde{\beta}\|_2^2 - t\right) = 0$
  3. $\tilde{\mu} \geq 0$
  4. $\|\tilde{\beta}\|_2^2 \leq t$

- Thus,
  - if

$$t = \|\hat{\beta}\|_2^2$$

  - Then

$$\tilde{\mu} = \lambda, \qquad \tilde{\beta} = \hat{\beta}$$

  - Satisfy the KKT conditions.

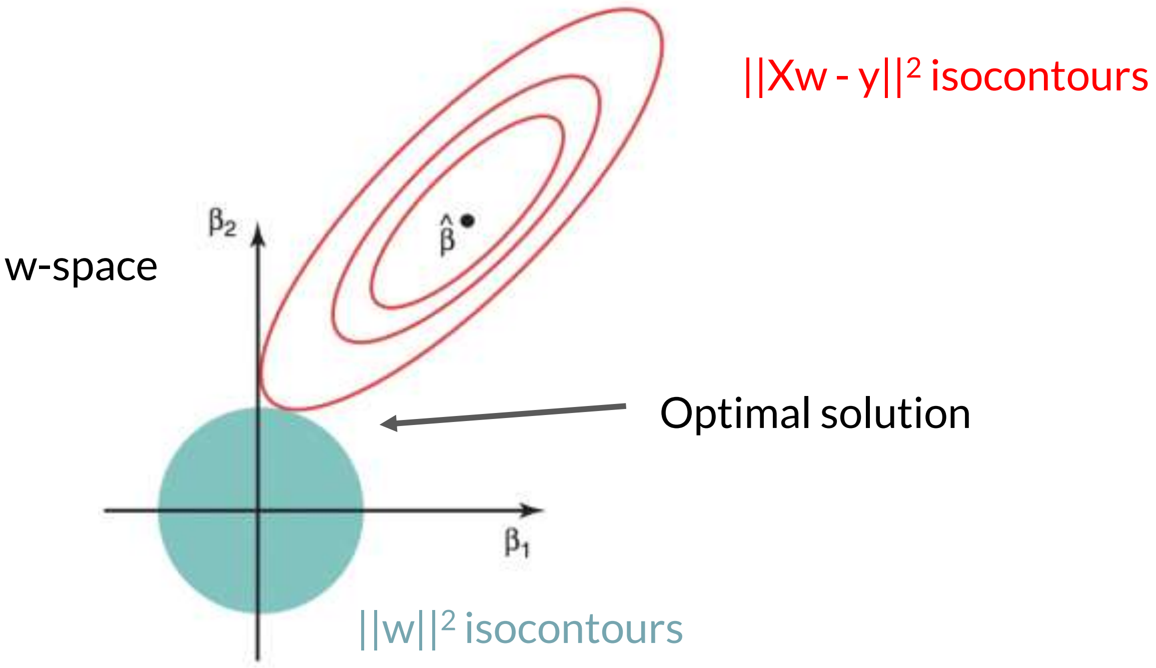
- Step 2: conversely, assume that P2 is solved
- The optimal solution $(\tilde{\beta}, \tilde{\mu})$ must satisfies KKT conditions. Therefore, let $\lambda = \tilde{\mu}$, we always have $\hat{\beta} = \tilde{\beta}$.

Strong duality holds for P2:

$(\tilde{\beta}, \tilde{\mu})$ is the optimal solution of P2 $\Longleftrightarrow$ $(\tilde{\beta}, \tilde{\mu})$ satisfies KKT conditions

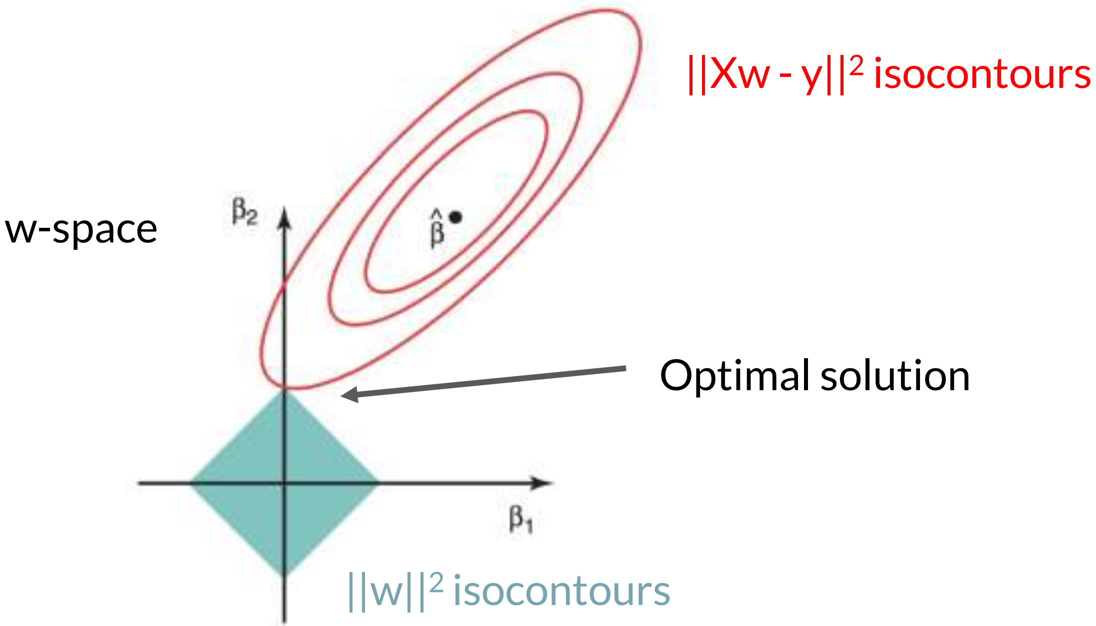
Ridge Regression (L_2 Regularization)

$$\boxed{\text{Find } w \text{ that minimizes } ||Xw - y||^2 + \lambda\,||w'||^2}$$

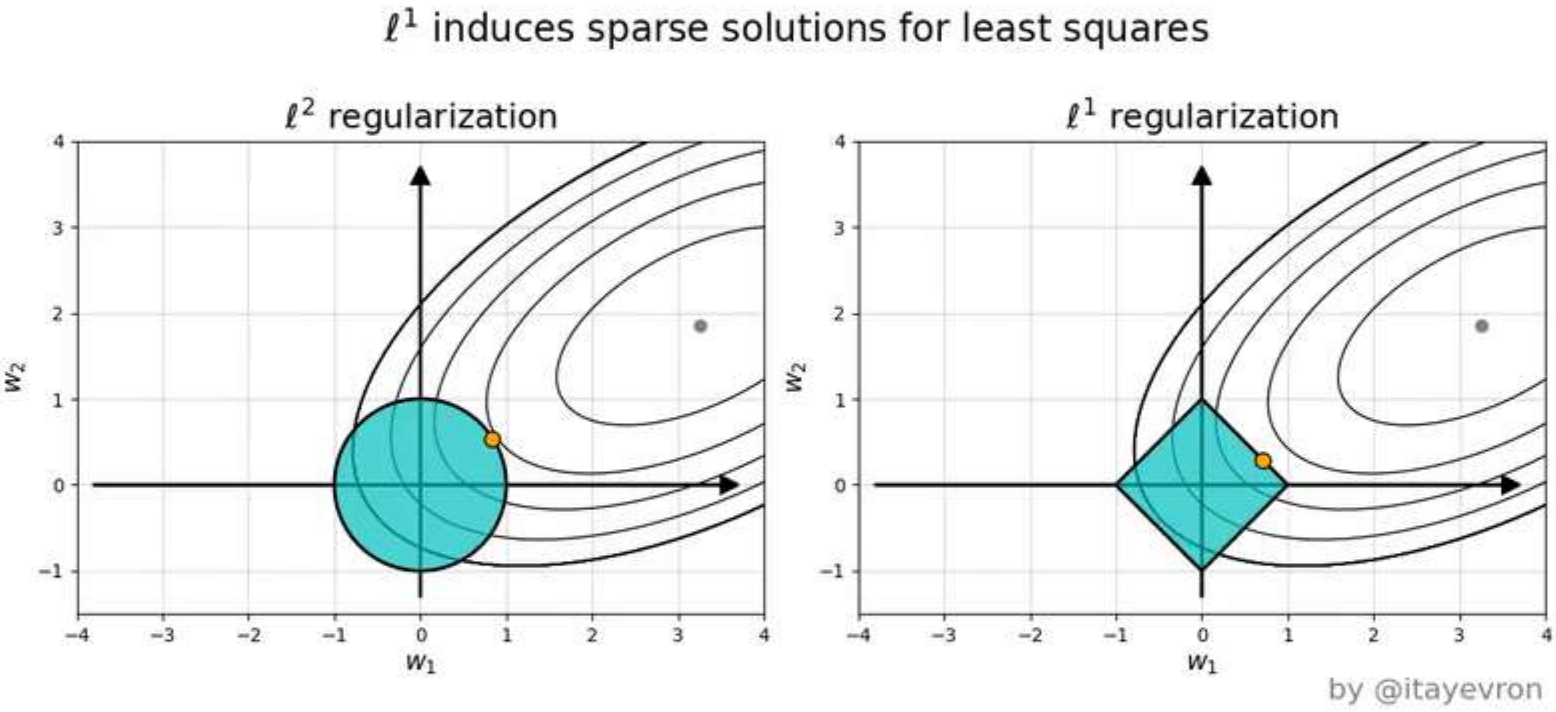


LASSO Regression (L_1 Regularization)

$$\boxed{\text{Find } w \text{ that minimizes } \|Xw - y\|^2 + \lambda \|w'\|_1}$$



L_1 norm induces sparsity



Why is sparsity useful?

It can help us get rid of unnecessary features! If we're predicting the price of a house:

$$y = \begin{pmatrix} \text{\# of bedrooms} \\ \text{Square footage} \\ \text{\# of apple trees} \end{pmatrix} \bullet \begin{pmatrix} 1001.3 \\ 21.2 \\ 0 \end{pmatrix}$$

x $\qquad$ w^*_{LASSO}

Irrelevant feature