Foundations of Machine Learning (ECE 5984)

- Regularization -

Eunbyung Park

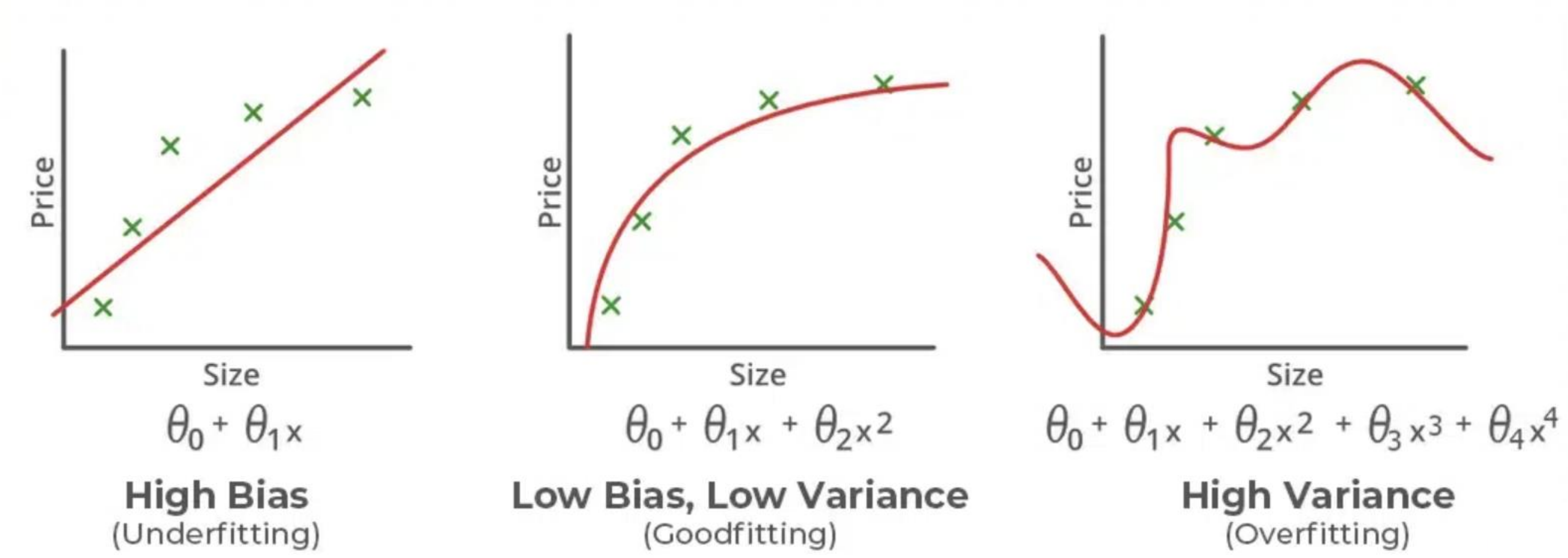
Assistant Professor

School of Electronic and Electrical Engineering

Eunbyung Park (silverbottlep.github.io)

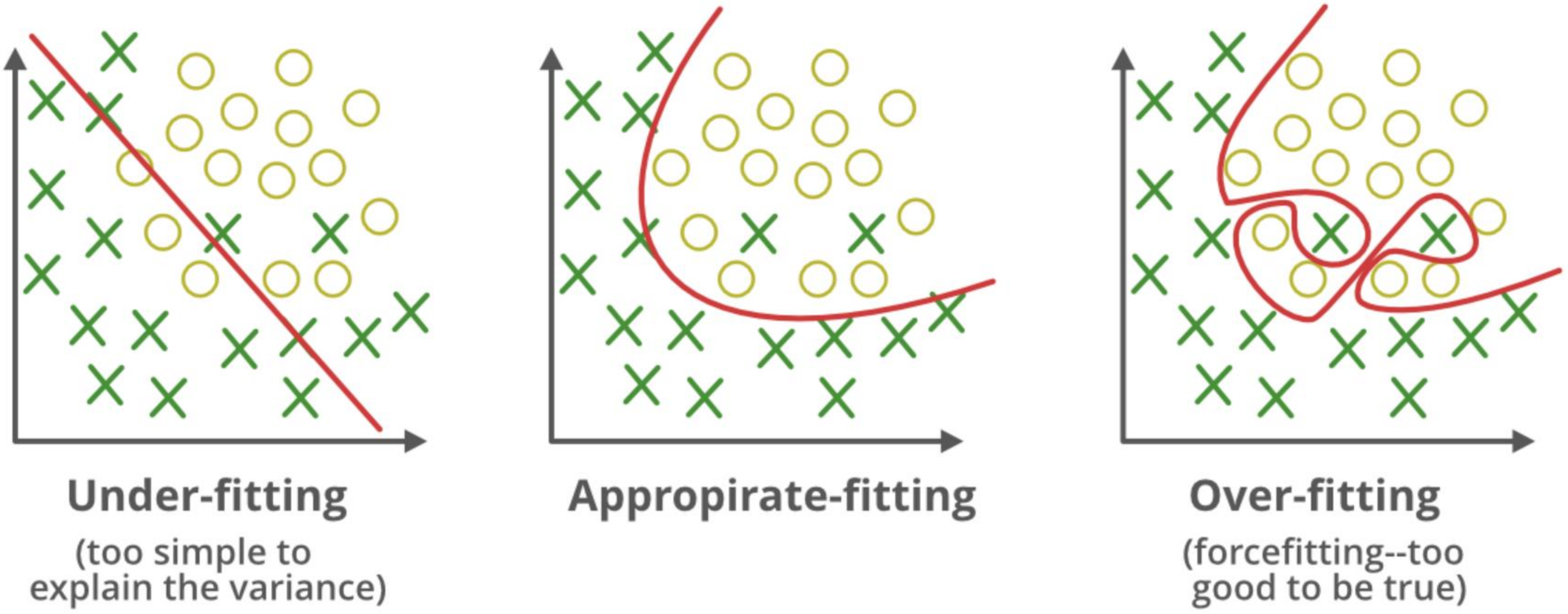
The Problem of Overfitting

Regression Example



[ML | Underfitting and Overfitting - GeeksforGeeks]

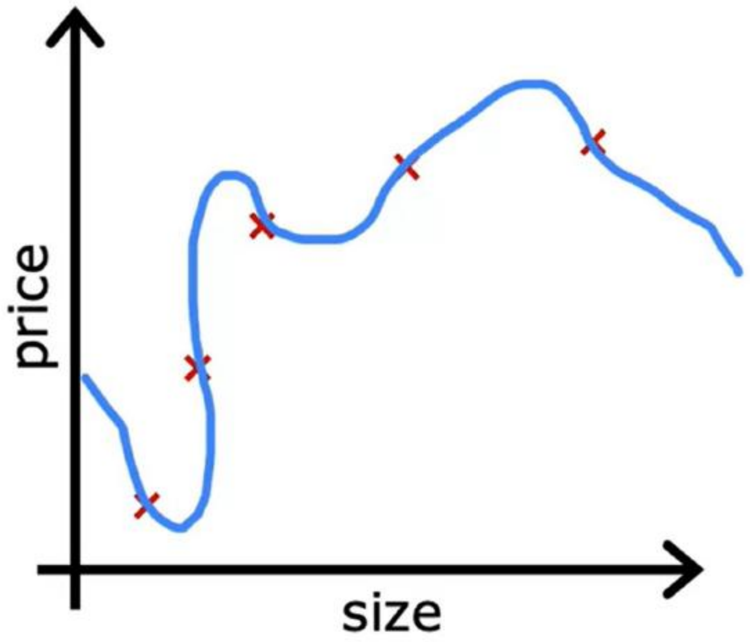
Regression Example

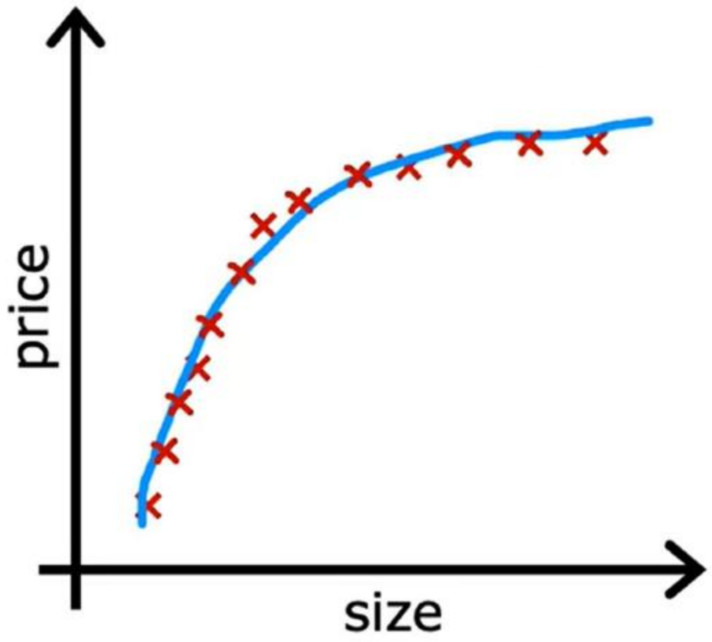


[ML | Underfitting and Overfitting - GeeksforGeeks]

Addressing Overfitting

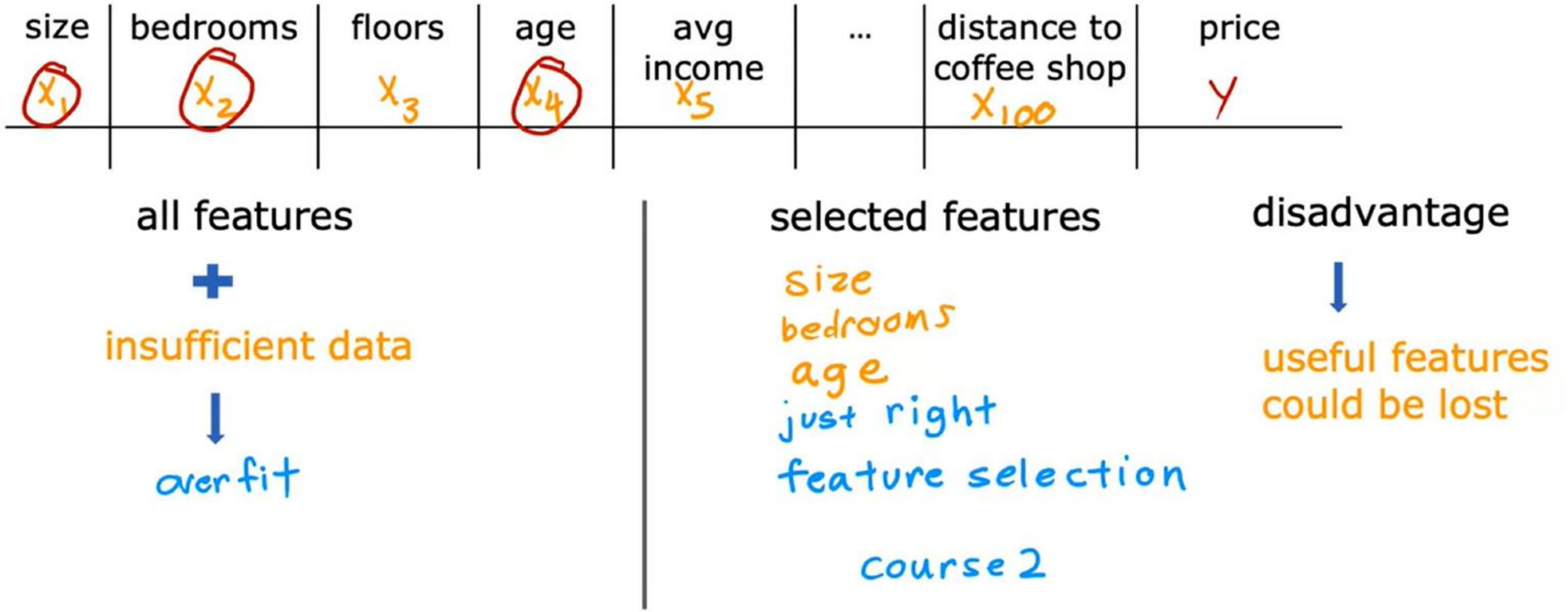
- Collect more data!



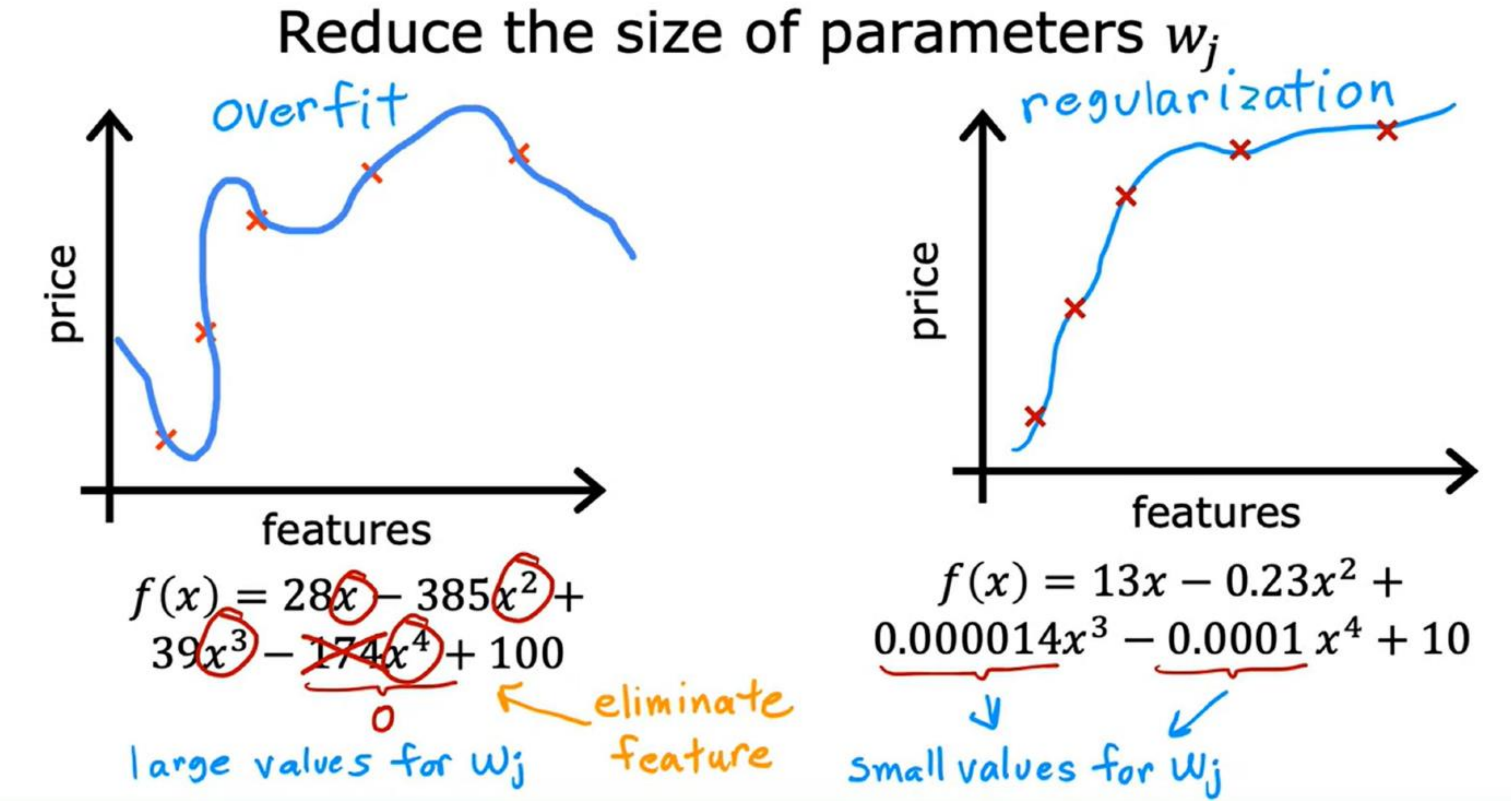


Select Features to Include/Exclude

- Feature selection



Regularization



Ridge Regression

Linear Regression vs Ridge Regression

$$D = \left\{\left(x^{(1)}, y^{(1)}\right), \dots, (x^{(N)}, y^{(N)})\right\}, x^{(i)} \in \mathbb{R}^d, y^{(i)} \in \mathbb{R}, w \in \mathbb{R}^d, X \in \mathbb{R}^{N\times d}, Y \in \mathbb{R}^N$$

Linear Regression

$$L(w) = \frac{1}{2}\sum_{i=1}^{N}\left(w^\top x^{(i)} - y^{(i)}\right)^2$$

Ridge Regression

$$L(w) = \frac{1}{2}\sum_{i=1}^{N}\left(w^\top x^{(i)} - y^{(i)}\right)^2 + \frac{\lambda}{2}||w||^2$$

Linear Regression vs Ridge Regression

$$L(w) = \frac{1}{2}\sum_{i=1}^{N}\left(w^\top x^{(i)} - y^{(i)}\right)^2 + \frac{\lambda}{2}||w||^2$$

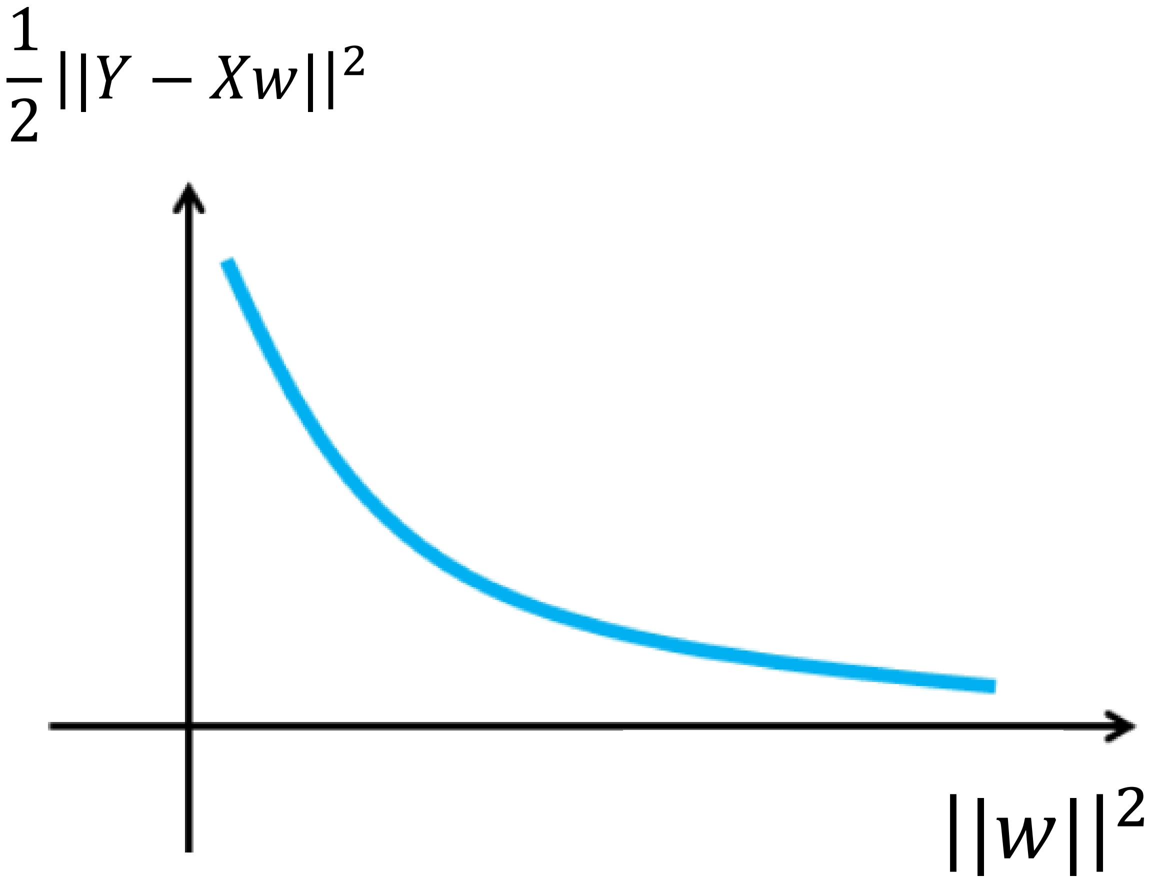
Trade-off

- If $\lambda \to 0,$

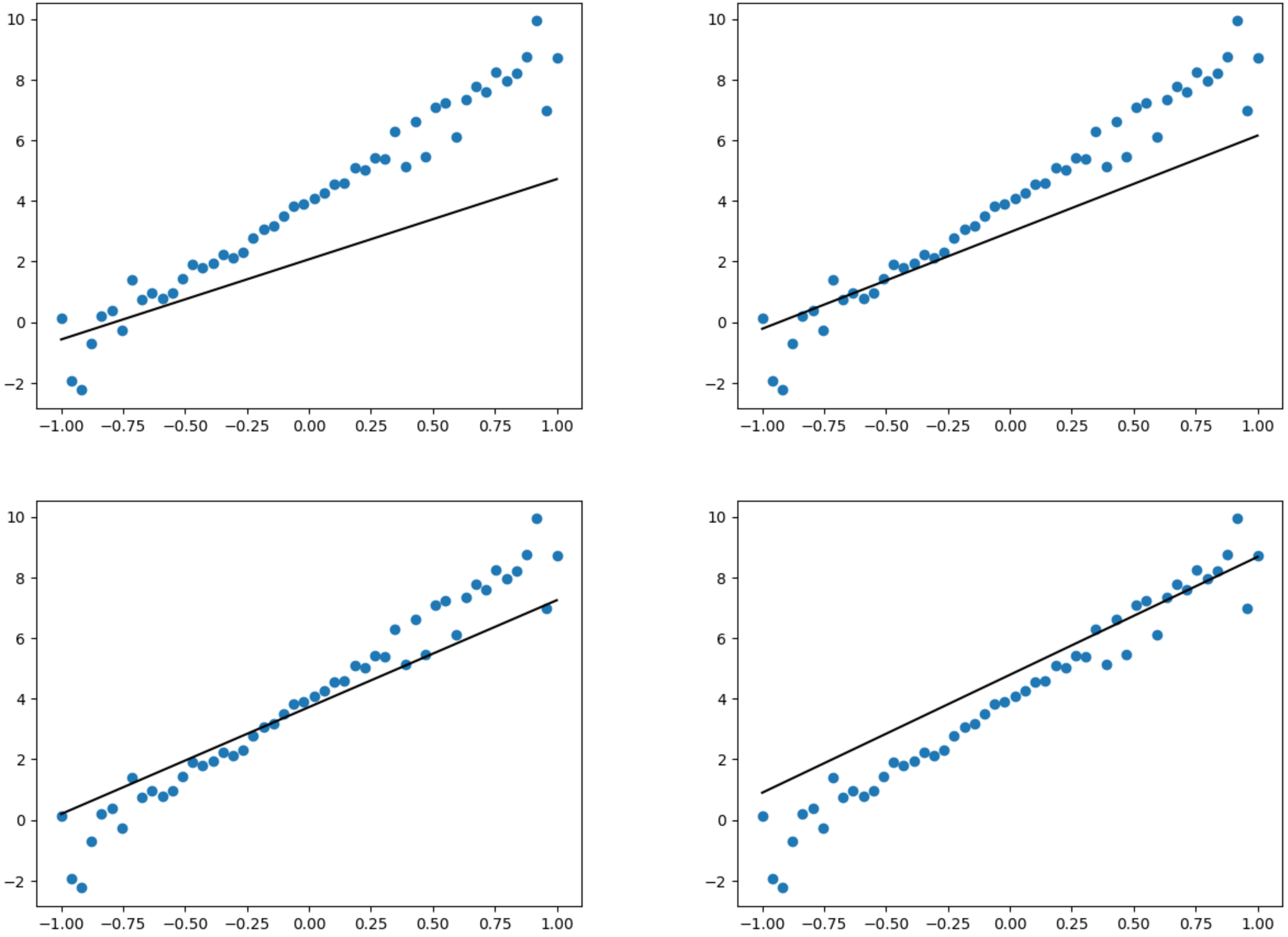
$$L(w) = \frac{1}{2}\sum_{i=1}^{N}\left(w^{\top}x^{(i)} - y^{(i)}\right)^2 + \cancel{\lambda||w||^2}$$

- If $\lambda \to \infty,$

$$L(w) = \cancel{\frac{1}{2}\sum_{i=1}^{N}\left(w^{\top}x^{(i)} - y^{(i)}\right)^2} + \lambda||w||^2$$



Trade-off



Maximum a Posteriori (MAP)

Maximum a Posteriori (MAP)

- Model parameter θ is also random variable
- Can we bring in prior knowledge?
- Bayes Rule

Likelihood

Prior

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

Posterior
Distribution

Maximum a Posteriori (MAP)

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \propto p(D|\theta)p(\theta)$$

constant

$$\arg\max_{\theta} \; \log p(\theta|D) = \log p(D|\theta) + \log p(\theta) - \log p(D)$$

$$p(\theta_j) = N(0, b^2)$$

$$p(\theta) = \prod_{j=1}^{d} N(0, b^2) = \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi b^2}} e^{-\frac{(\theta_j)^2}{2b^2}}$$

Maximum a Posteriori (MAP)

$$p(\theta) = \prod_{j=1}^{d} N(0, b^2) = \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi b^2}} e^{-\frac{(\theta_j)^2}{2b^2}}$$

$$\log p(\theta) = \sum_{j=1}^{d} \log \frac{1}{\sqrt{2\pi b^2}} e^{-\frac{(\theta_j)^2}{2b^2}} = \sum_{j=1}^{d} -\frac{(\theta_j)^2}{2b^2} - \log\sqrt{2\pi b^2} = -\frac{1}{2b^2} ||\theta||_2^2 - d \log\sqrt{2\pi b^2}$$

$$\arg\max_{\theta} \log p(\theta|D) =$$

Maximum a Posteriori (MAP)

$$\arg\min_{\theta} \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(\theta^\top x^{(i)} - y^{(i)}\right)^2 + \frac{1}{2b^2} ||\theta||_2^2$$

$$\frac{\partial}{\partial \theta} \left(\frac{1}{2\sigma^2} (X\theta - Y)^\top (X\theta - Y) + \frac{1}{2b^2} \theta^\top \theta \right) = 0$$

Maximum a Posteriori (MAP)

$$\arg\min_{\theta} \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y^{(i)} - \theta^\top x^{(i)}\right)^2 + \frac{1}{2b^2} ||\theta||_2^2$$

Ridge regression

$$\sigma = 1, \qquad \frac{1}{b^2} = \lambda$$

Maximum a Posteriori (MAP)

$$\theta_{MLE} = (X^\top X)^{-1} X^\top Y \qquad\qquad \theta_{MAP} = (X^\top X + \lambda I)^{-1} X^\top Y$$

Linear Regression vs Ridge Regression

$$D = \left\{\left(x^{(1)}, y^{(1)}\right), \dots, \left(x^{(N)}, y^{(N)}\right)\right\}, x^{(i)} \in \mathbb{R}^d, y^{(i)} \in \mathbb{R}, w \in \mathbb{R}^d, X \in \mathbb{R}^{N\times d}, Y \in \mathbb{R}^N$$

Linear Regression

$$L(w) = \frac{1}{2}\sum_{i=1}^{N}\left(w^\top x^{(i)} - y^{(i)}\right)^2$$

$$L(w) = -\left(\sum_{i=1}^{N} \log p\left(y^{(i)} \middle| x^{(i)}; w\right)\right)$$

$$w_{MLE} = (X^\top X)^{-1} X^\top Y$$

Ridge Regression

$$L(w) = \frac{1}{2}\sum_{i=1}^{N}\left(w^\top x^{(i)} - y^{(i)}\right)^2 + \frac{1}{2}\lambda||w||^2$$

$$L(w) = -\left(\sum_{i=1}^{N} \log p\left(y^{(i)} \middle| x^{(i)}; w\right) + \log p(w)\right)$$

$$w_{MAP} = (X^\top X + \lambda I)^{-1} X^\top Y$$

Gradient Descent of Ridge Regression

$$D = \left\{\left(x^{(1)}, y^{(1)}\right), \dots, (x^{(N)}, y^{(N)})\right\}, x^{(i)} \in \mathbb{R}^d, y^{(i)} \in \mathbb{R}, w \in \mathbb{R}^d, X \in \mathbb{R}^{N \times d}, Y \in \mathbb{R}^N$$

$$L(w) = \frac{1}{2}\sum_{i=1}^{N}\left(w^\top x^{(i)} - y^{(i)}\right)^2 + \frac{\lambda}{2}||w||^2$$

$$\frac{\partial L(w)}{\partial w_j} = \sum_{i=1}^{N}\left(w^\top x^{(i)} - y^{(i)}\right) x_j^{(i)} + \lambda w_j$$

$$\frac{\partial L(w)}{\partial w} = X^\top (Xw - Y) + \lambda w$$

$$w_j := w_j - \alpha\left(\sum_{i=1}^{N}\left(w^\top x^{(i)} - y^{(i)}\right) x_j^{(i)} + \lambda w_j\right)$$

$$w := w - \alpha(X^\top (Xw - Y) + \lambda w)$$

No need to regularize the ‘bias’ term

Gradient Descent of Ridge Regression

$$D = \{(x^{(1)}, y^{(1)}), \dots, (x^{(N)}, y^{(N)})\}, x^{(i)} \in \mathbb{R}^d, y^{(i)} \in \mathbb{R}, w \in \mathbb{R}^d, b \in \mathbb{R}, X \in \mathbb{R}^{N\times d}, Y \in \mathbb{R}^N$$

$$L(w) = \frac{1}{2}\sum_{i=1}^{N}\left((w^\top x^{(i)} + b) - y^{(i)}\right)^2 + \frac{\lambda}{2}||w||^2$$

$$\frac{\partial L(w)}{\partial w_j} = \sum_{i=1}^{N}\left((w^\top x^{(i)} + b) - y^{(i)}\right) x_j^{(i)} + \lambda w_j$$

$$\frac{\partial L(w)}{\partial b} = \sum_{i=1}^{N}\left((w^\top x^{(i)} + b) - y^{(i)}\right)$$

$$w_j \coloneqq w_j - \alpha\left(\sum_{i=1}^{N}\left((w^\top x^{(i)} + b) - y^{(i)}\right) x_j^{(i)} + \lambda w_j\right)$$

$$b \coloneqq b - \alpha\sum_{i=1}^{N}\left((w^\top x^{(i)} + b) - y^{(i)}\right)$$

No need to regularize the ‘bias’ term

Regularized Logistic Regression

$$D = \left\{\left(x^{(1)}, y^{(1)}\right), \dots, (x^{(N)}, y^{(N)})\right\}, x^{(i)} \in \mathbb{R}^d, y^{(i)} \in \{0,1\}, w \in \mathbb{R}^d, X \in \mathbb{R}^{N\times d}, Y \in \{0,1\}^N$$

$$L(w) = -\sum_{i=1}^{N} y^{(i)} \log\left(\hat{y}^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - \hat{y}^{(i)}\right) + \frac{\lambda}{2}||w||^2$$

$$\hat{y}^{(i)} = \sigma\left(w^\top x^{(i)}\right)$$

$$\frac{\partial L(w)}{\partial w_j} = \sum_{i=1}^{N} \left(\hat{y}^{(i)} - y^{(i)}\right) x_j^{(i)} + \lambda w_j$$

$$\frac{\partial L(w)}{\partial w} = X^\top(\sigma(Xw) - Y) + \lambda w$$

$$w_j \coloneqq w_j - \alpha \left(\sum_{i=1}^{N} \left(\hat{y}^{(i)} - y^{(i)}\right) x_j^{(i)} + \lambda w_j\right)$$

$$w \coloneqq w - \alpha(X^\top(\sigma(Xw) - Y) + \lambda w)$$

Convex Optimization

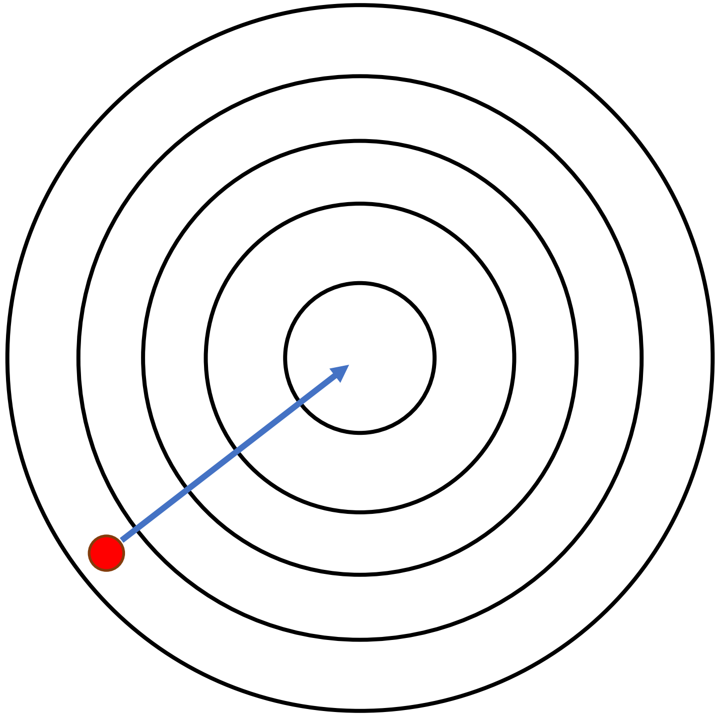
Unconstrained Optimization

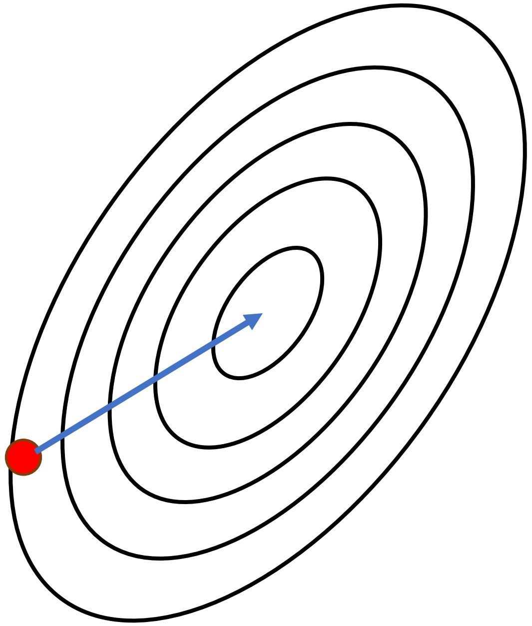
$$\min_{w} L(w)$$

$$L(w) = \frac{1}{2}\sum_{i=1}^{N}\left(w^\top x^{(i)} - y^{(i)}\right)^2$$

In convex optimization,
how can you solve it?

Unconstrained Optimization





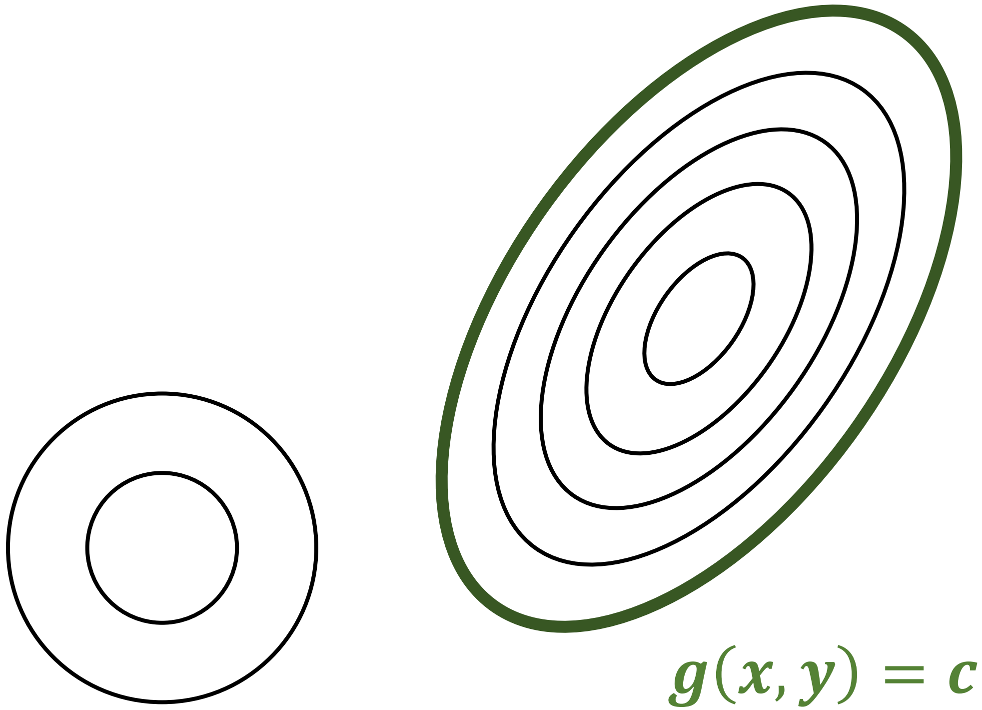
Constrained Optimization

- Equality constraints

$$\min_{x,y} f(x,y) \qquad s.t \quad g(x,y) = c$$

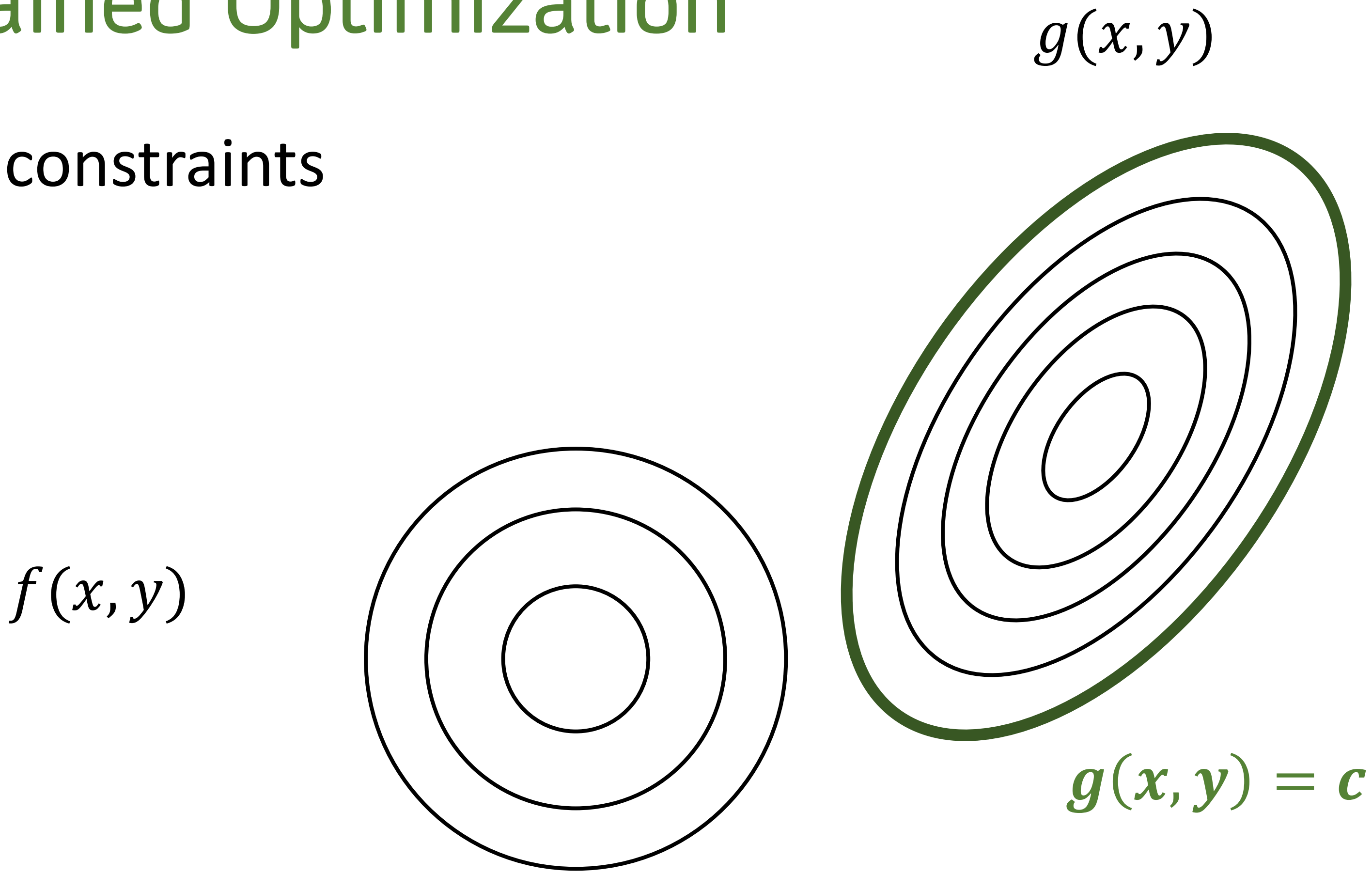
Constrained Optimization

- Equality constraints



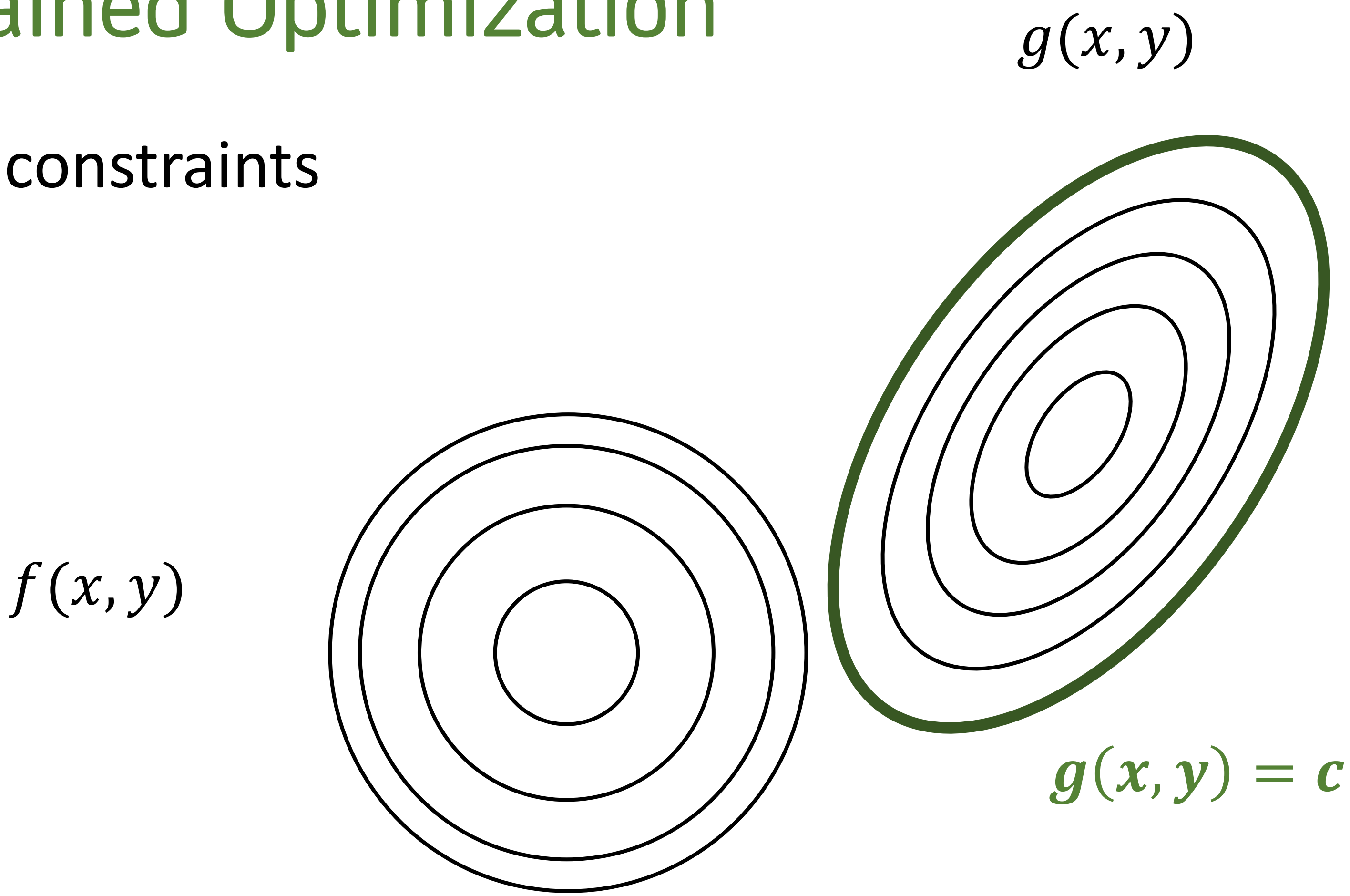
Constrained Optimization

- Equality constraints



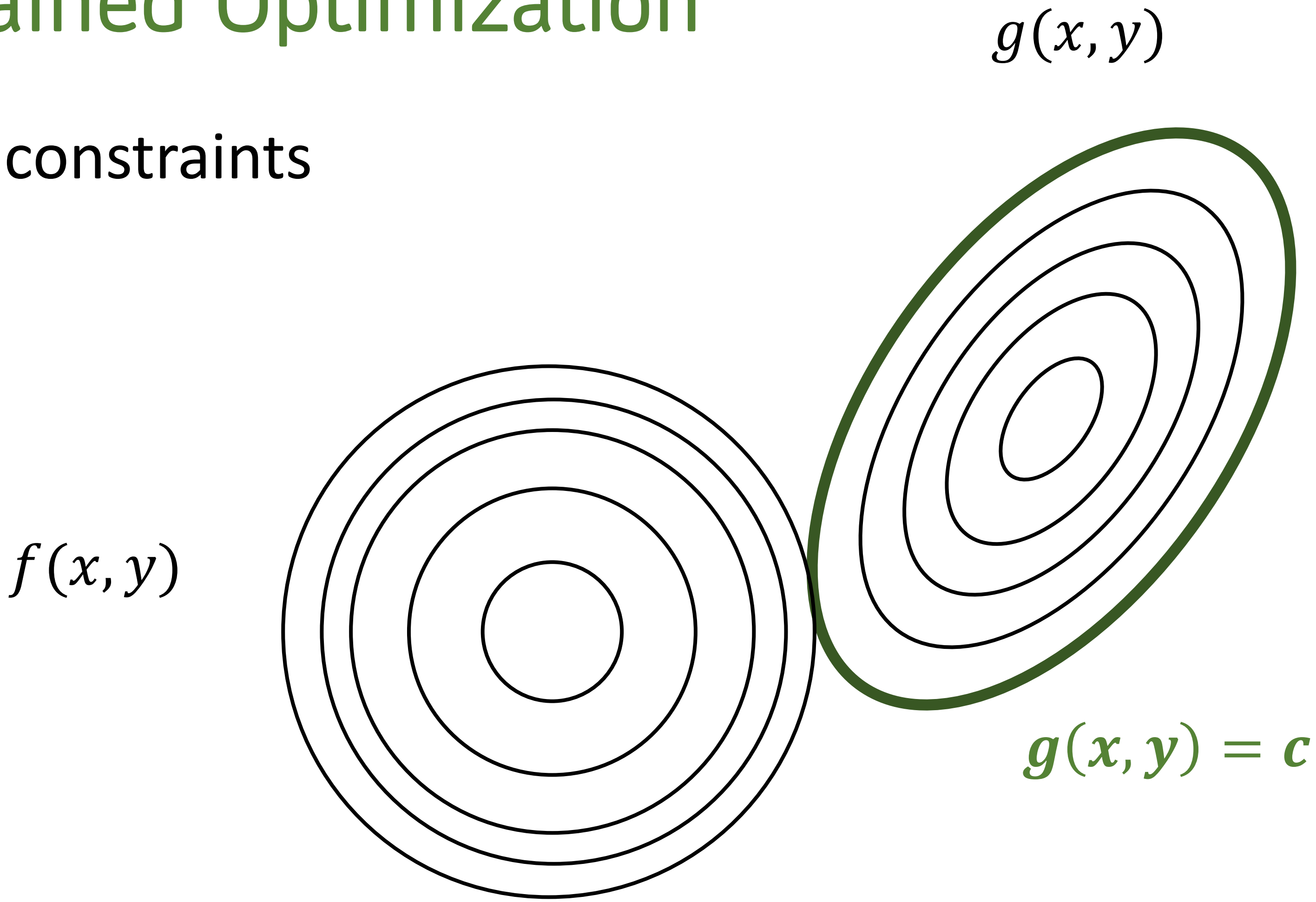
Constrained Optimization

- Equality constraints



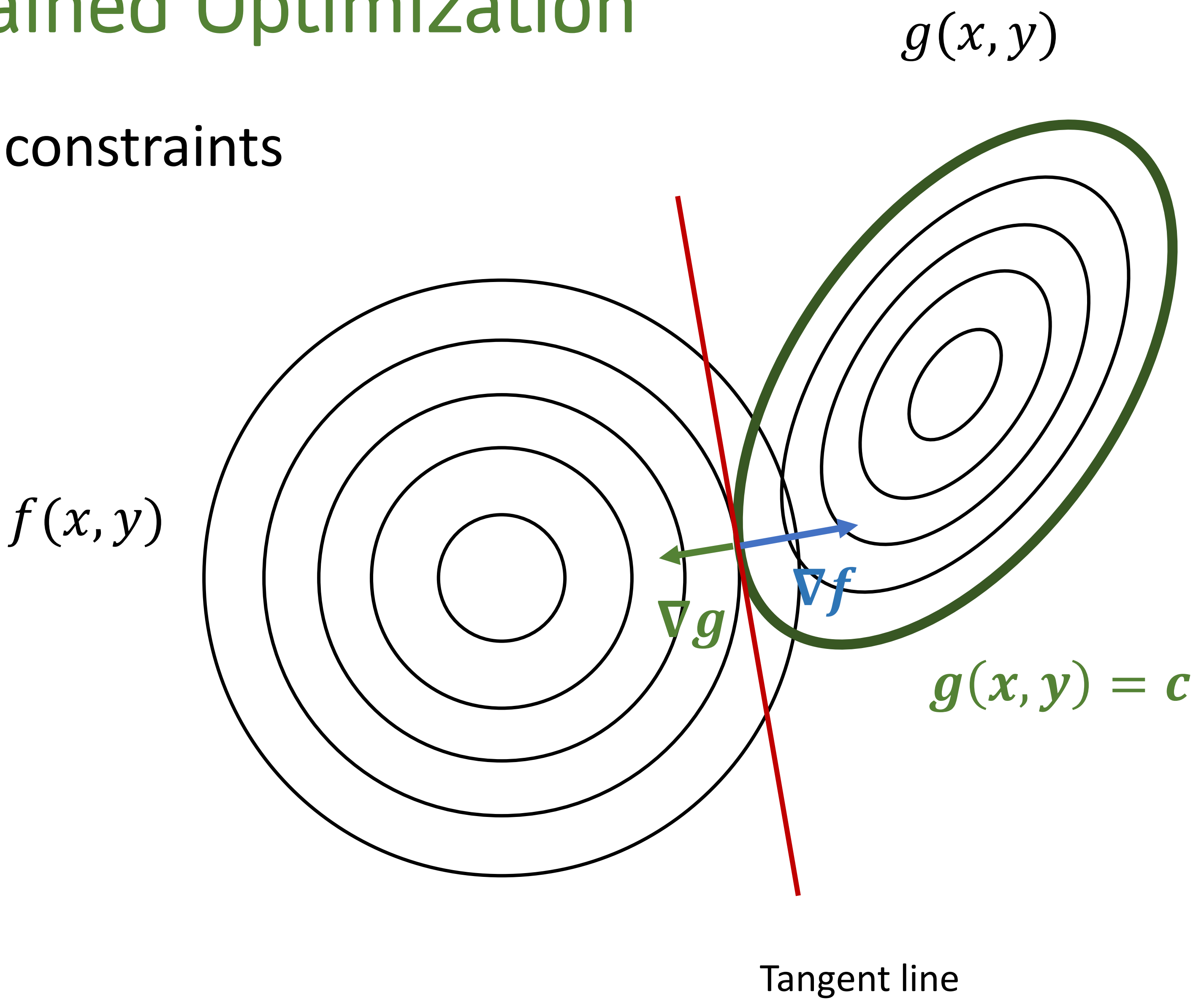
Constrained Optimization

- Equality constraints



Constrained Optimization

- Equality constraints



Optimality condition

$$-\nabla f = \lambda \nabla g$$
$$g(x, y) = c$$

Lagrange Multiplier

- Lagrangian

$$L(x,y,\lambda) = f(x,y) + \lambda(g(x,y) - c)$$

$$\min_{x,y} f(x,y)$$

$$s.t \quad g(x,y) = c$$

$$\nabla L(x,y,\lambda) = 0$$

$$\nabla_x L = \frac{\partial f}{\partial x} + \lambda\left(\frac{\partial g}{\partial x}\right) = 0$$

$$\nabla_y L = \frac{\partial f}{\partial y} + \lambda\left(\frac{\partial g}{\partial y}\right) = 0$$

$$\nabla_\lambda L = g(x,y) - c = 0$$

Lagrange Multiplier

- Lagrangian

$$\min_{x,y} f(x,y)$$

$$s.t \quad g(x,y) = c$$

$$L(x,y,\lambda) = f(x,y) + \lambda(g(x,y) - c)$$

$$\nabla L(x,y,\lambda) = 0$$

$$\nabla_x L = \frac{\partial f}{\partial x} + \lambda\left(\frac{\partial g}{\partial x}\right) = 0$$

$$\nabla_y L = \frac{\partial f}{\partial y} + \lambda\left(\frac{\partial g}{\partial y}\right) = 0$$

$$\nabla_\lambda L = g(x,y) - c = 0$$

Optimality condition

$$-\nabla f = \lambda \nabla g$$
$$g(x,y) = c$$

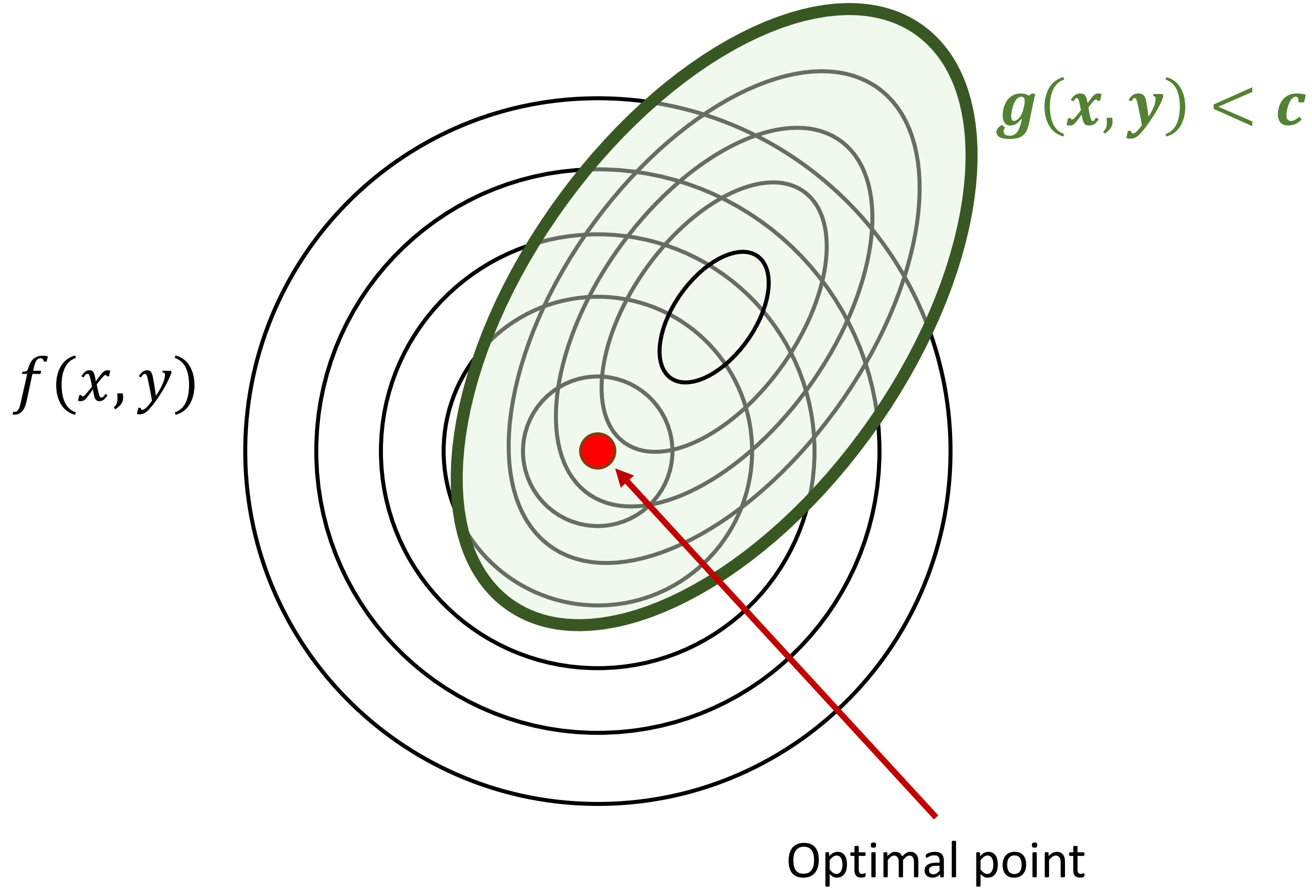
Constrained Optimization

- Inequality constraints

$$\min_{x,y} f(x,y) \qquad s.t \quad g(x,y) \leq c$$

Inequality constraints

- Case 1: Optimal point is in $\boldsymbol{g(x, y) < c}$

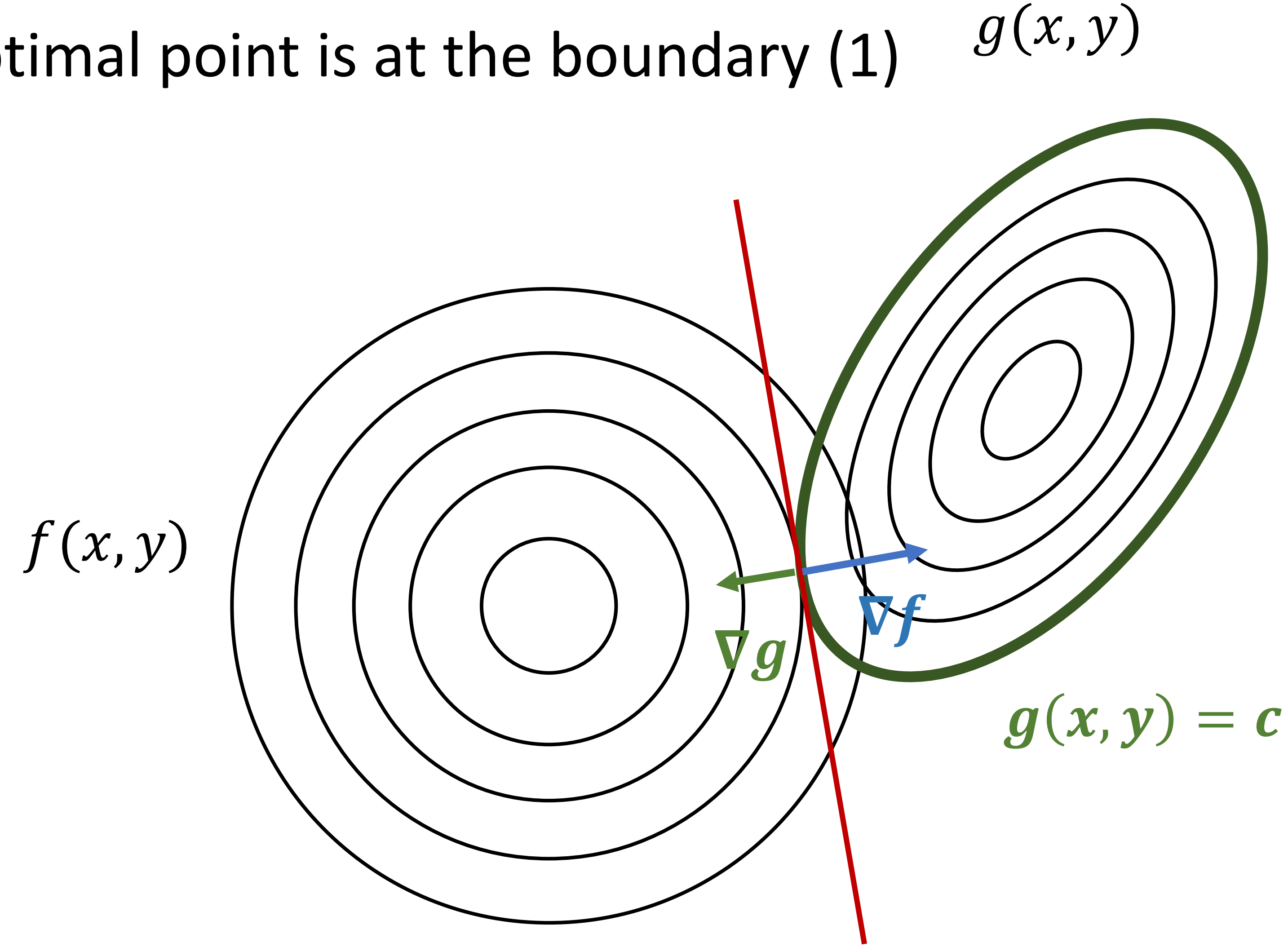


Optimality condition 1

$$\nabla f = 0$$
$$g(x, y) < c$$

Inequality constraints

- Case 2: optimal point is at the boundary (1)

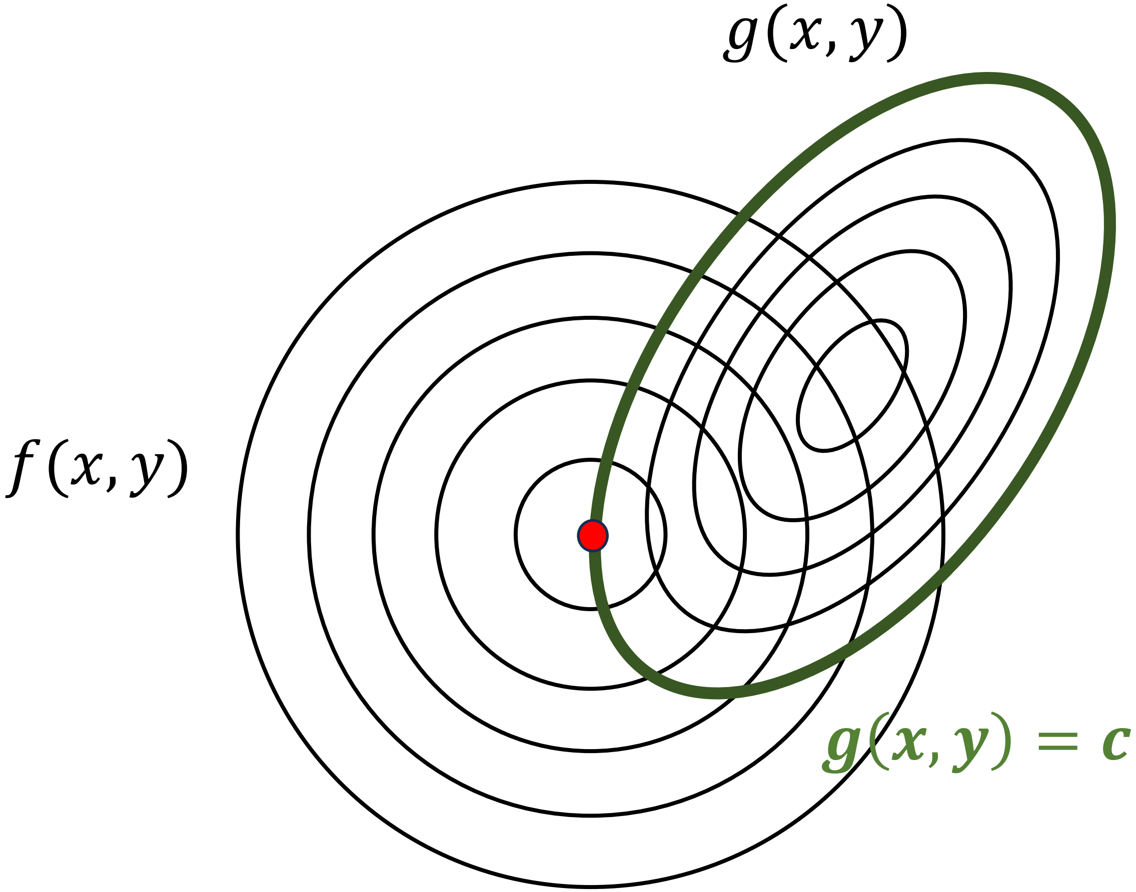


Optimality condition 2

$$-\nabla f = \lambda \nabla g$$
$$g(x,y) = c$$

Inequality constraints

- Case 3: optimal point is at the boundary (2)



Optimality condition 3

$$\nabla f = 0$$
$$g(x,y) = c$$

KKT Conditions

- Karush-Kuhn-Tucker conditions

$$1.\ \nabla L = 0, \qquad \nabla f + \lambda \nabla g = 0$$

$$2.\ g(x, y) - c \leq 0$$

$$3.\ \lambda(g(x, y) - c) = 0$$ (Complementary Condition)

$$4.\ \lambda \geq 0$$

Regularization and Constrained Optimization

Ridge Regression

$$\min_{w} \frac{1}{2} \sum_{i=1}^{N} \left(w^\top x^{(i)} - y^{(i)} \right)^2 + \frac{\lambda}{2} ||w||^2$$

$$w_{MAP} = (X^\top X + \lambda I)^{-1} X^\top Y$$

Constrained Optimization

$$\min_{w} \frac{1}{2}\sum_{i=1}^{N}\left(w^\top x^{(i)} - y^{(i)}\right)^2, \qquad s.t.\ ||w||^2 \leq c$$

KKT Conditions

$$L(w,\mu) = \frac{1}{2}\sum_{i=1}^{N}\left(y^{(i)} - w^\top x^{(i)}\right)^2 + \mu(||w||^2 - c)$$

$$\nabla_w L(w,\mu) = 0$$

1. $\nabla L = 0, \quad \nabla f + \mu \nabla g = 0$
2. $g(x,y) \leq c$
3. $\mu(g(x,y) - c) = 0$
4. $\mu \geq 0$

KKT Conditions

$$L(w,\mu) = \frac{1}{2}\sum_{i=1}^{N}\left(y^{(i)} - w^{\top}x^{(i)}\right)^2 + \mu(||w||^2 - c)$$

$$\mu^*(||w^*||^2 - c) = 0$$

1. $\nabla L = 0, \quad \nabla f + \mu \nabla g = 0$
2. $g(x,y) \le c$
3. $\mu(g(x,y) - c) = 0$
4. $\mu \ge 0$

KKT Conditions

$$L(w, \mu) = \frac{1}{2}\sum_{i=1}^{N}\left(y^{(i)} - w^{\top}x^{(i)}\right)^2 + \mu(||w||^2 - c)$$

1. $\nabla L = 0, \quad \nabla f + \mu \nabla g = 0$
2. $g(x, y) \leq c$
3. $\mu(g(x, y) - c) = 0$
4. $\mu \geq 0$

$$\mu^*(||w^*||^2 - c) = 0$$

Case 1: $||w^*||^2 < c, \ \mu^* = 0$

$$\nabla f + \mu \nabla g = \nabla f = 0$$

No regularization effect!

KKT Conditions

$$L(w,\mu) = \frac{1}{2}\sum_{i=1}^{N}\left(y^{(i)} - w^\top x^{(i)}\right)^2 + \mu(||w||^2 - c)$$

$$\mu^*(||w^*||^2 - c) = 0$$

Case 2: $||w^*||^2 = c, \mu^* \neq 0$

$$\nabla f = -\mu \nabla g$$

Solution at the boundary of the constraint!

1. $\nabla L = 0, \quad \nabla f + \mu \nabla g = 0$
2. $g(x,y) \leq c$
3. $\mu(g(x,y) - c) = 0$
4. $\mu \geq 0$

Ridge Regression and Constrained Optimization

Problem 1 $$\min_w \frac{1}{2}\sum_{i=1}^{N}\left(y^{(i)} - w^\top x^{(i)}\right)^2 + \lambda||w||^2$$

Is equivalent to

Problem 2 $$\min_w \frac{1}{2}\sum_{i=1}^{N}\left(y^{(i)} - w^\top x^{(i)}\right)^2, \qquad s.t.\ ||w||^2 \leq c$$

Ridge Regression and Constrained Optimization

- λ, μ, c relationship?

1. Given any λ, find the solution of the problem 1 , then we get the solution $w^*(\lambda)$
2. It is equivalent to solve the problem 2 with $c = ||w^*(\lambda)\ ||^2$

Ridge Regression and Constrained Optimization

- λ, μ, c relationship?

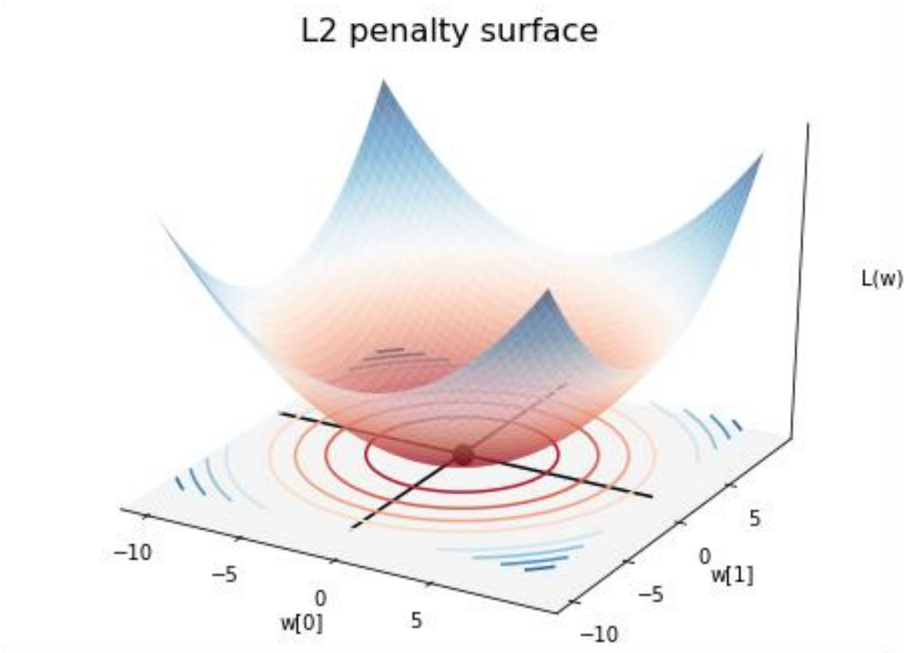
1. Given any λ, find the solution of the problem 1 , then we get the solution $w^*(\lambda)$
2. It is equivalent to solve the problem 2 with $c = ||w^*(\lambda)\,||^2$

$$w^* = (X^\top X + \lambda I)^{-1} X^\top Y$$

$$||w^*(\lambda)\,||^2 = \left((X^\top X + \lambda I)^{-1} X^\top Y\right)^\top \left((X^\top X + \lambda I)^{-1} X^\top Y\right)$$

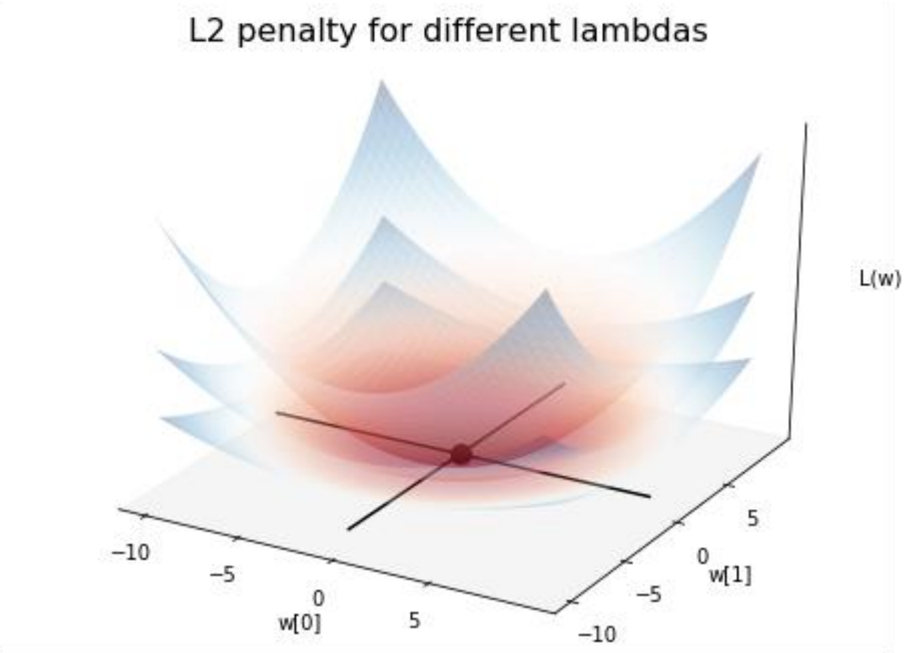
$$\frac{1}{\lambda^2} \propto c$$

Visualization



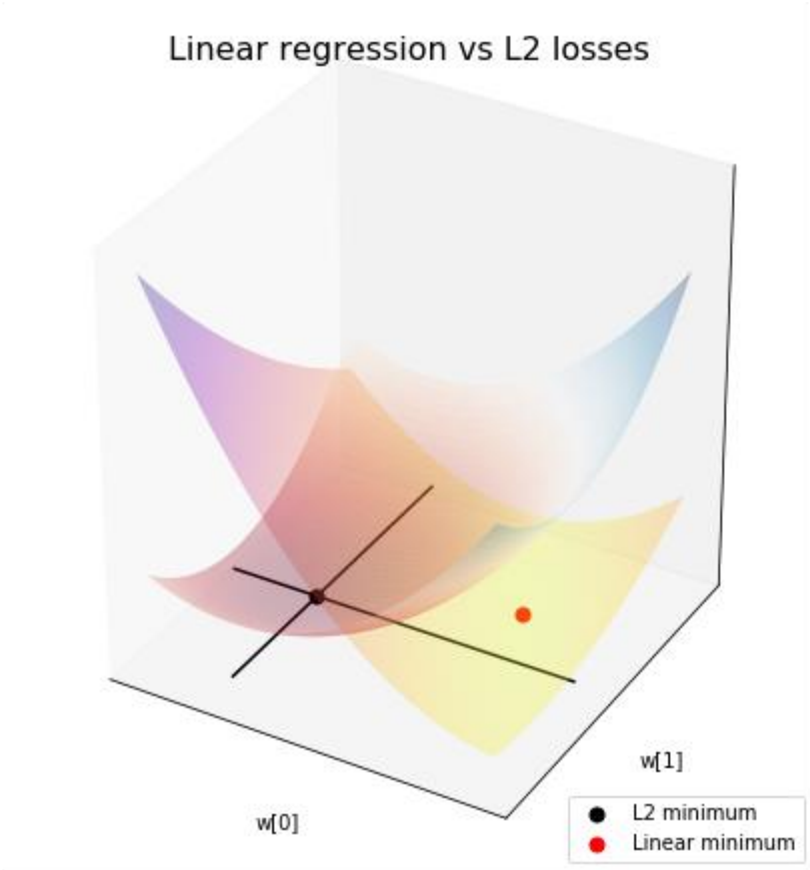
[Linear Regression with Regularization (aunnnn.github.io)](https://aunnnn.github.io)

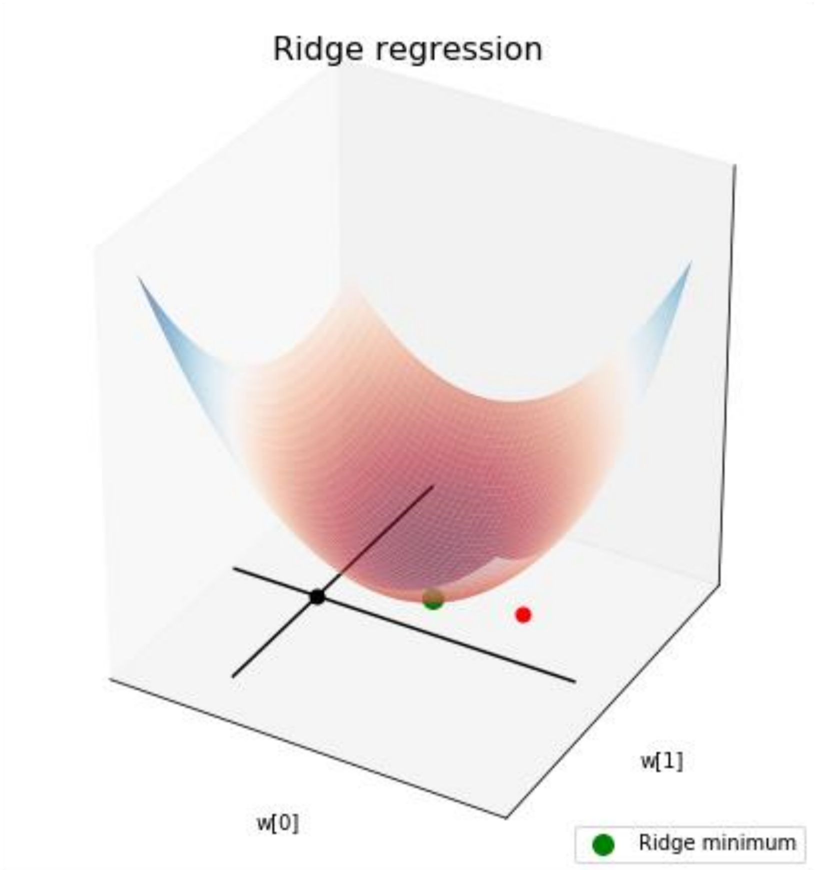
Visualization



Linear Regression with Regularization (aunnnn.github.io)

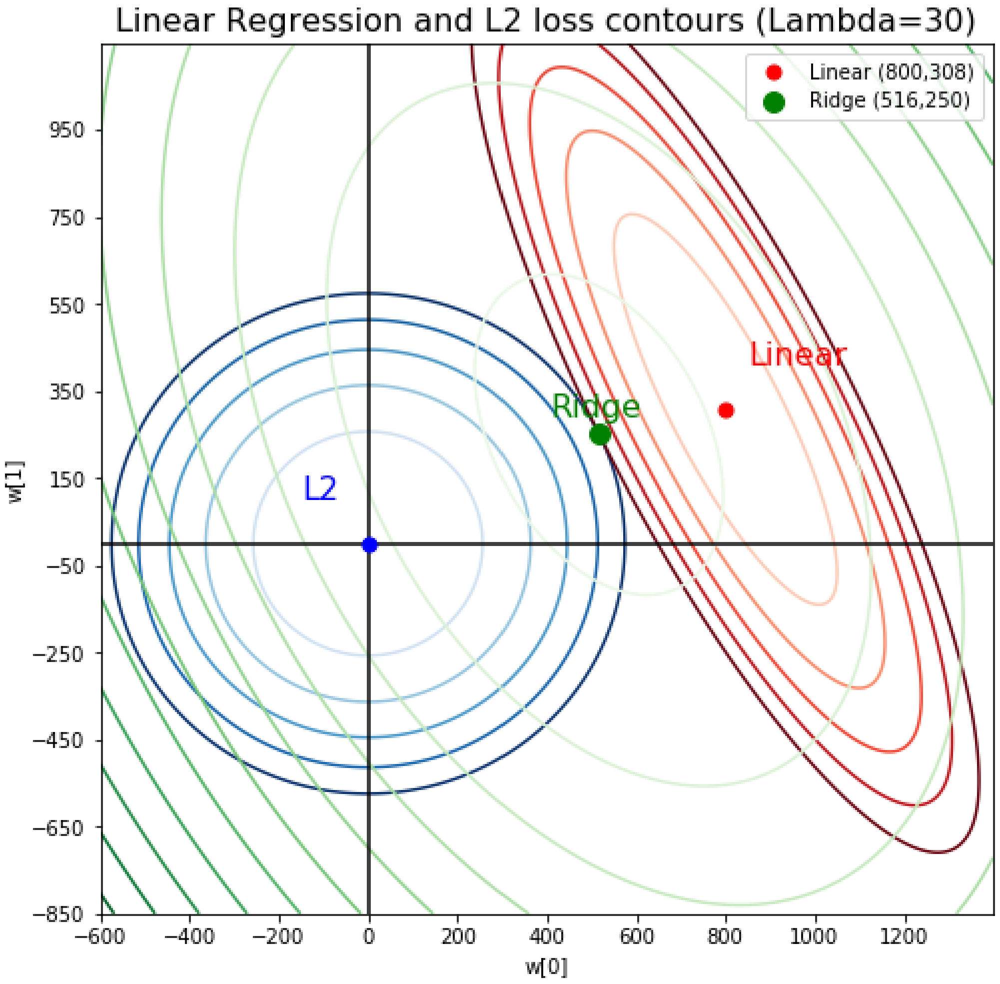
Visualization

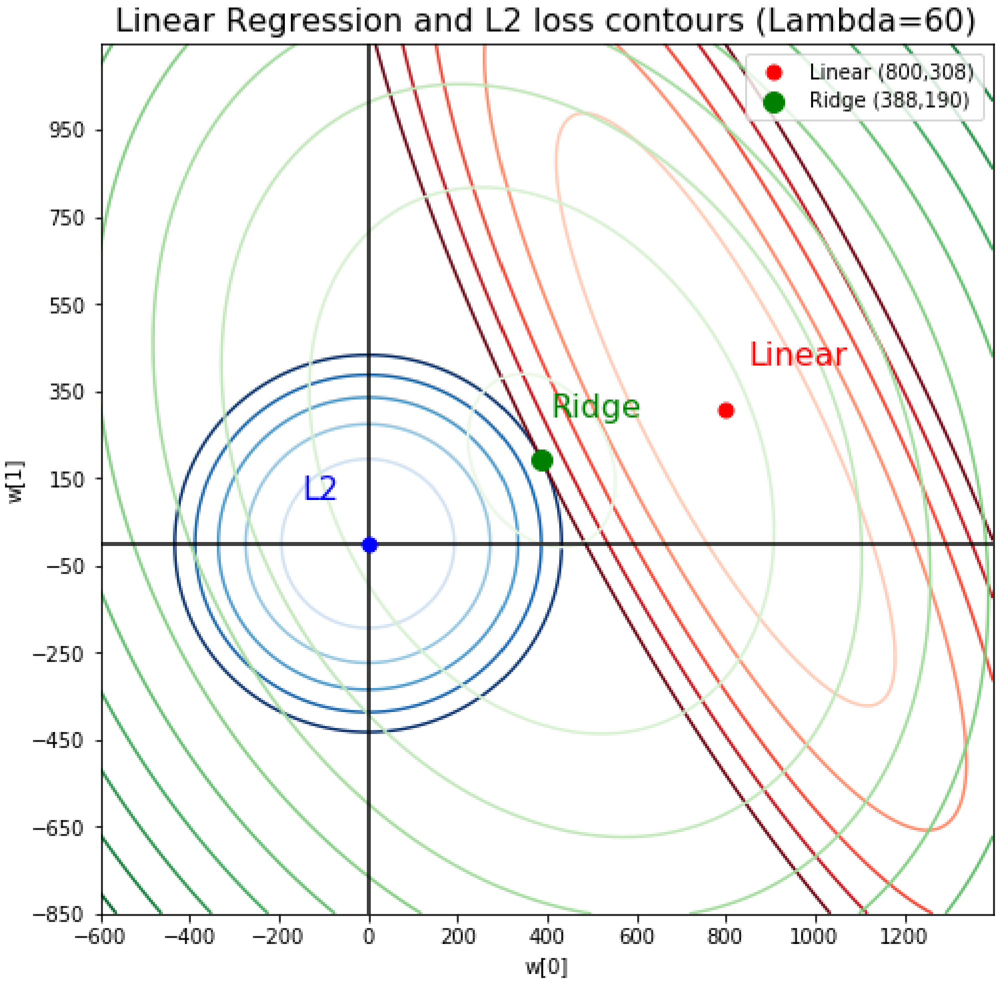




[Linear Regression with Regularization (aunnnn.github.io)](https://aunnnn.github.io)

Visualization





[Linear Regression with Regularization (aunnnn.github.io)]

Lasso

Problem 1

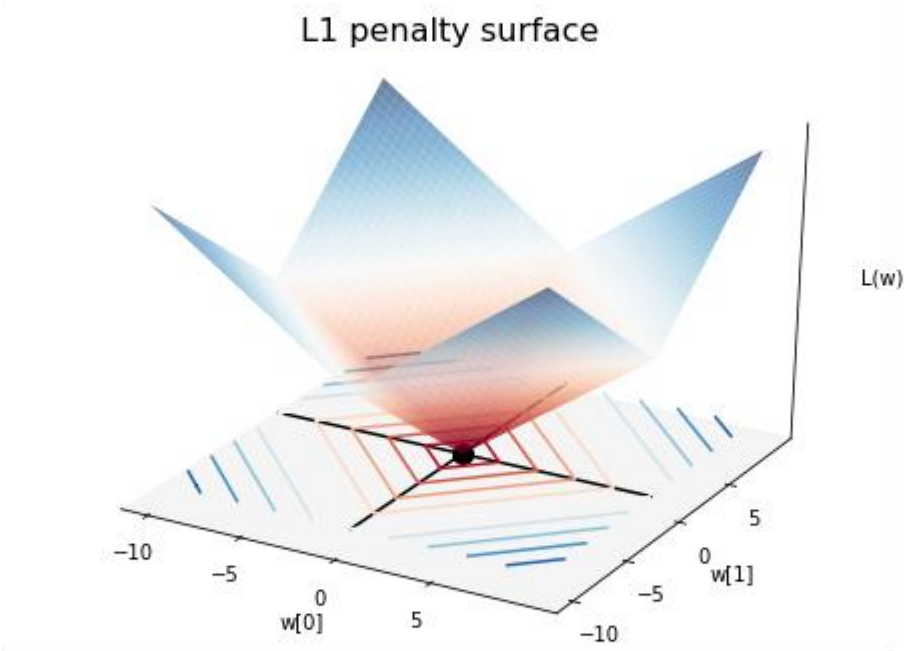
$$\min_{w} \frac{1}{2}\sum_{i=1}^{N}\left(y^{(i)} - w^{\top}x^{(i)}\right)^2 + \lambda||w||_1$$

Is equivalent to

Problem 2

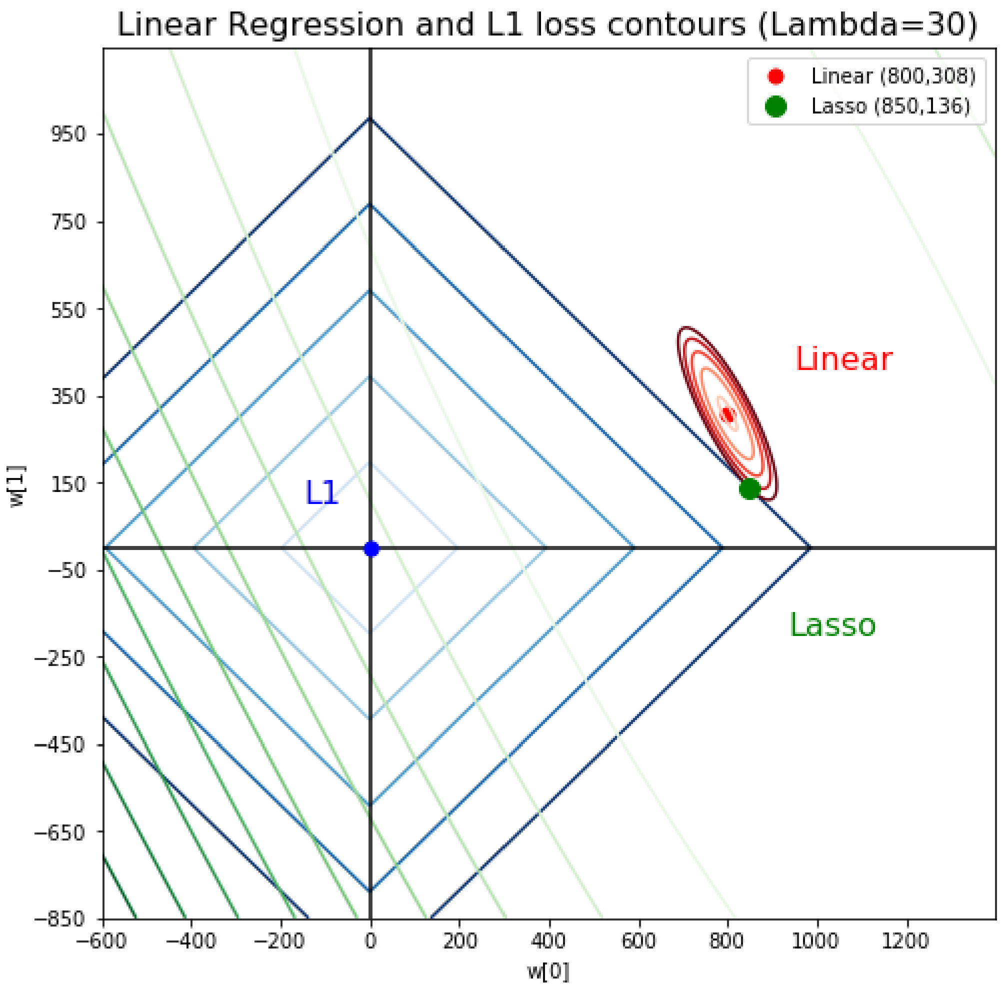
$$\min_{w} \frac{1}{2}\sum_{i=1}^{N}\left(y^{(i)} - w^{\top}x^{(i)}\right)^2, \qquad s.t. \, ||w||_1 \leq c$$

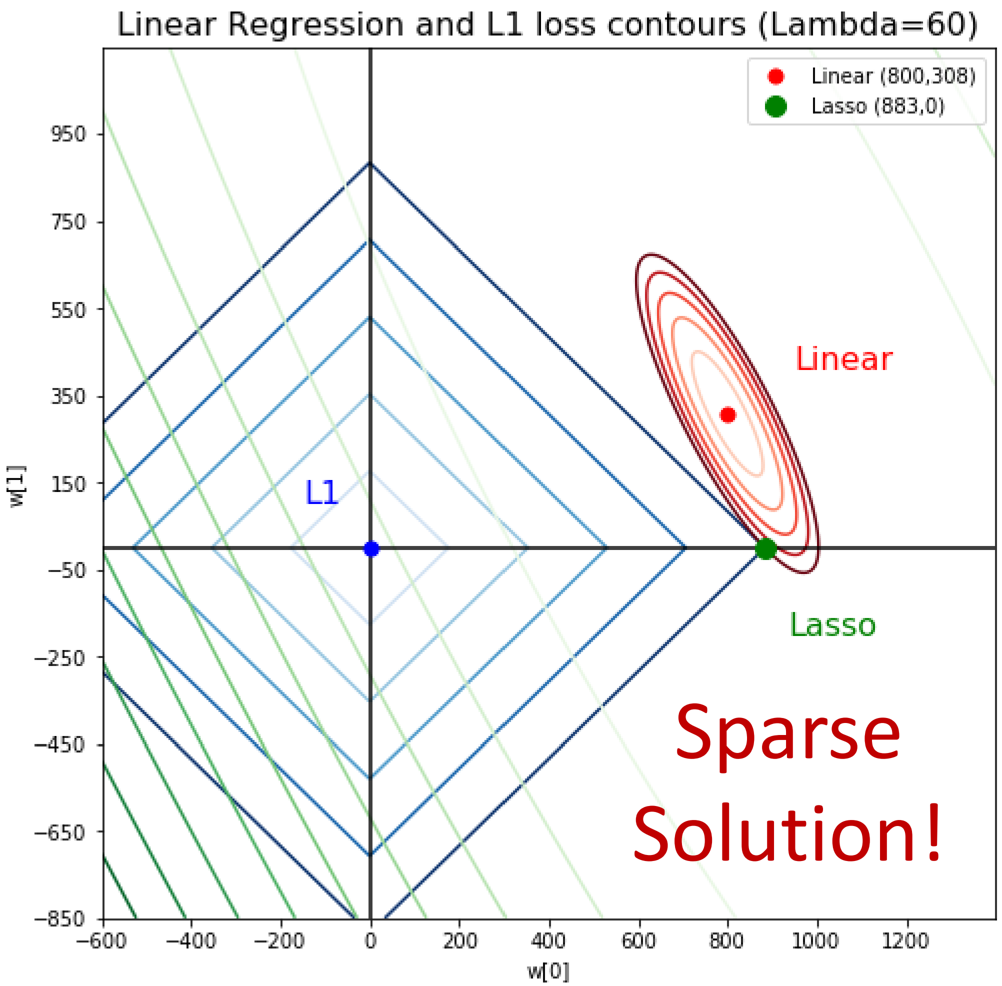
Lasso



[Linear Regression with Regularization (aunnnn.github.io)]

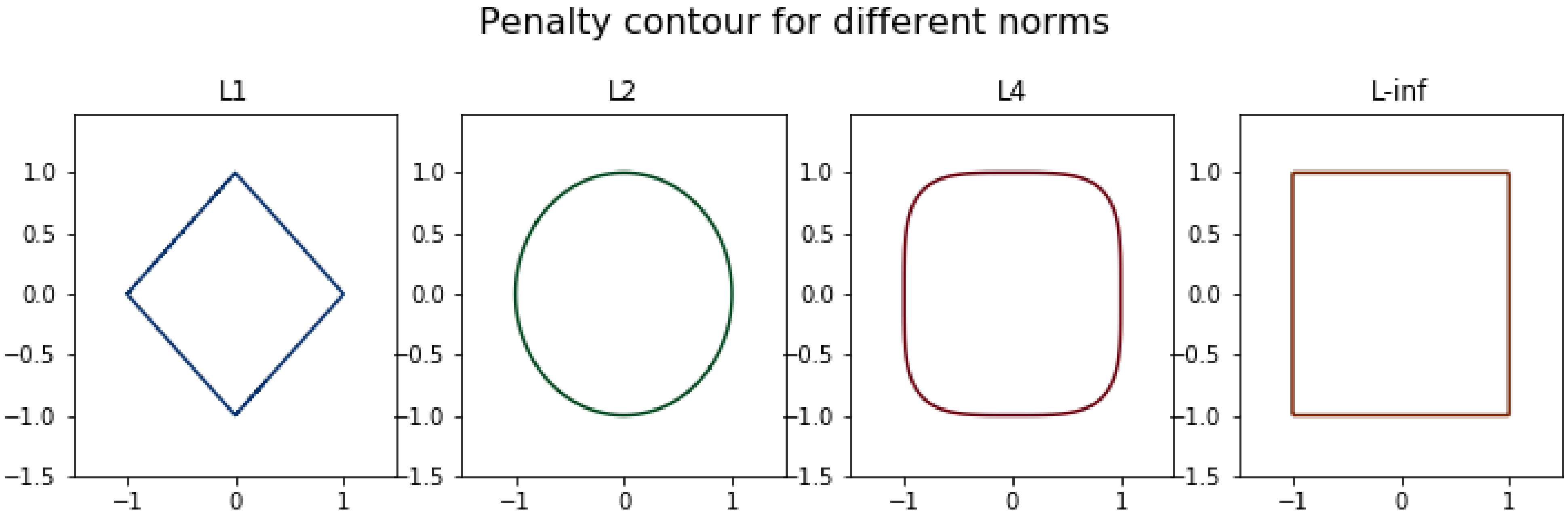
Lasso





[Linear Regression with Regularization (aunnnn.github.io)](aunnnn.github.io)

Various Norms



[Linear Regression with Regularization (aunnnn.github.io)](https://aunnnn.github.io)