# Regularized Linear Regression

# TAs

- **Binghui (Benjamin) Zhang**
  - **Email**: bzhang57@hawk.iit.edu
  - **Office hour**: Wednesday 2:00 PM-3:00PM
  - **Zoom**: https://us05web.zoom.us/j/3420609882
- **Haoran Dai**
  - **Email**: hdai10@hawk.iit.edu
  - **Office hour**: Thursday 2:00 PM-3:00PM
  - **Zoom**: https://iit-edu.zoom.us/j/3761993166
- **Jiawen Wang**
  - **Email**: jwang306@hawk.iit.edu
  - **Office hour**: Tuesday 2:00 PM-3:00PM
  - **Zoom**: https://iit-edu.zoom.us/my/wangjiawen

# Recap

Minimize least square loss

$$\min L(\mathbf{w}) = \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T \mathbf{w}\right)^2$$

$$= \min_{\mathbf{w}} \|\mathbf{X}^T\mathbf{w} - \mathbf{y}\|_2^2$$

Maximize likelihood estimate (MLE)

$$\max \mathcal{L}(\mathbf{w}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\left(y_i - \mathbf{x}_i^T\mathbf{w}\right)^2}{2\sigma^2}\right)$$

$$y_i = \mathbf{x}_i^T\mathbf{w} + \epsilon_i, \forall i \qquad p(\epsilon_i) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\epsilon_i^2}{2\sigma^2}\right)$$

Normal equation: $\mathbf{X}\mathbf{X}^T\mathbf{w} = \mathbf{X}\mathbf{y}$ $\Longrightarrow$ $\mathbf{w}^\star = \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\mathbf{X}\mathbf{y}$

(1) $\mathbf{X}\mathbf{X}^T$ should be full rank ($\#samples > \#features$) — Ridge regression

(2) $\mathbf{w}^*$ : a dense parameter vector (most entries have nonzero values) — LASSO

(3) Overfitting vs. Underfitting — Regularized linear regression

# Underfitting vs Overfitting: Polynomial Curve Fitting

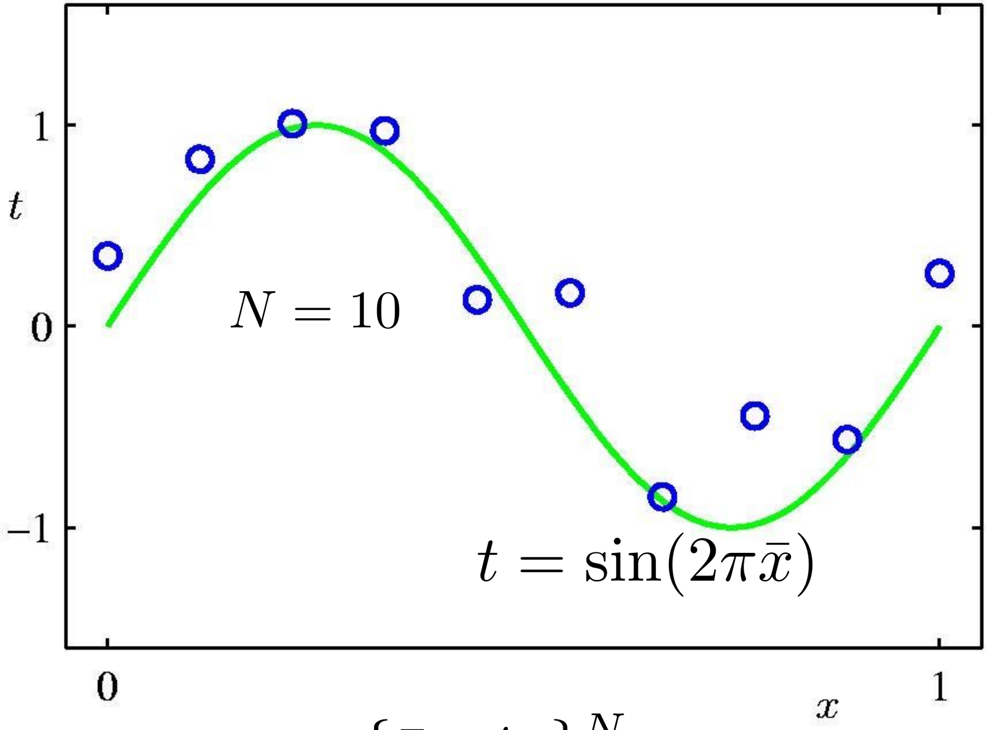
$f(\mathbf{x};\mathbf{w}) = \mathbf{x}^T\mathbf{w} = \sum_{j=0}^{M} w_j x_j$

Let $x_j = \bar{x}^j$

$\mathbf{x} = [1; \bar{x}; \bar{x}^2; \cdots ; \bar{x}^M]$

$f(\mathbf{x};\mathbf{w}) \ = f(\bar{x};\mathbf{w}) = \sum_{j=0}^{M} w_j \bar{x}^j$

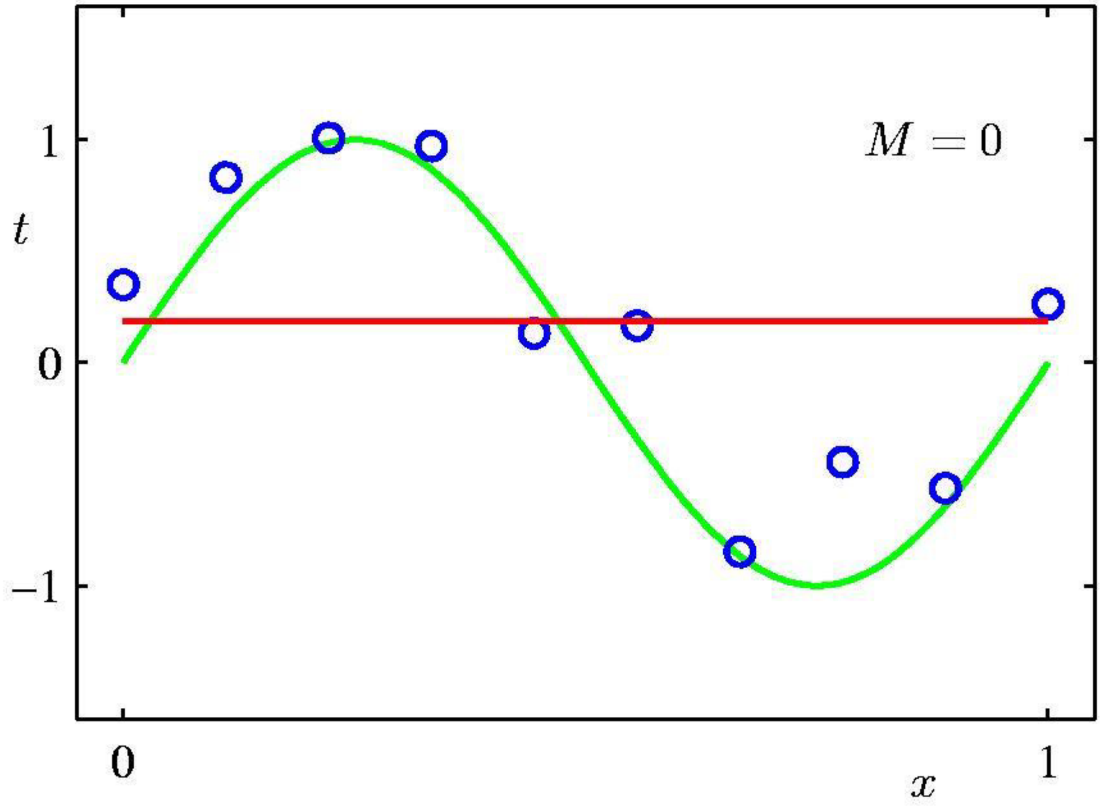
$$\min_{\mathbf{w}} L(\mathbf{w}) = \sum_{n=1}^{N} (f(\bar{x}_n;\mathbf{w}) - t_n)^2$$




$\{\bar{x}_n, t_n\}_{n=1}^{N}$

$t_n = \sin(2\pi\bar{x}_n) + \epsilon_n$

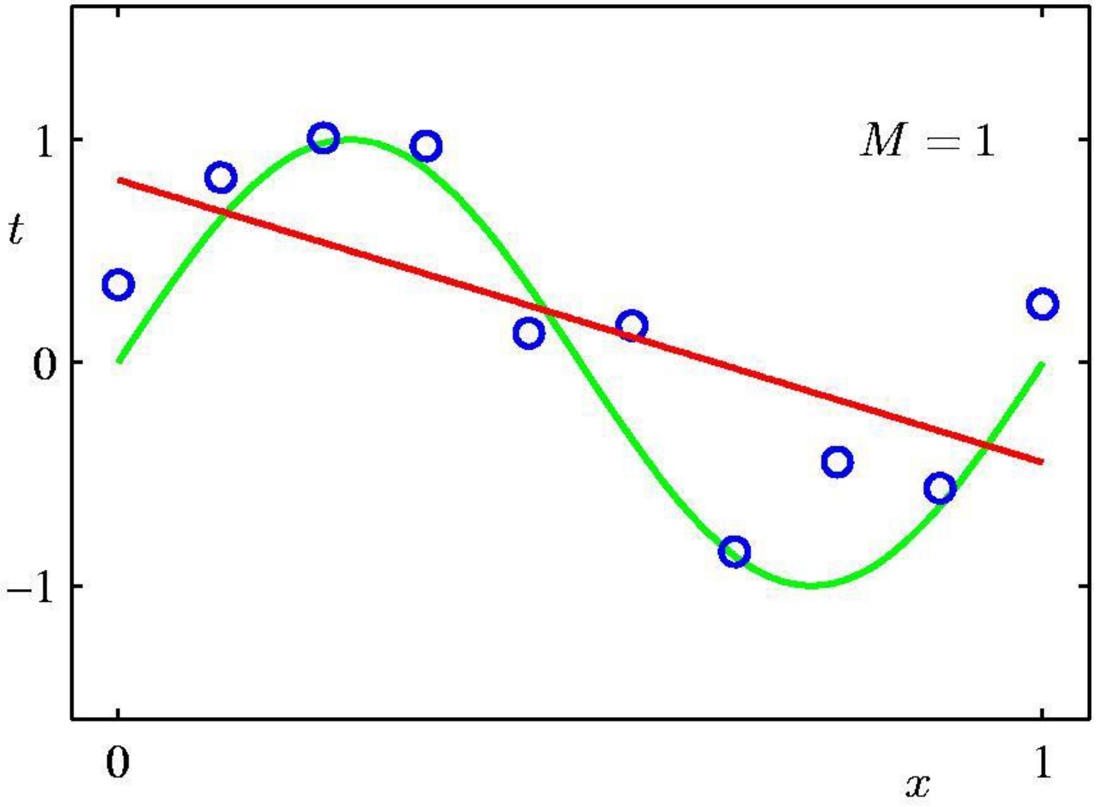
# 0th Order Polynomial (M=0)




$$f(\bar{x}; \mathbf{w}) = w_0$$

Underfitting

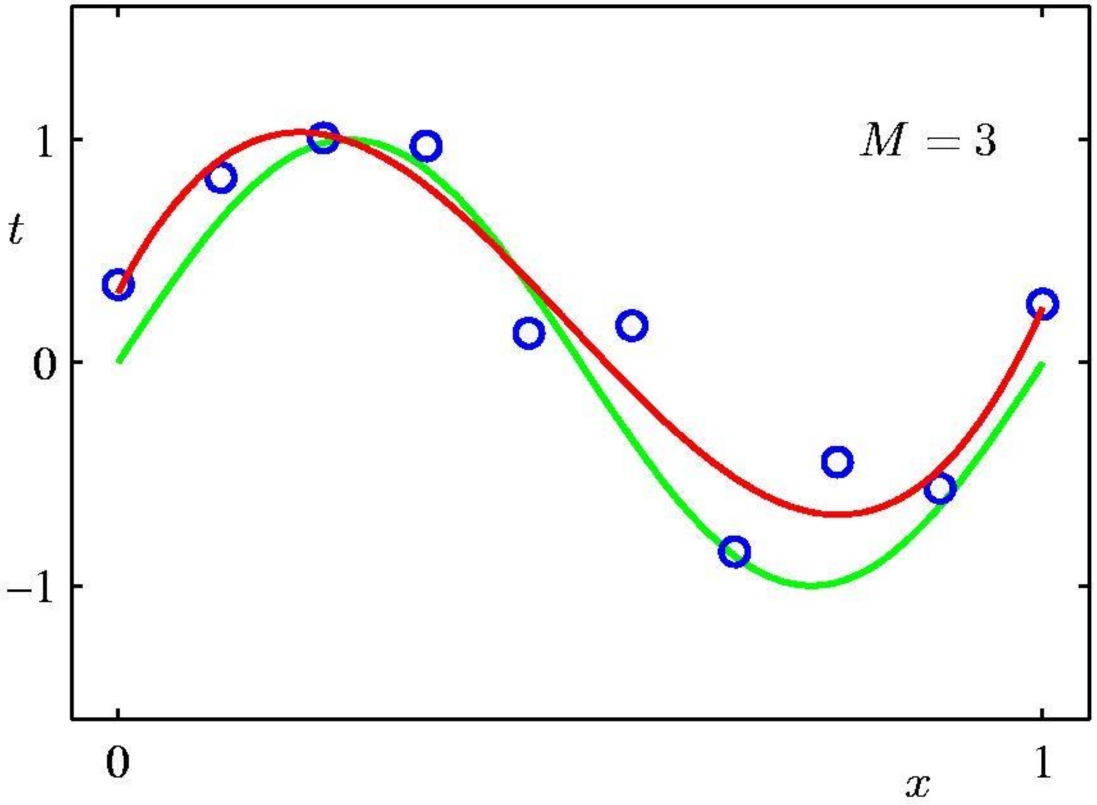
# 1st Order Polynomial (M=1)




$$f(\bar{x}; \mathbf{w}) = w_0 + w_1 \bar{x}$$

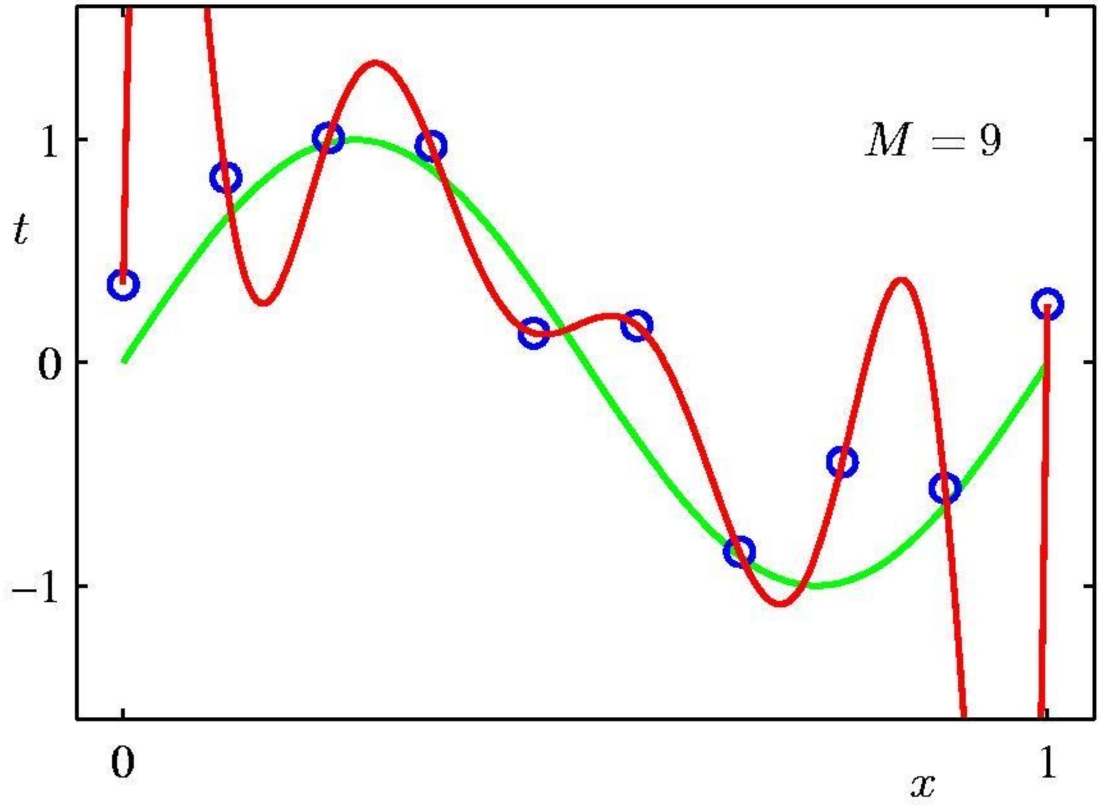
Underfitting

# 3rd Order Polynomial (M=3)




$$f(\bar{x}; \mathbf{w}) = w_0 + w_1\bar{x} + w_2\bar{x}^2 + w_3\bar{x}^3$$ Looks good

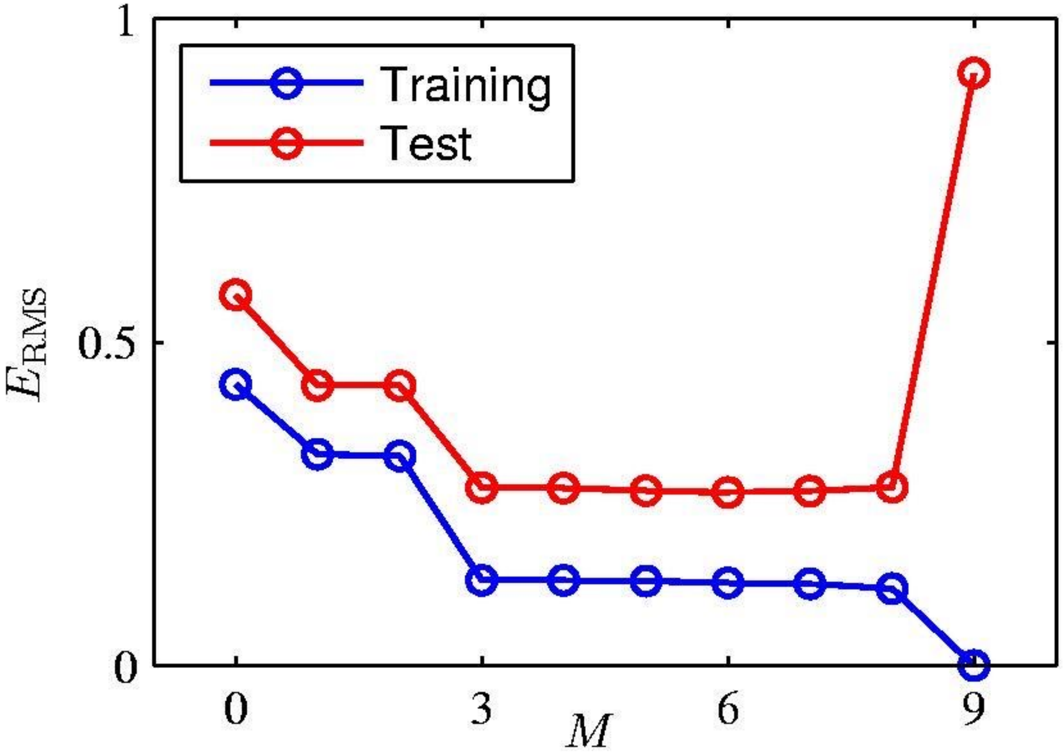
# 9th Order Polynomial (M=9)




$$f(\bar{x}; \mathbf{w}) = w_0 + w_1\bar{x} + \cdots + w_9\bar{x}^9$$

Overfitting

# Underfitting vs. Overfitting




Root-Mean-Square (RMS) Error:

$$\mathrm{E}_{RMS} = \sqrt{2L(\mathbf{w}^{\star})/N}$$

Small M, simple model;
Large training RMSE and large testing RMSE;
Model underfitting

Large M, powerful model;
Small training RMSE, but large testing RMSE;
Model overfitting

# Polynomial Parameter Values

| | $M=0$ | $M=1$ | $M=3$ | $M=9$ |
|---|---|---|---|---|
| $w_0^\star$ | 0.19 | 0.82 | 0.31 | 0.35 |
| $w_1^\star$ | | -1.27 | 7.99 | 232.37 |
| $w_2^\star$ | | | -25.43 | -5321.83 |
| $w_3^\star$ | | | 17.37 | 48568.31 |
| $w_4^\star$ | | | | -231639.30 |
| $w_5^\star$ | | | | 640042.26 |
| $w_6^\star$ | | | | -1061800.52 |
| $w_7^\star$ | | | | 1042400.18 |
| $w_8^\star$ | | | | -557682.99 |
| $w_9^\star$ | | | | 125201.43 |

As M increases, the magnitude of the parameters  becomes larger

- M=9, parameter values are too large!

We expect a powerful model as well as reducing overfitting

Penalize large parameters

- Constraint the norm of parameter vector

# Regularized Linear Regression

Constrained optimization problem

$$\min_{\mathbf{w}} L(\mathbf{w}) = \sum_{n=1}^{N} (f(\mathbf{x}_n; \mathbf{w}) - t_n)^2$$

$$\text{s.t.,} \ \|\mathbf{w}\|_p^p \leq \gamma$$

Lagrangian multiplier

Unconstrained optimization problem

$$\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = L(\mathbf{w}) + \lambda \|\mathbf{w}\|_p^p$$

Regularization term

$$\lambda > 0$$

If $\lambda$ is large, focus on smaller norm of w, but loss L(**w**) may be large (underfitting)

If $\lambda$ is small, focus on reducing L(**w**), but norm of w could be large (overfitting)

- $\lambda$=0 => ordinary least square

$\lambda$ trade-offs loss (underfitting) and norm of w (overfitting)

# Ridge Regression / Penalized Least Square

$$\min_{\mathbf{w}} L(\mathbf{w}) = \sum_{n=1}^{N} (f(\mathbf{x}_n; \mathbf{w}) - t_n)^2 \quad \text{s.t.,} \ \|\mathbf{w}\|_2^2 \leq \gamma$$

Lagrangian multiplier

$$\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = \sum_{n=1}^{N} (f(\mathbf{x}_n; \mathbf{w}) - t_n)^2 + \lambda \|\mathbf{w}\|_2^2$$

L2 Regularization

$$f(\mathbf{x}_n; \mathbf{w}) = \mathbf{x}_n^T \mathbf{w} \qquad = \min_{\mathbf{w}} \|\mathbf{X}^T \mathbf{w} - \mathbf{t}\|_2^2 + \lambda \|\mathbf{w}\|_2^2 \qquad \|\mathbf{w}\|_2 = \sqrt{\sum_i w_i^2}$$

# Ridge Regression: Normal Equation

$$\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = \|\mathbf{X}^T\mathbf{w} - \mathbf{t}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$$

$$= \mathbf{w}^T\mathbf{X}\mathbf{X}^T\mathbf{w} - 2\mathbf{w}^T\mathbf{X}\mathbf{t} + \mathbf{t}^T\mathbf{t} + \lambda\mathbf{w}^T\mathbf{w}$$

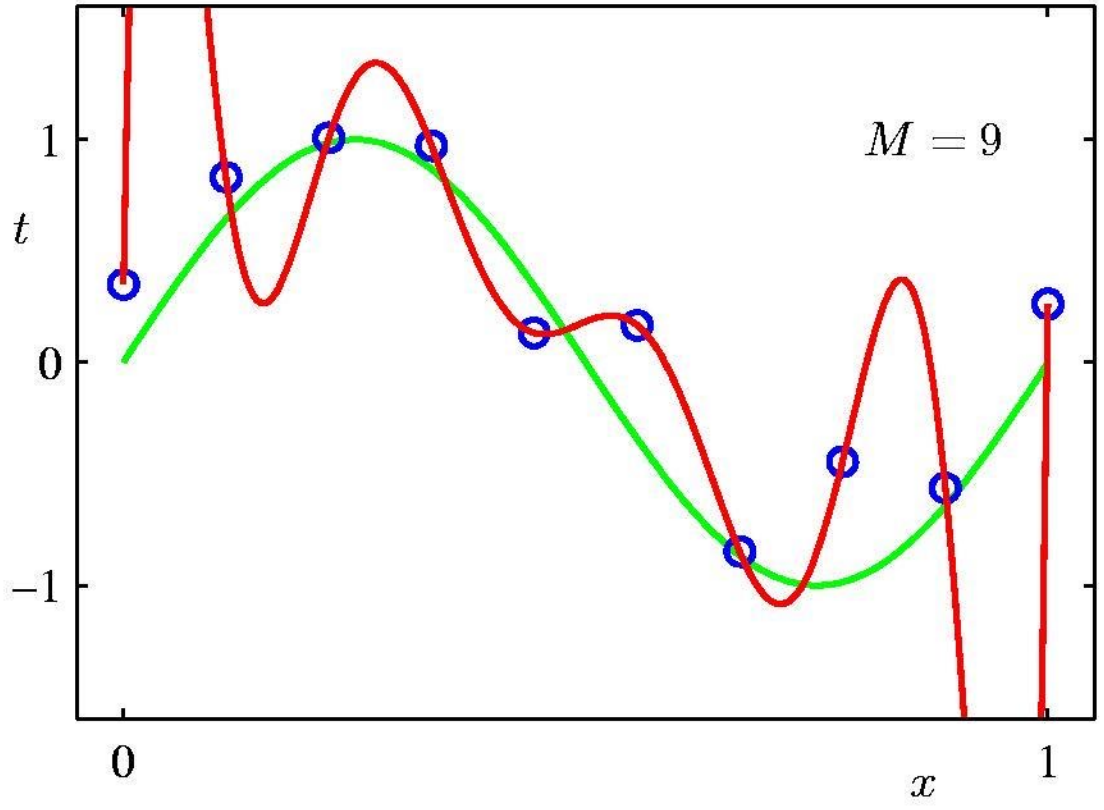
First-order optimality: $\dfrac{\partial \mathcal{L}(\mathbf{w})}{\partial \mathbf{w}} = \mathbf{0}$ $\Longrightarrow$ $2\mathbf{X}\mathbf{X}^T\mathbf{w} - 2\mathbf{X}\mathbf{t} + 2\lambda\mathbf{w} = \mathbf{0}$

Normal equation: $(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})\mathbf{w} = \mathbf{X}\mathbf{t}$ $\Longrightarrow$ $\mathbf{w}^\star = \left(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I}\right)^{-1}\mathbf{X}\mathbf{t}$

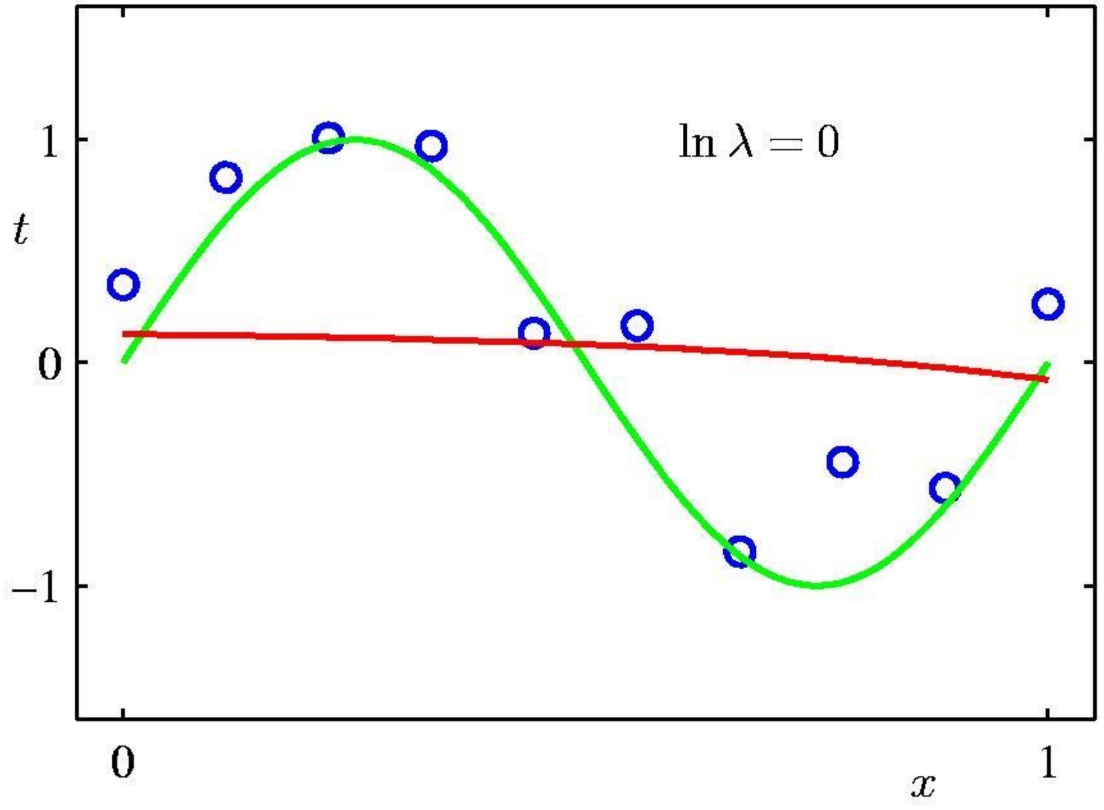
Identity matrix: $\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \mathrm{diag}[1, 1, \cdots, 1]$

$\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I}$ is full rank!

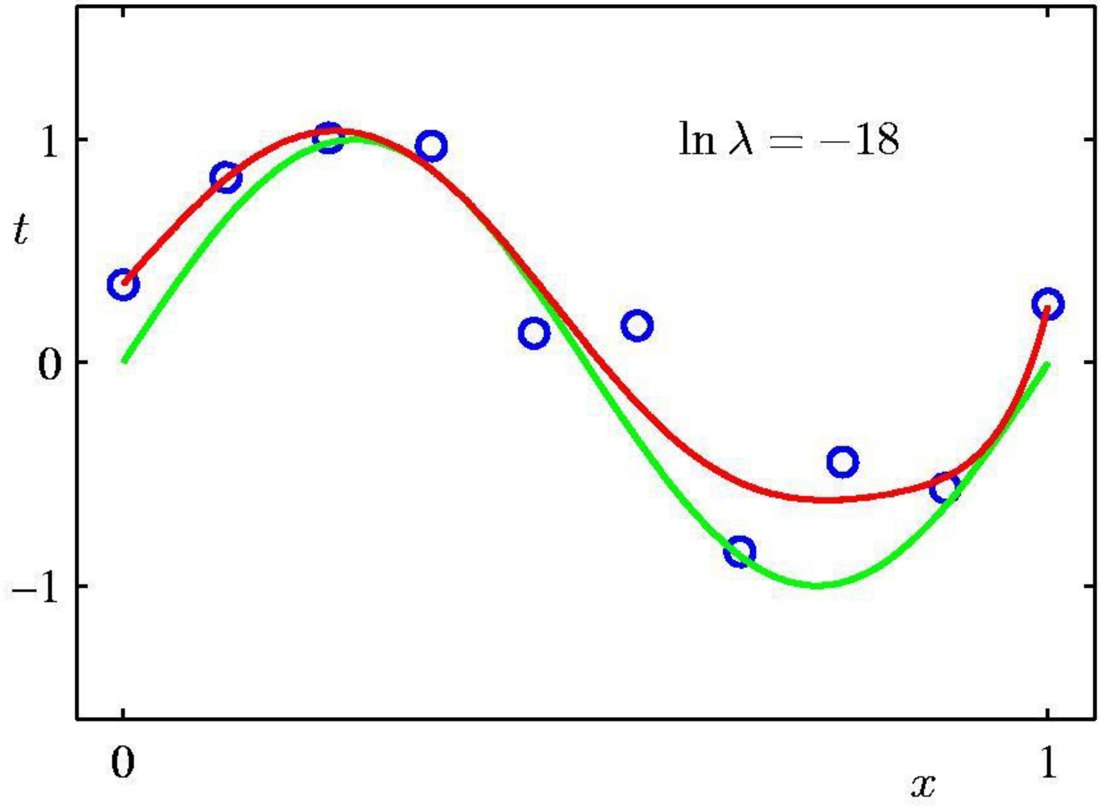
# Regularization (M=9): $\ln \lambda = -\infty$



# Regularization (M=9): $\ln \lambda = 0$

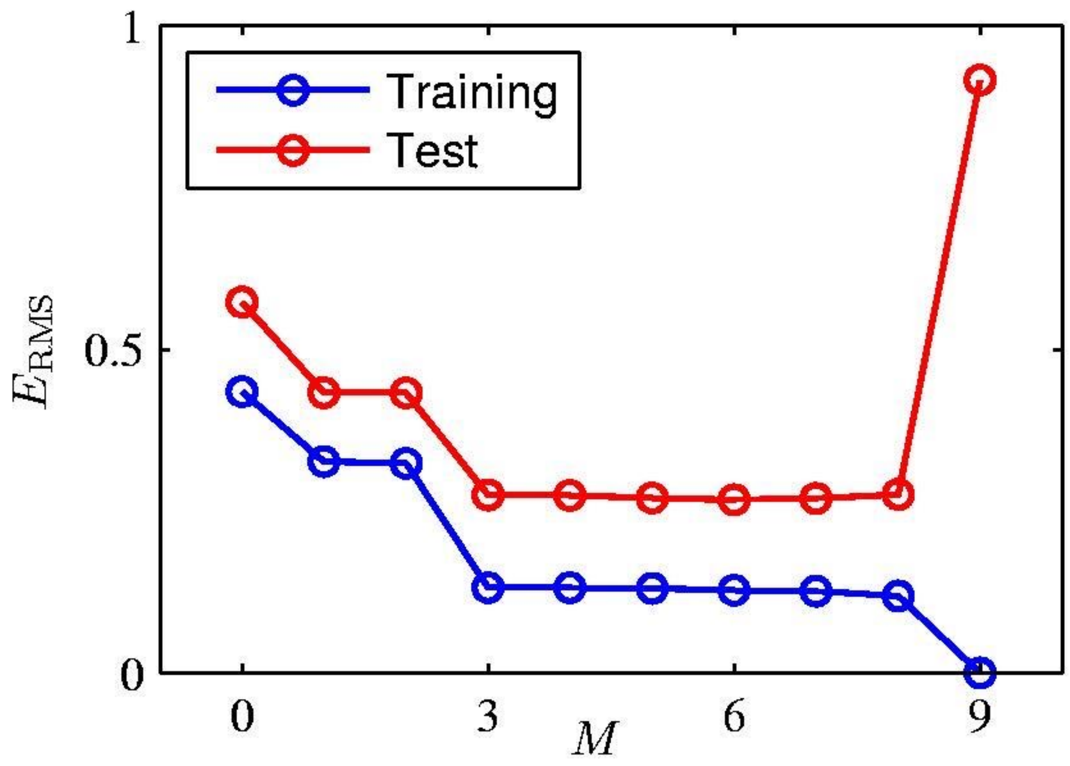


# Regularization (M=9): $\ln \lambda = -18$

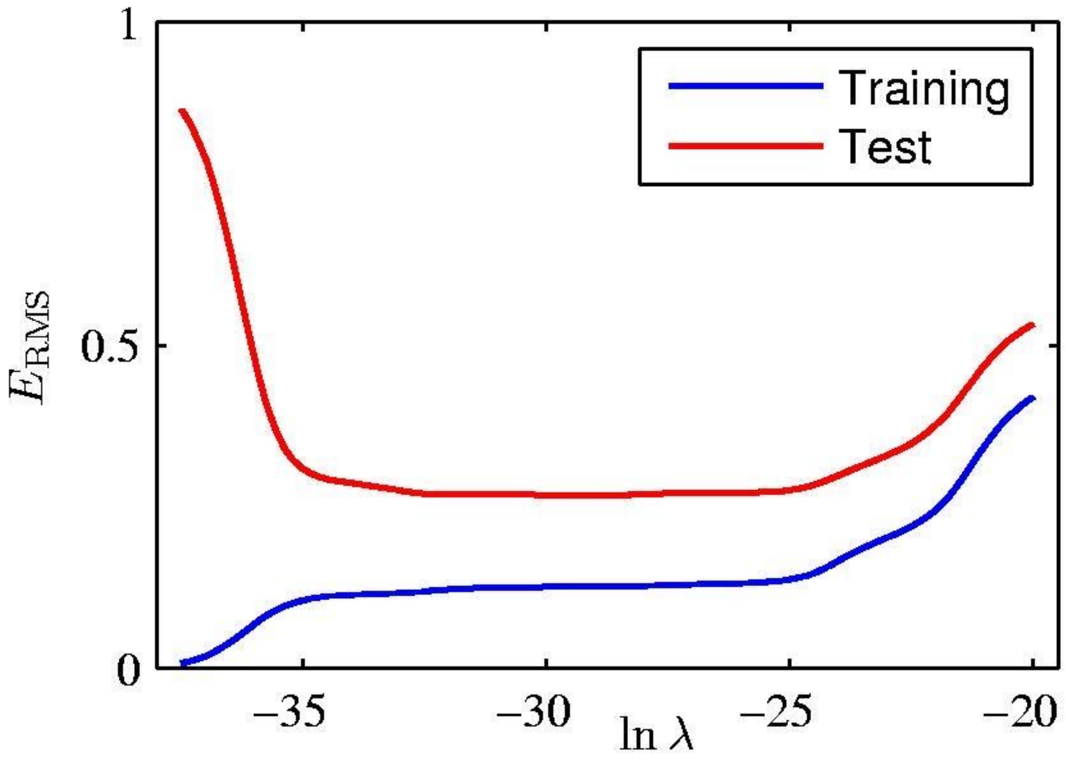


# Underfitting vs. Overfitting

**Ordinary Linear Regression**




**Ridge Regression (M=9)**




Increasing $\lambda$ to certain value reduces overfitting

# Polynomial Parameter Values

| | $\ln \lambda = -\infty$ | $\ln \lambda = -18$ | $\ln \lambda = 0$ |
|---|---|---|---|
| $w_0^\star$ | 0.35 | 0.35 | 0.13 |
| $w_1^\star$ | 232.37 | 4.74 | -0.05 |
| $w_2^\star$ | -5321.83 | -0.77 | -0.06 |
| $w_3^\star$ | 48568.31 | -31.97 | -0.05 |
| $w_4^\star$ | -231639.30 | -3.89 | -0.03 |
| $w_5^\star$ | 640042.26 | 55.28 | -0.02 |
| $w_6^\star$ | -1061800.52 | 41.32 | -0.01 |
| $w_7^\star$ | 1042400.18 | -45.95 | -0.00 |
| $w_8^\star$ | -557682.99 | -91.53 | 0.00 |
| $w_9^\star$ | 125201.43 | 72.68 | 0.01 |

As $\lambda$ increases, the magnitude of the para. gets smaller

Good model produces a dense parameter vector

# LASSO Regression

$$\min_{\mathbf{w}} L(\mathbf{w}) = \sum_{n=1}^{N} (f(\mathbf{x}_n; \mathbf{w}) - t_n)^2 \qquad \text{s.t.,} \; \|\mathbf{w}\|_1 \le \gamma$$

Lagrangian multiplier

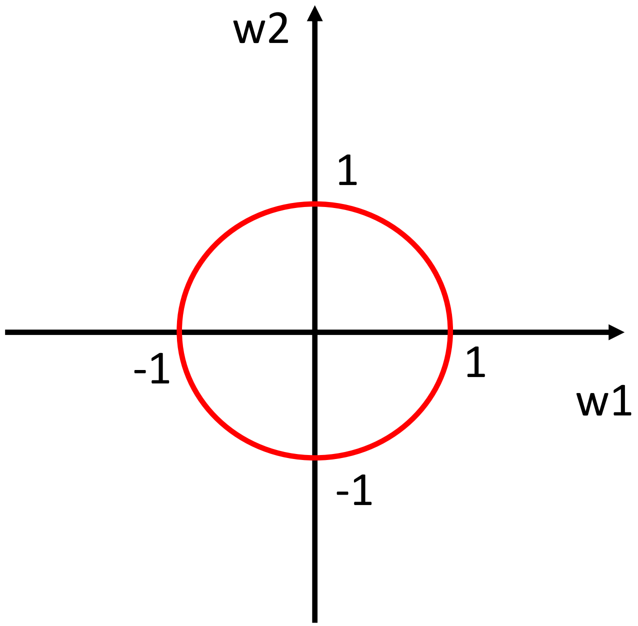
$$\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = \sum_{n=1}^{N} (f(\mathbf{x}_n; \mathbf{w}) - t_n)^2 + \lambda \|\mathbf{w}\|_1$$

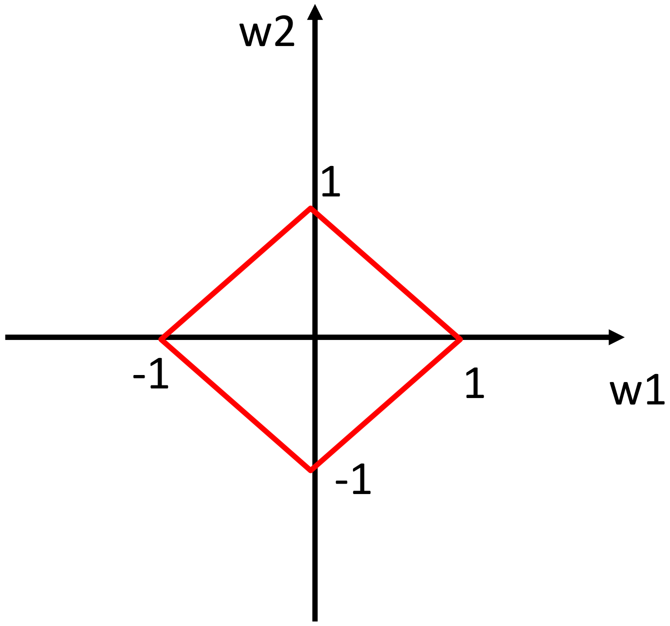
L1 Regularization

$$= \min_{\mathbf{w}} \|\mathbf{X}^T \mathbf{w} - \mathbf{t}\|_2^2 + \lambda \|\mathbf{w}\|_1 \qquad \|\mathbf{w}\|_1 = \sum_i |w_i|$$

# L2 vs. L1 Norm

$$\mathbf{w} = [w_1; w_2]$$



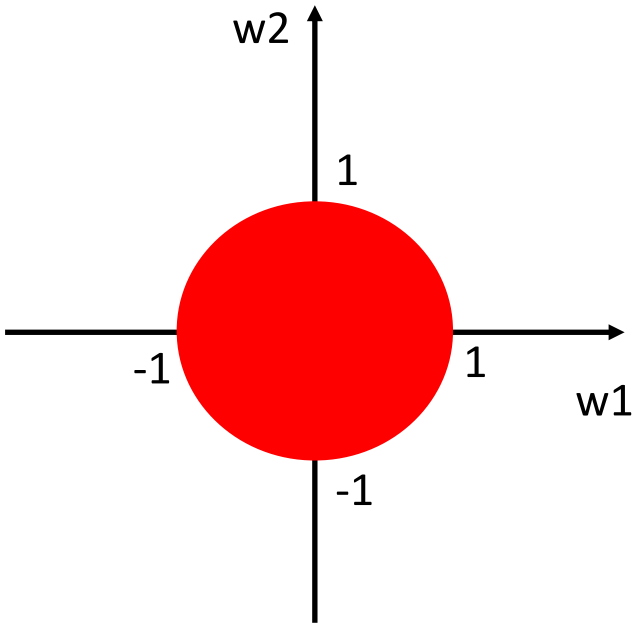




$$\|\mathbf{w}\|_2^2 = w_1^2 + w_2^2 = 1$$

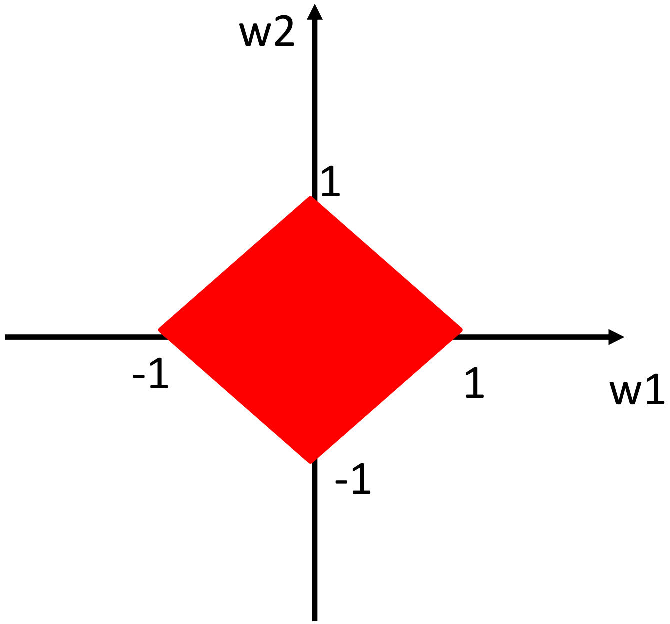
$$\|\mathbf{w}\|_1 = |w_1| + |w_2| = 1$$

# L2 vs. L1 Norm

$$\mathbf{w} = [w_1; w_2]$$

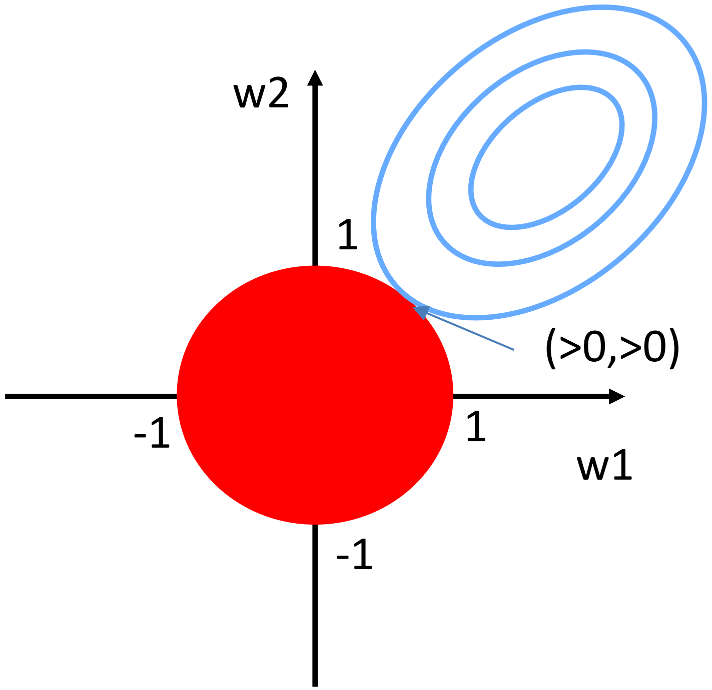


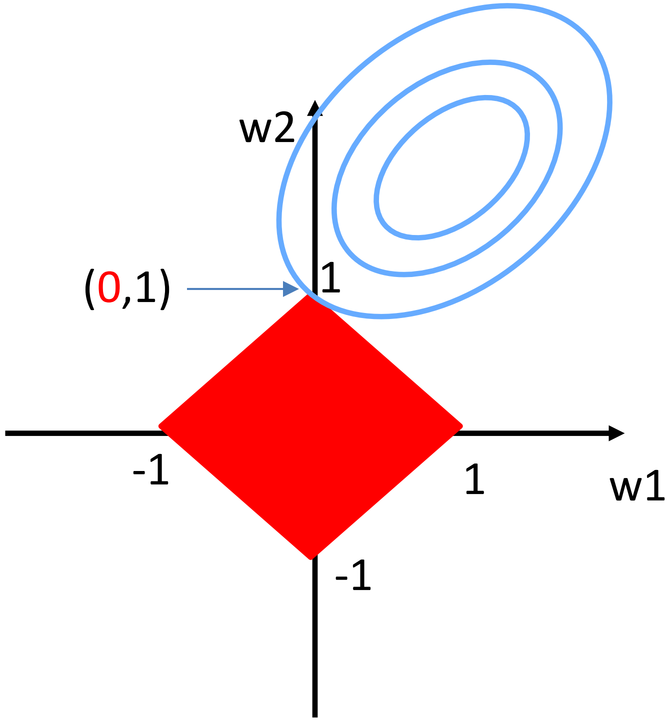

$$\|\mathbf{w}\|_2^2 = w_1^2 + w_2^2 \leq 1$$




$$\|\mathbf{w}\|_1 = |w_1| + |w_2| \leq 1$$

# L1 Regularization Yields Sparse Solutions




$$\min_{w_1,w_2} L(w_1,w_2) = (t - (x_1 w_1 + x_2 w_2))^2$$

$$\text{s.t., } w_1^2 + w_2^2 \le 1$$

$$\min_{w_1,w_2} L(w_1,w_2) = (t - (x_1 w_1 + x_2 w_2))^2$$

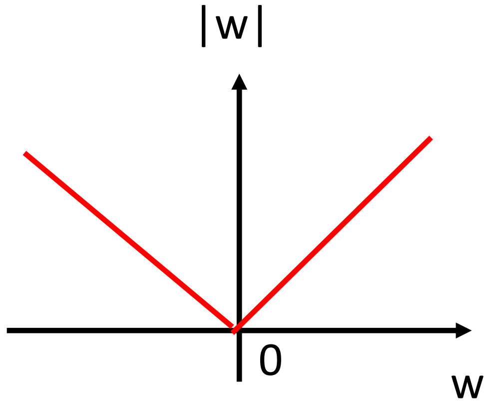
$$\text{s.t., } |w_1| + |w_2| \le 1$$

# Optimality Condition for LASSO

$$\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = L(\mathbf{w}) + \lambda \|\mathbf{w}\|_1$$

First-order optimality:

$$\partial_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = \partial_{\mathbf{w}} L(\mathbf{w}) + \lambda \partial_{\mathbf{w}} \|\mathbf{w}\|_1 = \mathbf{0}$$




However, $\|\boldsymbol{w}\|_1$ is not differentiable when $w_j = 0$

**Non-smooth/differentiable** optimization problem

Gradient/Derivative $\Rightarrow$ Subgradient/Subderivative

# Gradient, Convex & Differentiable Function

A differentiable function *f* is convex, *if*

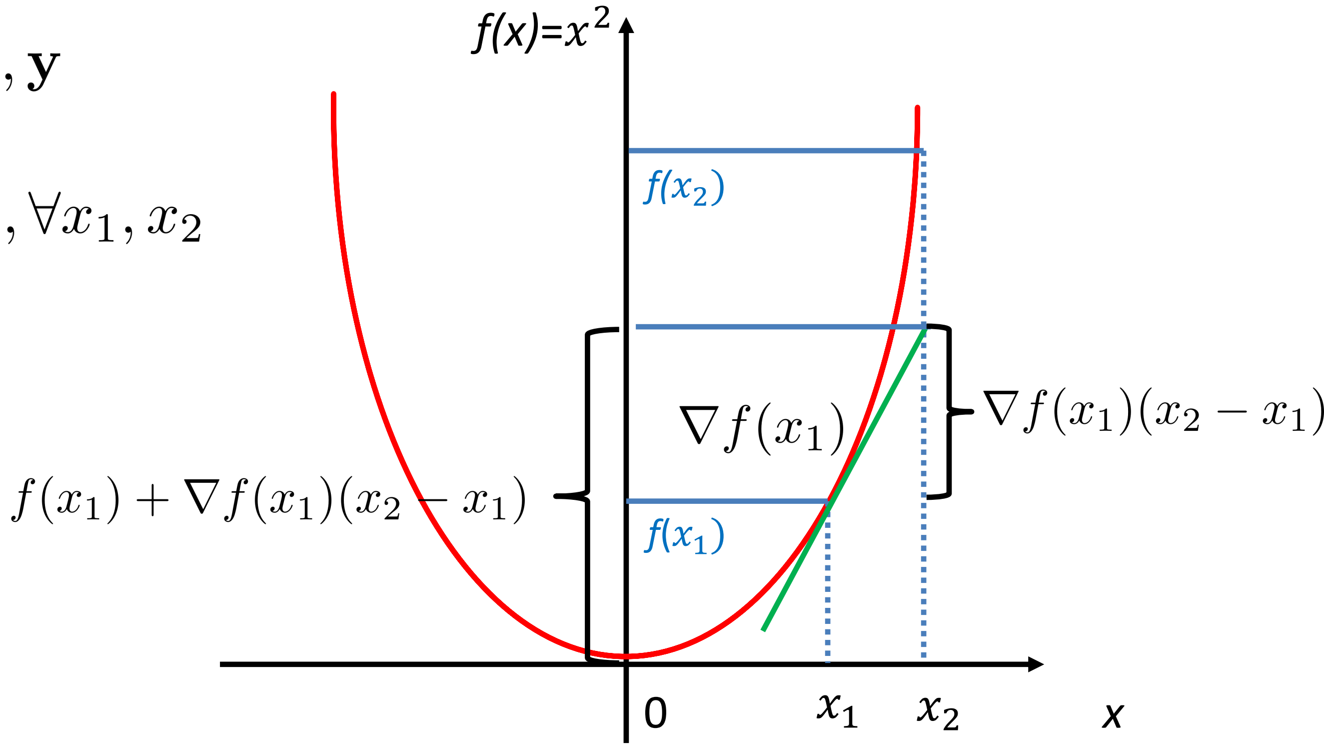
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}),\ \forall \mathbf{x}, \mathbf{y}$$

1-dim: $f(x_2) \geq f(x_1) + \nabla f(x_1)\cdot(x_2 - x_1), \forall x_1, x_2$

$\nabla f(x_1) = 2x_1, f(x_1) = x_1^2, f(x_2) = x_2^2$

$$\begin{aligned}\text{R.H.S.} &= f(x_1) + \nabla f(x_1)\cdot(x_2-x_1)\\ &= x_1^2 + 2x_1\cdot(x_2-x_1)\\ &= 2x_1x_2 - x_1^2\end{aligned}$$

L.H.S - R.H.S. $= x_2^2 - (2x_1x_2 - x_1^2) = (x_2 - x_1)^2 \geq 0$




$f(x) = x^2$

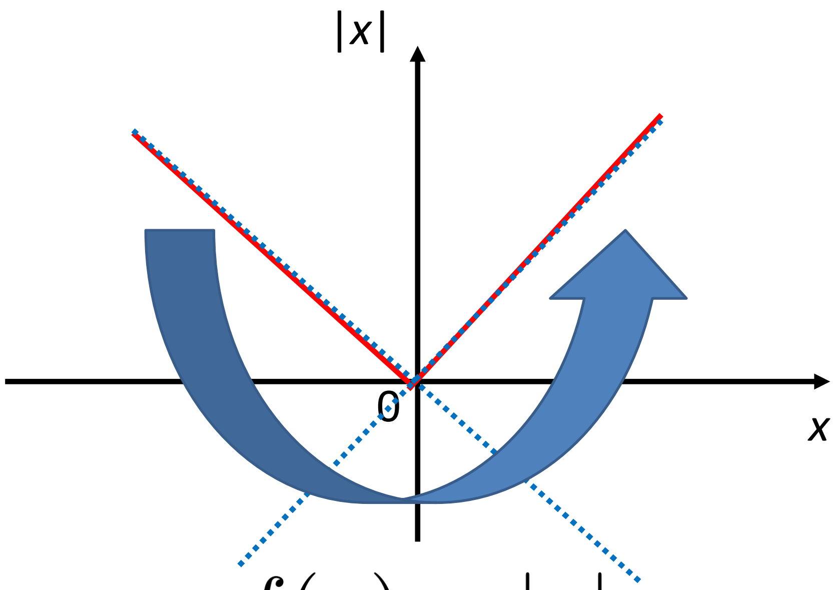
# Subgradient, Convex & Non-Diff. Function

A subgradient $\partial f(x)$ of a convex function $f$ at $x$ is any $g \in R^n$ such that

$$f(\mathbf{y}) \geq f(\mathbf{x}) + g^T(\mathbf{y} - \mathbf{x}), \forall \mathbf{y}$$

$$\text{1-dim: } f(y) \geq f(x) + g \cdot (y - x), \forall y$$

$$|y| - |x| \geq g \cdot (y - x), \forall y$$




$$f(x) = |x|$$

$$\text{If } x > 0: \; |y| - gy \geq (1 - g)x, \forall y \;\Rightarrow\; g = 1$$

$$\text{If } x < 0: \; |y| - gy \geq -(1 + g)x, \forall y \;\Rightarrow\; g = -1$$

$$\text{If } x = 0: \; |y| \geq gy, \forall y \;\Rightarrow\; g \in [-1, 1]$$

$$\partial_x |x| = \begin{cases} -1, \text{ if } x < 0 \\ [-1, 1], \text{ if } x = 0 \\ 1, \text{ if } x > 0 \end{cases}$$

# Soft Thresholding Algorithm

$$\partial_{w_j}\mathcal{L}(\mathbf{w}) = \partial_{w_j}L(\mathbf{w}) + \lambda\partial_{w_j}\|\mathbf{w}\|_1$$

$$\frac{\partial L(\mathbf{w})}{\partial w_j} = a_j w_j - c_j$$

$$a_j = 2\|\mathbf{X}_{j,:}\|_2^2$$

$$c_j = 2\langle \mathbf{X}_{j,:}, \mathbf{t} - \mathbf{X}_{-j,:}\mathbf{w}_{-j}\rangle$$

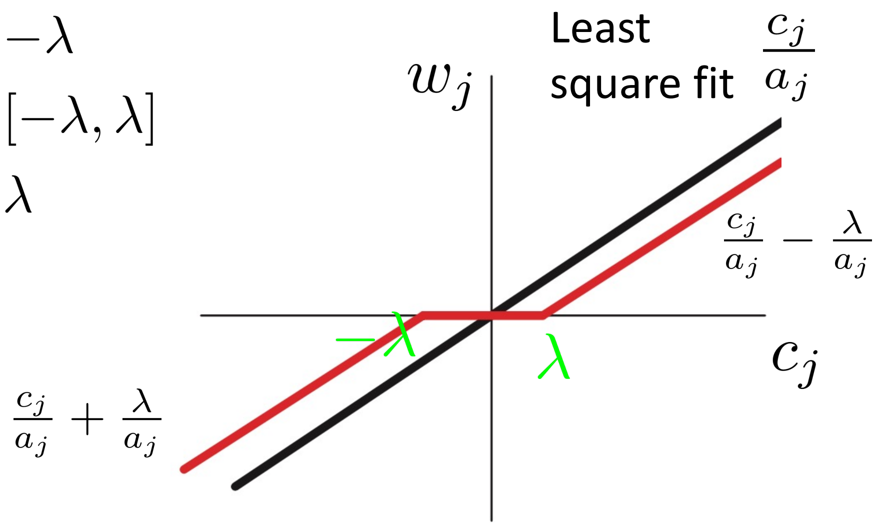
$$\partial_{w_j}\|\mathbf{w}\|_1 = \begin{cases} -1, \text{ if } w_j < 0 \\ [-1,1], \text{ if } w_j = 0 \\ 1, \text{ if } w_j > 0 \end{cases}$$

$$= \begin{cases} a_j w_j - (c_j + \lambda), \text{ if } w_j < 0 \\ [-c_j - \lambda, -c_j + \lambda], \text{ if } w_j = 0 \\ a_j w_j - (c_j - \lambda), \text{ if } w_j > 0 \end{cases} = 0$$

$$w_j = \begin{cases} (c_j + \lambda)/a_j < 0, \text{ if } c_j < -\lambda \\ 0, \text{ if } c_j = [-\lambda, \lambda] \\ (c_j - \lambda)/a_j > 0, \text{ if } c_j > \lambda \end{cases}$$

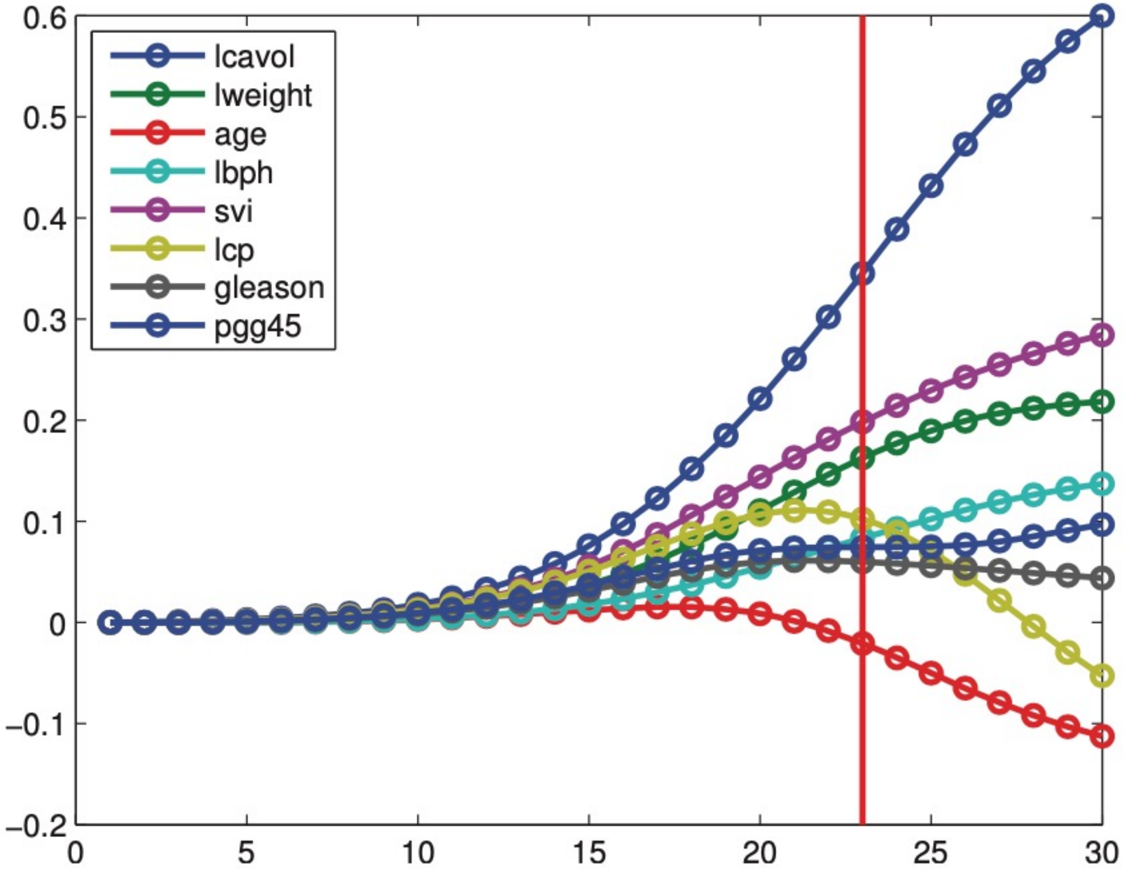
$$= \text{soft}\left(\frac{c_j}{a_j}; \frac{\lambda}{a_j}\right)$$

$$\text{soft}(a;\delta) := \text{sign}(a)(|a| - \delta)_+$$



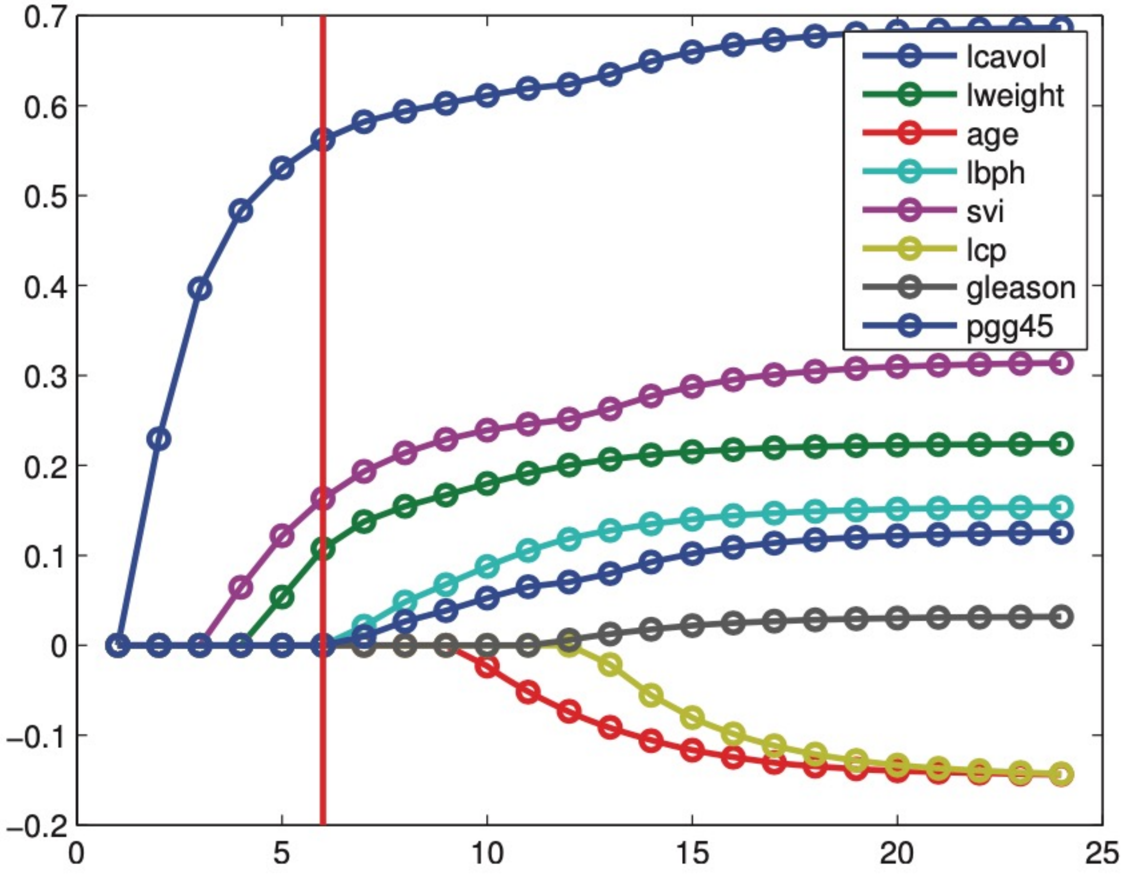

LASSO: biased estimator

# Regularization Path




Dense solution

Ridge Regression




Sparse solution

LASSO

# Comparison: Least Square, Ridge & LASSO

**Assuming data features orthonormal**

$$\mathbf{X}\mathbf{X}^T = \mathbf{I}_D \qquad \|\mathbf{X}_{j,:}\|_2^2 = 1, \forall j$$

Ordinary least square $\quad \mathbf{w}^{OLS} = \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\mathbf{X}^T\mathbf{t} = \mathbf{X}^T\mathbf{t} \qquad w_j^{OLS} = \frac{c_j}{a_j}$

Ridge regression $\quad \mathbf{w}^{Ridge} = \left(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I}_D\right)^{-1}\mathbf{X}^T\mathbf{t} = \frac{1}{1+\lambda}\mathbf{X}^T\mathbf{t}$

$$\mathbf{w}^{Ridge} = \frac{1}{1+\lambda}\mathbf{w}^{OLS}$$

Scaled (biased) estimator

LASSO regression $\quad a_j = 2\|\mathbf{X}_{j,:}\|_2^2 \; = 2 \qquad w_j = \mathrm{soft}\left(\frac{c_j}{a_j}; \frac{\lambda}{a_j}\right) \; = \mathrm{soft}\left(w_j^{OLS}; \frac{\lambda}{2}\right)$

$$\mathbf{w}^{LASSO} = \mathrm{soft}\left(\mathbf{w}^{OLS}; \frac{\lambda}{2}\right)$$

Biased estimator

# Acknowledgement

Some slides are from **Christ Bishop**

**PATTERN RECOGNITION AND MACHINE LEARNING**

https://www.microsoft.com/en-us/research/wp-content/uploads/2016/05/prml-slides-1.pdf