. To deal with overfitting

. So called Ridge Regression

Truth:

- Data may have outliers, they will pull the line away from main trend line to minimize squared error

L2 Regularization:

. Penalize large weights by adding their squared magnitude to the cost:

$$\boxed{J = \sum (y_n - \hat{y})^2 + \lambda |w|^2}$$

$$|w|^2 = w^T \cdot w = w_1^2 + w_2^2 + \dots w_D^2$$

. Probabilistic perspective:

. $P(\text{data} \mid w)$: likelihood

$$\boxed{\exp(J) = \left[\prod \exp\left\{-(y_n - w^T x_n)^2\right\}\right] \cdot \exp\left\{-\lambda \cdot w^T w\right\}}$$

↳ new Gaussian: w is the random variable

↳ likelihood

$P(Y \mid X, w)$
$\sim \mathcal{N}(w^T x_n, \sigma^2)$

↳ Prior : $P(w)$
$= \frac{\lambda}{\sqrt{2\pi}} \exp\left\{-\frac{\lambda}{2} w^T \cdot w\right\}$
$\sim \mathcal{N}(0, \frac{1}{\lambda})$

. Solving for w:

$$J = (Y - Xw)^T \cdot (Y - Xw) + \lambda w^T \cdot w$$

take $\frac{\partial J}{\partial w} = 0 \rightarrow \boxed{w = (\lambda I + X^T X)^{-1} X^T \cdot Y}$