Regularization

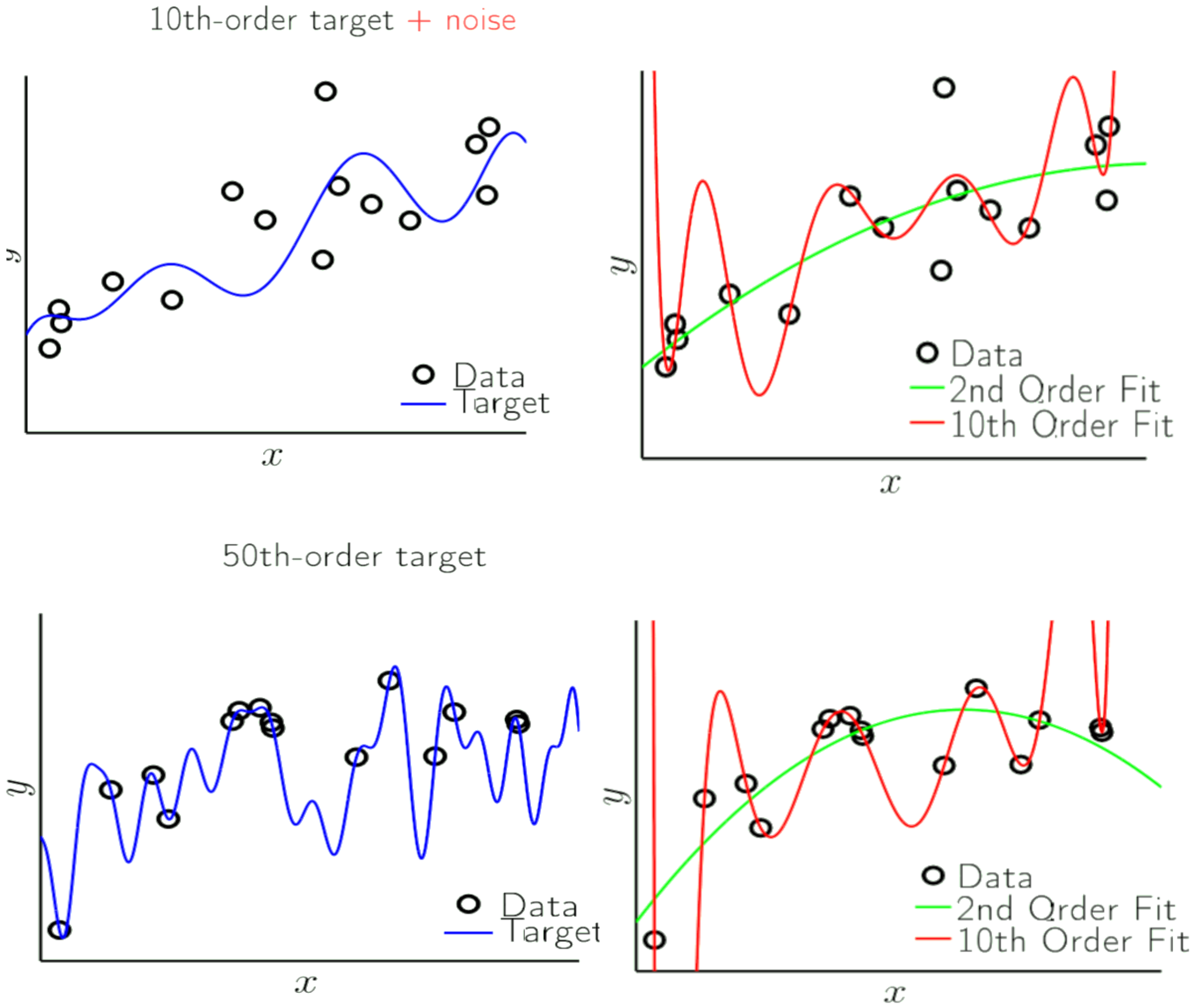
Anahita Zarei, Ph.D.

Reading

- Learning from Data: 4.2

Review: Bias-Variance and Noise

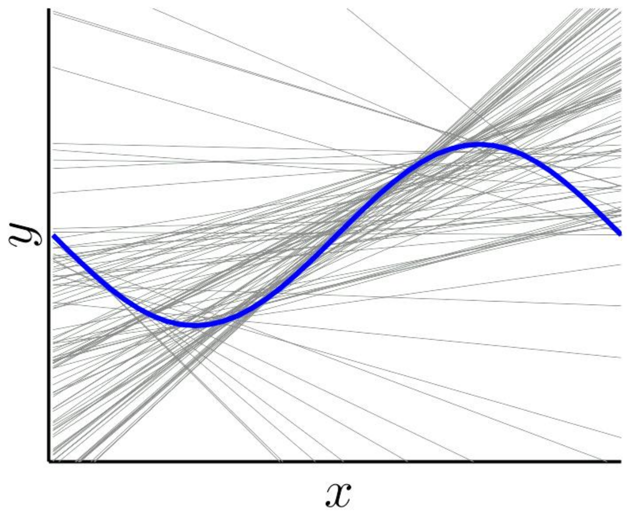
- $Total\ Error = Bias^2 + Var + \sigma^2$
- The underlying root of overfitting is:
  - Stochastic Noise
  - Deterministic Noise
- When you fit noise, you extrapolate out of sample to a pattern that does not exists and therefore take you away from the target function.



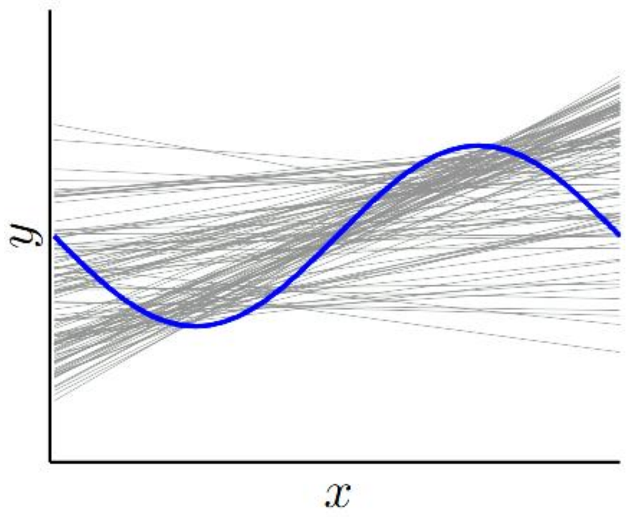
Regularization

- Without regularizations, the lines have high variance that directly results in overfitting.
- With regularization, we improve the variance by restricting the parameters (slope and intercept) of the lines.
- In doing so, we sacrifice the perfect fit on the training set (i.e. we increase the bias)
- Therefore, the new lines are not as crazy, but they don’t fit the two points perfectly because they’re under constraint.

$without\ regularization$



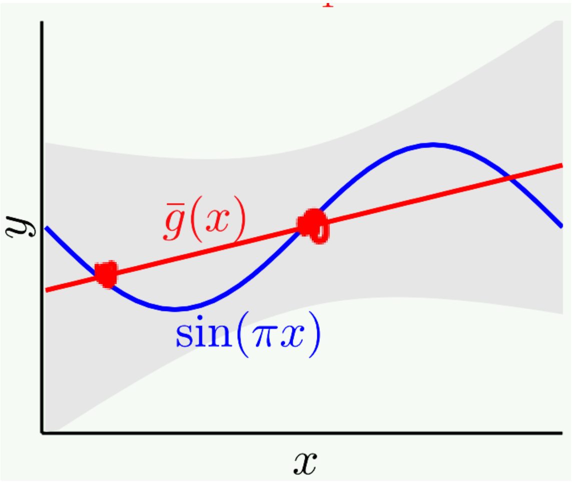
$with\ regularization$



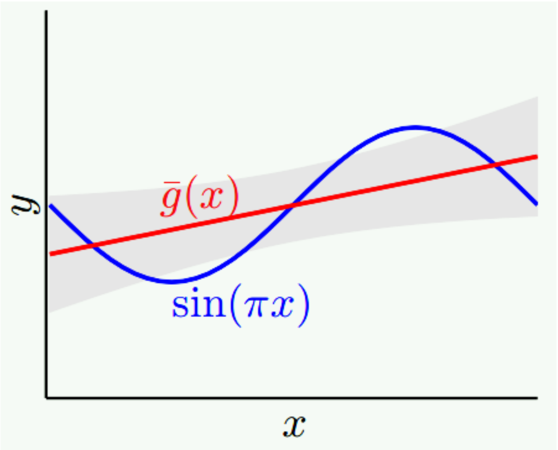
Effect of Regularization on Bias and Variance

- Without regularization:
  - The average hypothesis isn't too bad.
  - But depending on what two points you get, the variance can have a severe effect on total error.
- With regularization:
  - The average hypothesis isn't as perfect. There is a bit of added bias.
  - Reduction in variance is dramatic.
- The regularized linear model outperforms the constant model.
- Regularization works as an intermediate step between extremely restricted and extremely unrestricted.

without regularization



with regularization



Bias =?
Variance = ?

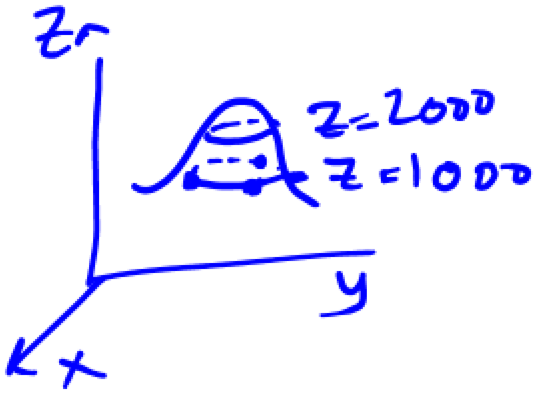
Unconstrained Solution: Review of Least Squares

- Given the training set $\left(x^{(1)}, y^{(1)}\right) \ldots \left(x^{(N)}, y^{(N)}\right)$, find the least square solution:
- $\begin{bmatrix} 1 & \cdots & x_d^{(1)} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_d^{(N)} \end{bmatrix} \begin{bmatrix} w_0 \\ \vdots \\ w_d \end{bmatrix} \approx \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{bmatrix}$
- $E_{in} = MSE = \frac{1}{N} \sum_{n=1}^{N} (w^T x^{(n)} - y^{(n)})^2$
- $E_{in} = \frac{1}{N} (Xw - y)^T (Xw - y)$

Gauss-Markov Theorem: BLUE

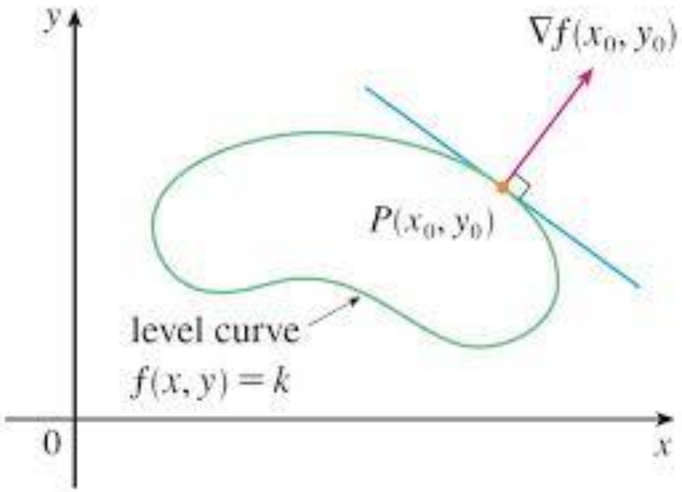
- The (Ordinary Least Square) OLS coefficients are **Best** Linear **Unbiased** Estimates.
- **Unbiased:** If you are given many different data sets D_k and use OLS to find beta coefficients for each set, then average of those betas will be equal to the population parameter beta.
- $E\left[\hat{\beta}\right] = \beta$ => $\hat{\beta}$ is an unbiased estimator of β.
- **Best:** The coefficients derived from OLS have smallest variance than any other unbiased coefficients (calculated using criteria other than minimizing OLS.)
- However, if we're willing to sacrifice a little bias, we'll be able to find coefficients that have lower variance than OLS.

Gradient Review

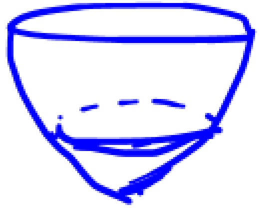


- Gradient is the vector of partial derivatives.
- Direction of gradient shows the direction of maximum change.
- It’s perpendicular to the level curves. (why?)
- Direction of gradient shows the direction of maximum change. Therefore, it has to be perpendicular to the level curves. Because if it is not, then it has a component on the tangent to the level curve. We know that the height on the level curve doesn’t change. So then gradient won’t be pointing to the direction of max change. Hence, it has to be perpendicular.
- It can be evaluated at a given point A = (x0,y0)
- Magnitude of $\nabla f\,(x_0, y_0)$ denotes the slope of the plane tangent to the surface at that point. So it shows how steep the surface is.

$$\|\nabla f(x_0, y_0)\| = \text{slope}$$

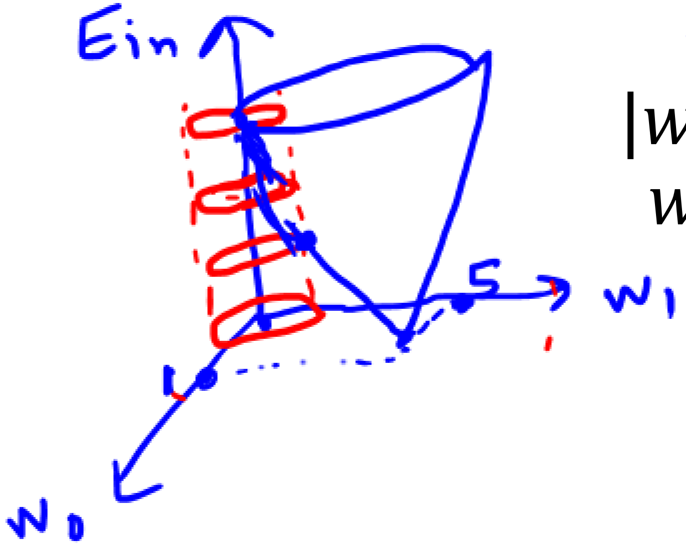


Constrained Solution



Minimize $SSE = \overbrace{(Xw - y)^T(Xw - y)}$ *subject to* ?

What would be a good choice?



$$\cancel{w_0 + w_1 + \cdots w_{d+1} \le c}$$

$$|w_0| + |w_1| + \cdots |w_{d+1}| \le c$$

$$w_0{}^2 + w_1{}^2 + \cdots w_{d+1}{}^2 \le c$$

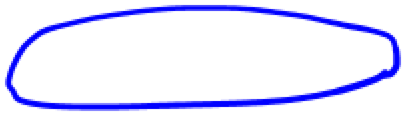
$$w^T w \le C$$

$$[w_0, w_1] \cdot \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = w_0^2 + w_1^2$$

$$\boxed{w_0^2 + w_1^2 \le C}$$

0.5

$$\begin{bmatrix} 1 \\ 5 \end{bmatrix}$$



Constrained Solution

In this figure, w isn't optimal, because gradient has some nonzero component along the circle, and by moving a small amount in the opposite direction of this component we can improve Ein, while still remaining on the circle.

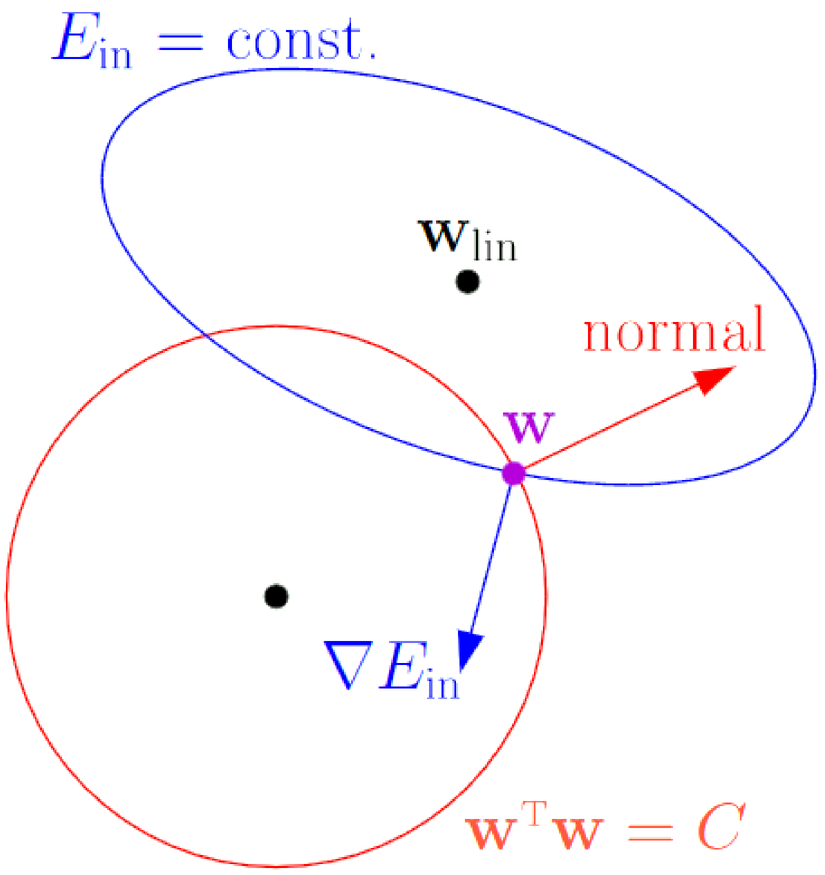
$$\nabla E_{\text{in}}(\mathbf{w}_{\text{reg}}) \propto -\mathbf{w}_{\text{reg}}$$

$$= -2\frac{\lambda}{N}\mathbf{w}_{\text{reg}}$$

$$\nabla E_{\text{in}}(\mathbf{w}_{\text{reg}}) + 2\frac{\lambda}{N}\mathbf{w}_{\text{reg}} = \mathbf{0}$$

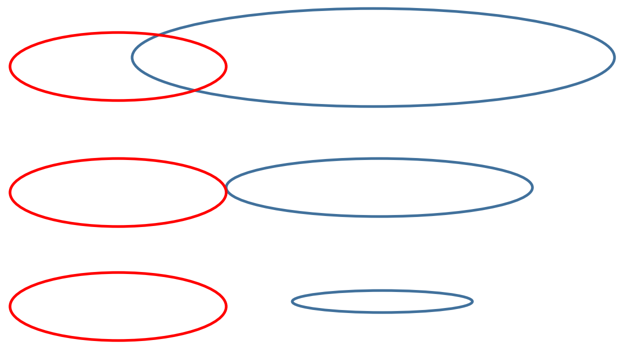
Minimize $\quad E_{\text{in}}(\mathbf{w}) + \frac{\lambda}{N}\mathbf{w}^{\mathsf{T}}\mathbf{w}$

The above expression provides equivalence of the constrained expression in the previous slide.



Relationship between lambda and C

- C and lambda are inversely proportional:
- When c is really large, then W_{ls} is the solution because the circle already contains W_{ls} which is consistent with minimizing Ein, only. That happens when lambda is really small.
- When c is really small regularization is severe. This happens when lambda is really large to put emphasis on the constraint part.

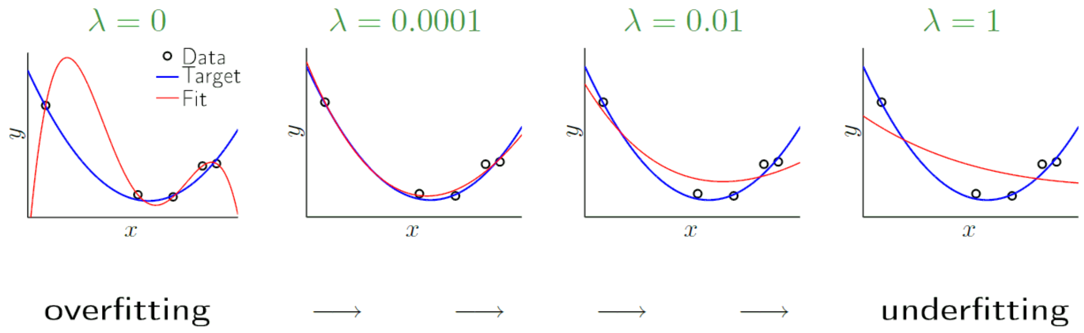


Constrained Solution

- Minmize $E_{in} + \frac{\lambda}{N} w^T w$
- $\frac{2}{N} X^T (Xw - y) + \frac{2\lambda}{N} w = 0$
- $X^T X w - X^T y + \lambda w = 0$
- $(X^T X + \lambda I) w = X^T y$
- $w = (X^T X + \lambda I)^{-1} X^T y$
- Recall that without regularization the solution was:
- $w = (X^T X)^{-1} X^T y$

Bias-Variance Trade off

- $E_{in} + \frac{\lambda}{N} w^T w$
- Small lambda:
  - Low bias, high variance
  - Lambda = 0 => you get LS solution
  - Prone to overfitting
- Large lambda:
  - High bias, low variance
  - what's w when lambda goes to infinity?
  - Prone to underfitting
- Use cross-validation to determine a proper value for lambda



Coefficient Path

- This graph shows how the magnitude of weight changes as a function of lambda.
- When lambda was 0, we get least square solution.
- When lambda goes to infinity, we get very small coefficients approaching 0.
- In between, we get some other set of coefficients.
- One of those intermediate values is the desirable answer that's determined by cross-validation.