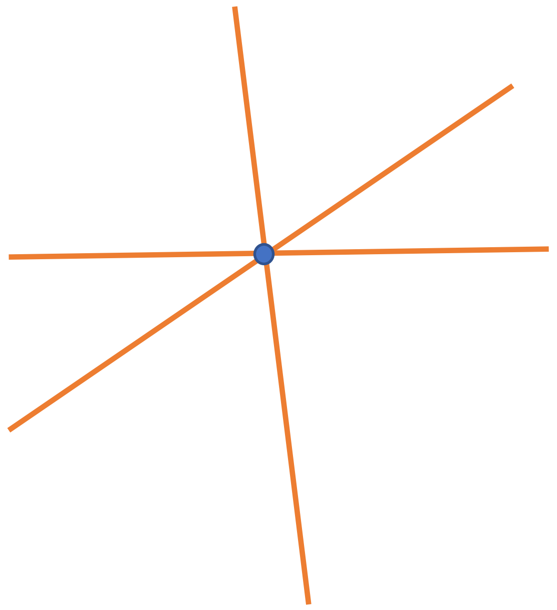
Linear Regression

ECE 449

Regression Problems

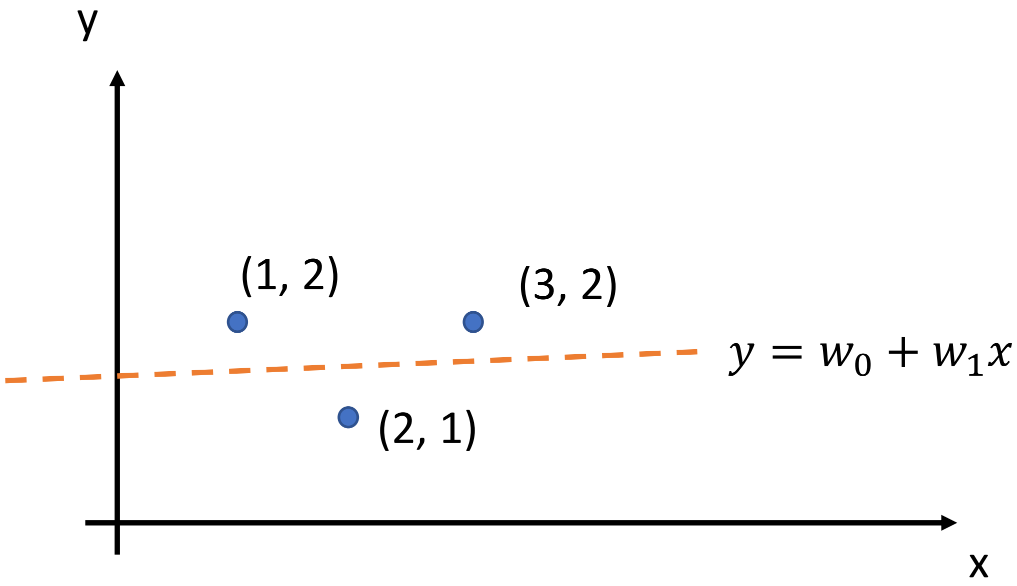
- Task: Given observation vector x, estimate real-valued $\hat{y} = f(x)$ to minimize error
- Examples
  - Predict temperature from weather statistics
  - Predict object position in space
- Linear if $f(x; w) = w^t x$
- Extend to non-linear $f(x; w) = w^t h(x)$, where $h(x)$ is a non-linear function
- Input features
  - Numerical $(\mathcal{R}^d, \mathcal{Z}^d)$, binary {0,1}, categorical
  - If categorical, map to a one-hot vector

Line Fitting



Line Fitting

- How about 3 points?
- $y = w_0 + w_1 x$
- What is a "**good**" choice for w_0 and w_1?
  - How to define "good"?
  - How to estimate w_0 and w_1?



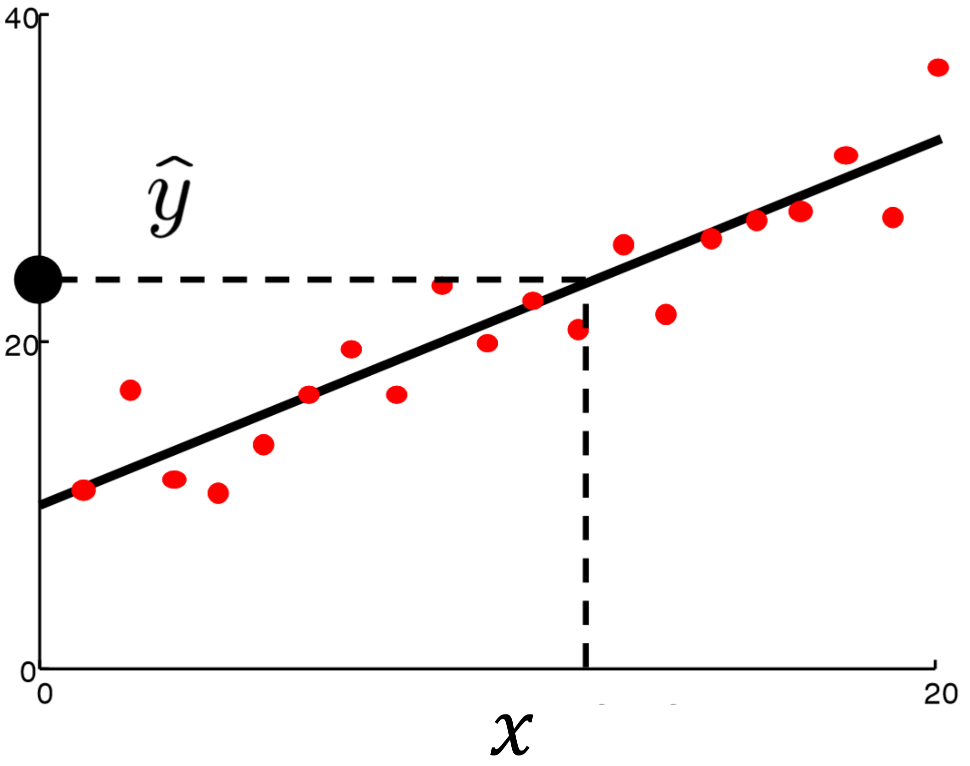
Line Fitting

- How to define "good"?
  - An appropriate loss function, like MSE
- How to estimate w_0 and w_1?

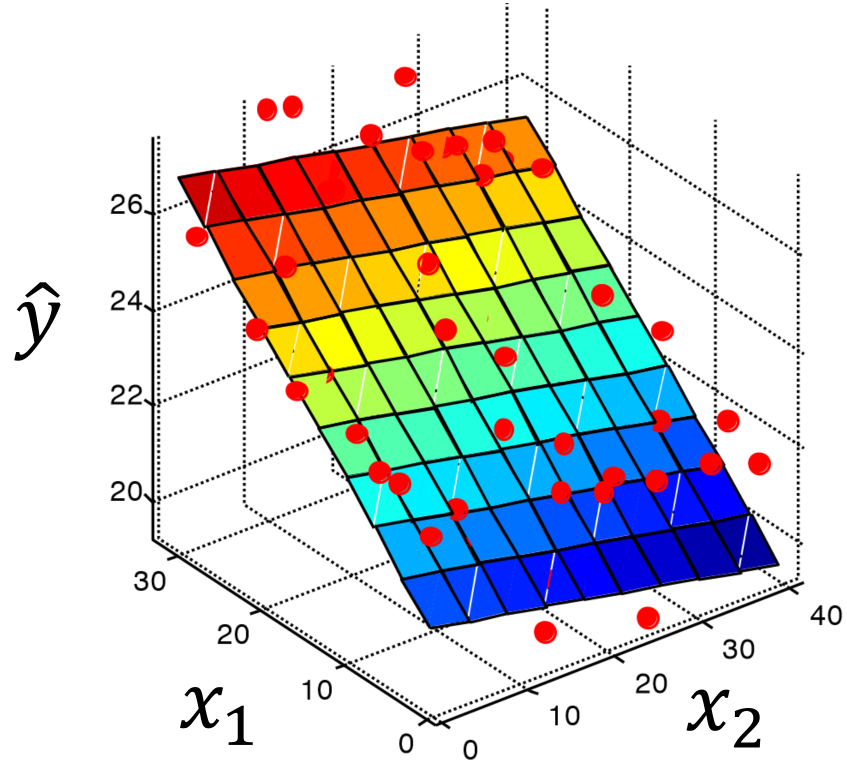
- $l = \left(y_1 - (w_0 + w_1 x_1)\right)^2 + \left(y_2 - (w_0 + w_1 x_2)\right)^2 + \left(y_3 - (w_0 + w_1 x_3)\right)^2$
- $\frac{\partial l}{\partial w_0} = 0, \frac{\partial l}{\partial w_1} = 0$

More Fitting Problems

$$\hat{y} = w_0 + w_1 x$$

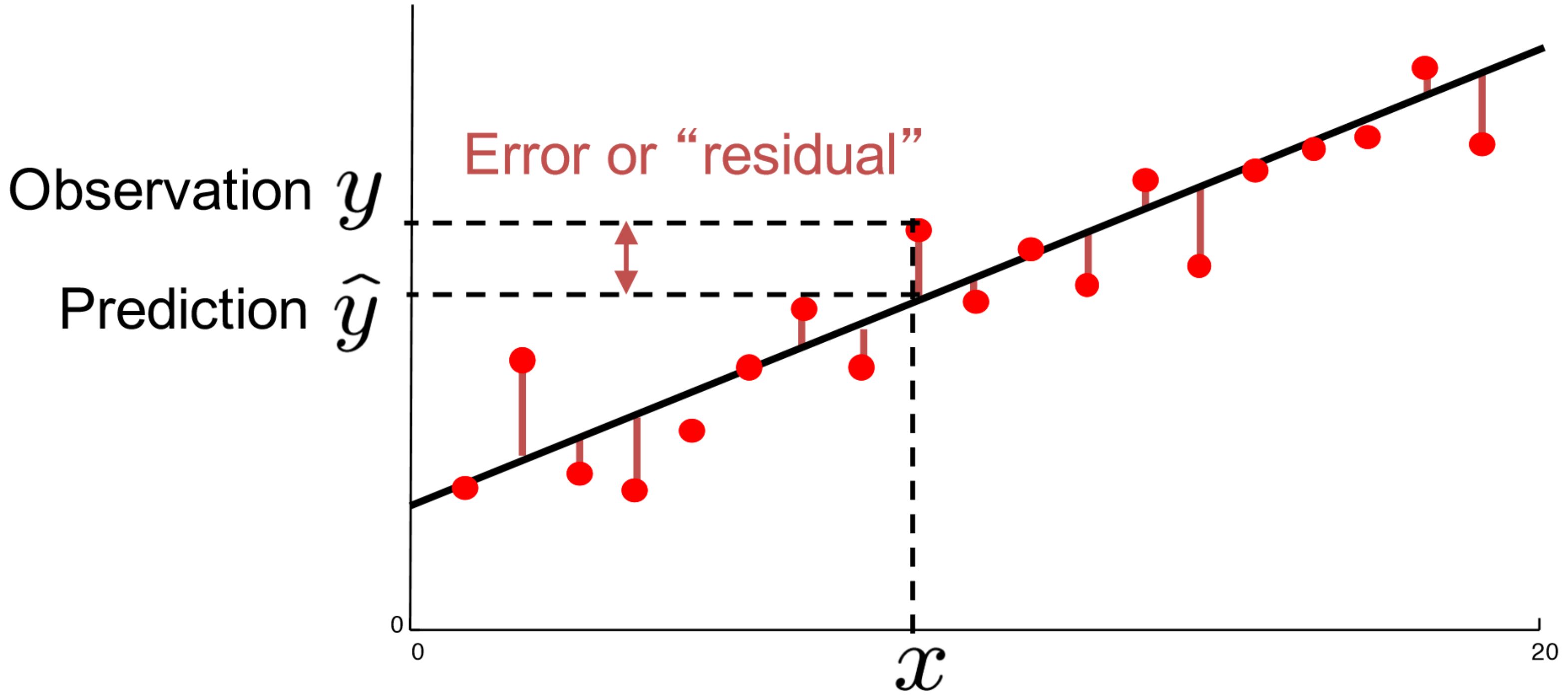


$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2$$



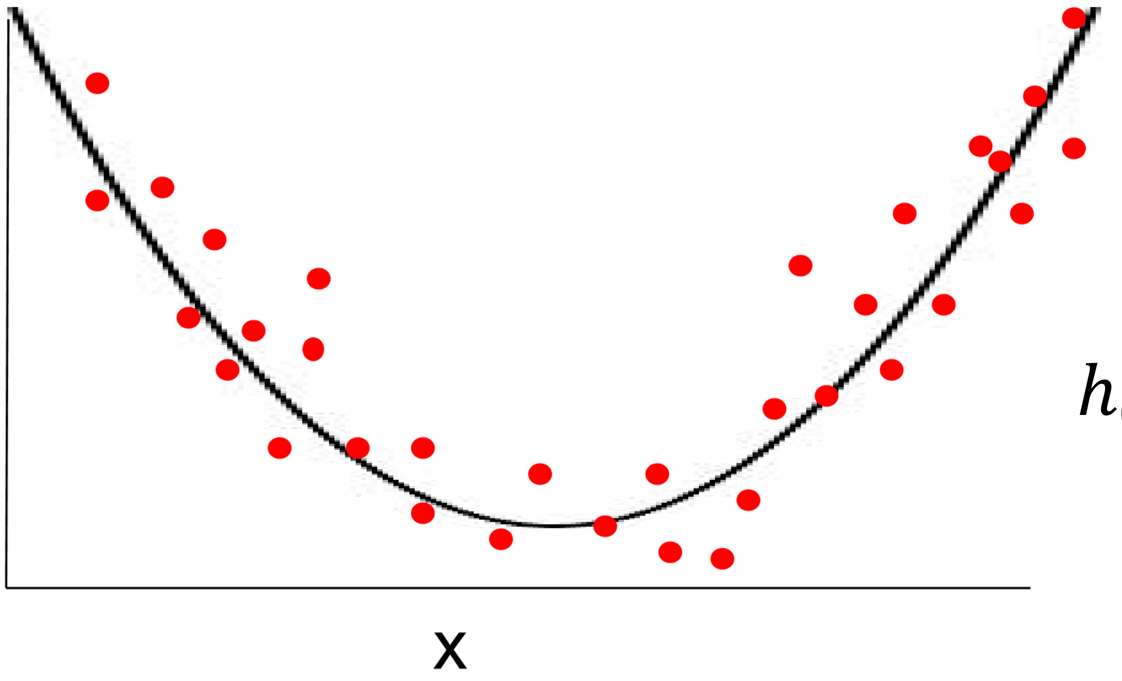
Ordinary Least Squares (OLS)

- Total error $= \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - \sum_k w_k x_{i,k})^2$



Extend to Non-Linear Function

- Map from x to $h(x)$
- Total error = $\sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - \sum_k w_k h_k\ (x_i))^2$

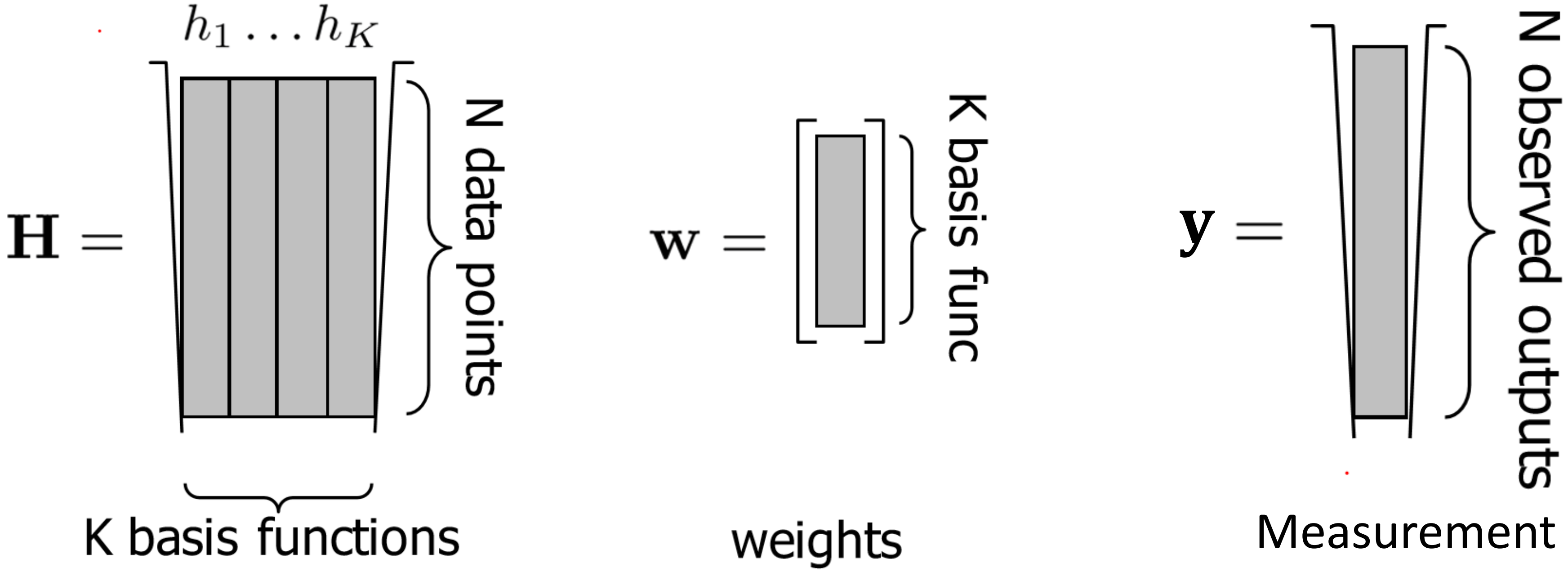


$$f(x) = w_0 + w_1\ x_1 + w_2 x_2^2$$

$$h_0(x) = 1, h_1(x) = x_1, h_2(x) = x_2^2$$

Regression with Matrix Notation

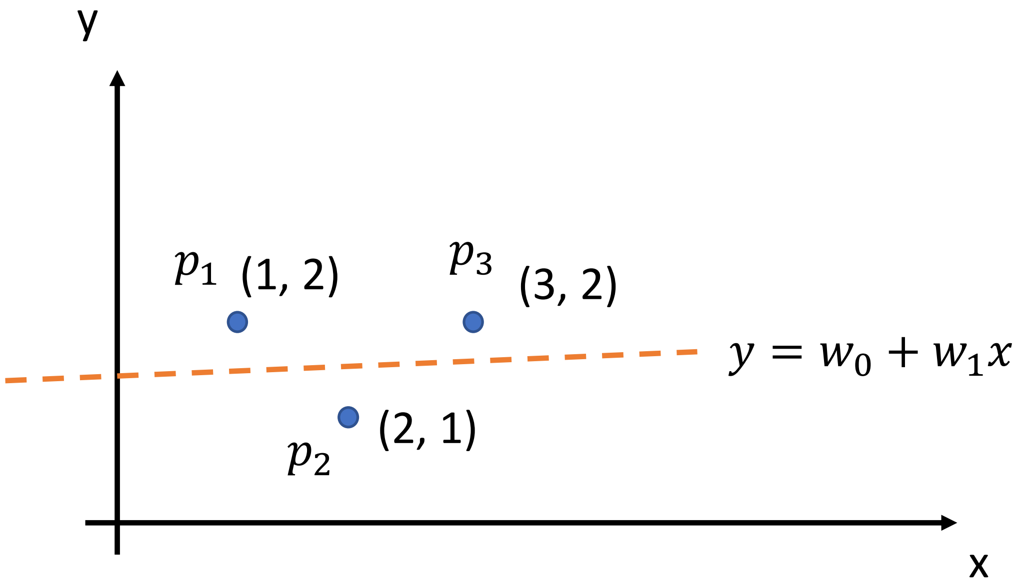
- $\hat{w} = \arg\min_{w} \sum_i (y_i - \sum_k w_k h_k\ (x_i))^2$
- $\hat{w} = \arg\min_{w} (\mathbf{Hw} - \mathbf{y})^T (\mathbf{Hw} - \mathbf{y})$



Matrix Notation

- $\mathbf{H}$=?, $\mathbf{y}$=?, $\mathbf{w}$=?

- $\mathbf{H} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix}$



Closed Form Solution

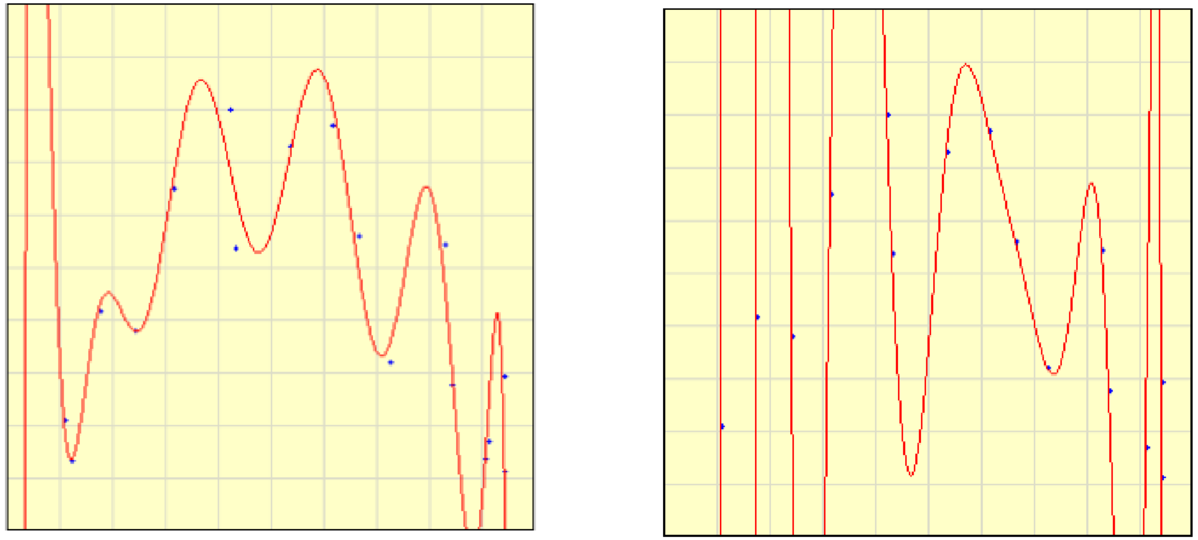
- The closed form solution
- $\hat{w} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}$

- $l = (\mathbf{Hw} - \mathbf{y})^T(\mathbf{Hw} - \mathbf{y})$
- $\frac{\partial l}{\partial \mathbf{w}} = 2\mathbf{H}^T(\mathbf{Hw} - \mathbf{y}) = 0$
- $\mathbf{H}^T\mathbf{Hw} = \mathbf{H}^T\mathbf{y}$
- $\hat{w} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}$

https://en.wikipedia.org/wiki/Matrix_calculus

Regularization in Linear Regression

- One sign of overfitting: large parameter values



- Regularized or penalized regressions modified
  - Learning object to penalize large parameters

Ridge Regression

- l_2 regularization
- $\widehat{w}_{ridge} = \arg\min_{w} \sum_i \left(y_i - \left(w_0 + \sum_{k=1}^{K} w_k h_k\,(x_i)\right)\right)^2 + \lambda \sum_{k=1}^{K} w_k^2$
- $\widehat{w}_{ridge} = \arg\min_{w} (\mathbf{Hw} - \mathbf{y})^T (\mathbf{Hw} - \mathbf{y}) + \lambda \mathbf{w}^T I_{0+K} \boldsymbol{w}$
- The closed form solution
- $\widehat{w}_{ridge} = (\mathbf{H}^T\mathbf{H} + \lambda I_{0+K})^{-1}\mathbf{H}^T\mathbf{y}$

Ridge Regression

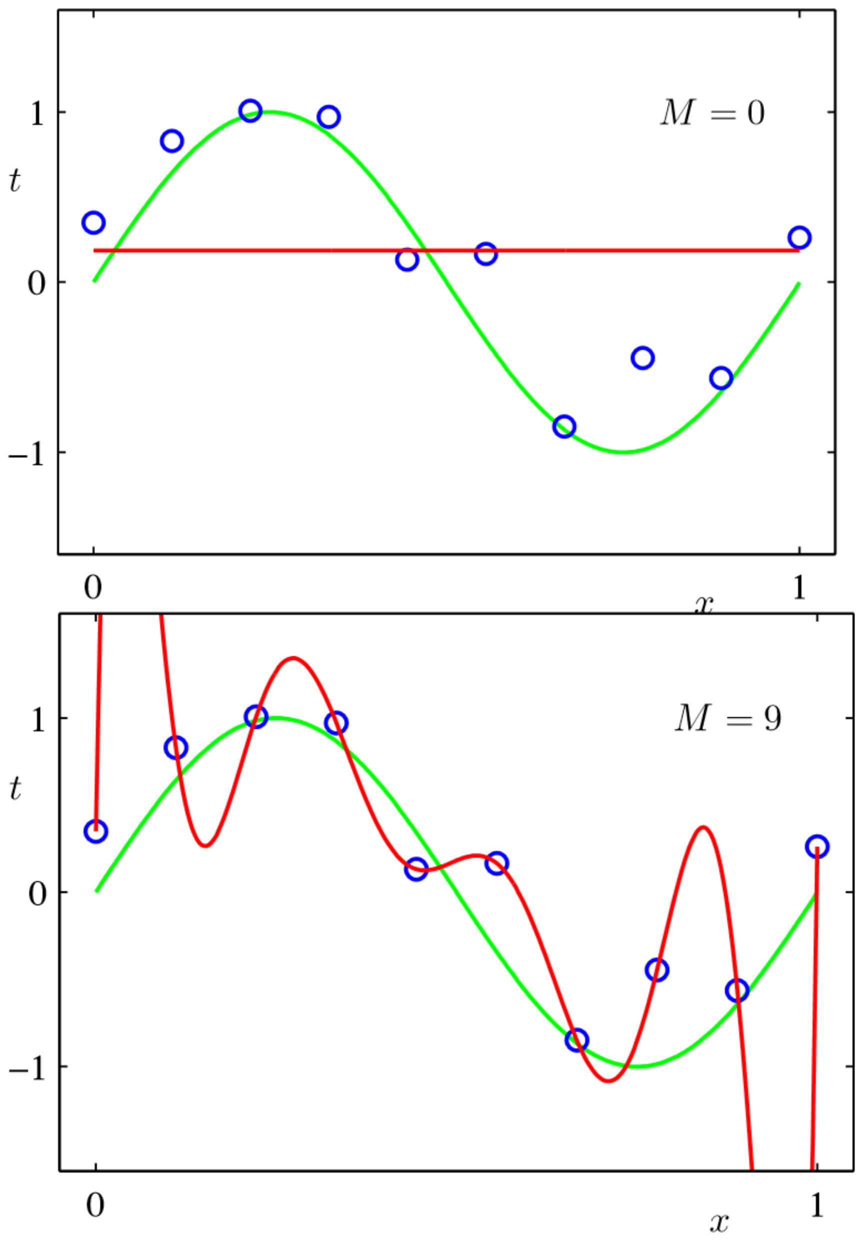
- How does varying λ change w?

- Larger λ? Smaller λ?
- As $\lambda \rightarrow 0$?
  - unregularized
- As $\lambda \rightarrow \infty$?
  - All weights will be 0

Bias-Variance Trade-off

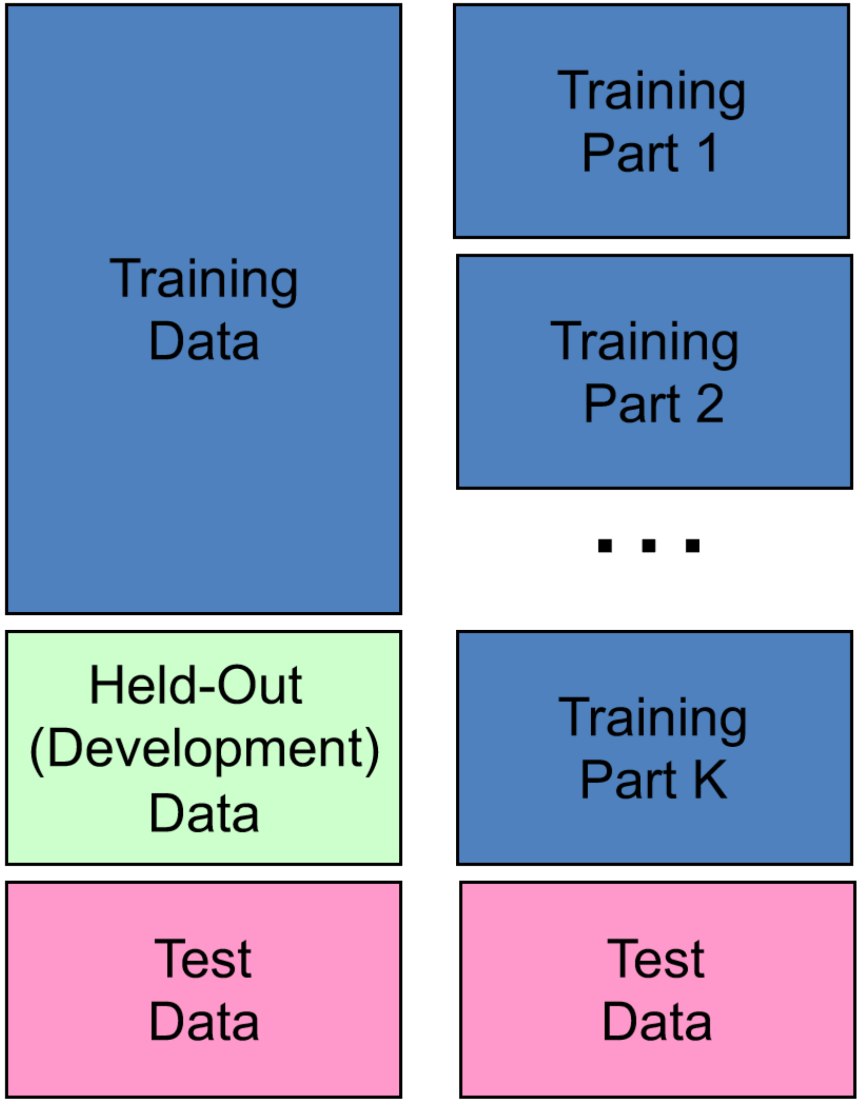
- Model too simple: does not fit the data well
  - A biased solution
- Model too complex: small changes to the data, solution changes a lot
  - A high-variance solution

- Regularization reduces variance at the cost of some bias



How to Pick λ

- Experimentation cycle
  - Select a hypothesis f to best match training set
  - Tune λ on held-out set
  - Try many different values of lambda, pick best one
- Or, can do k-fold cross validation
  - No held-out set
  - Divide training set into k subsets
  - Repeatedly train on k-1 and test on remaining one
  - Use average of λ's to retrain on full data set, OR use average of w's

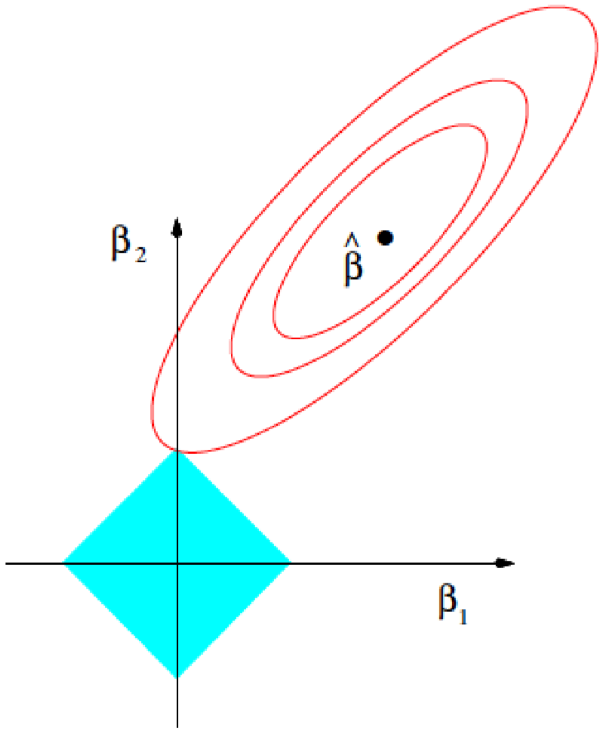


LASSO

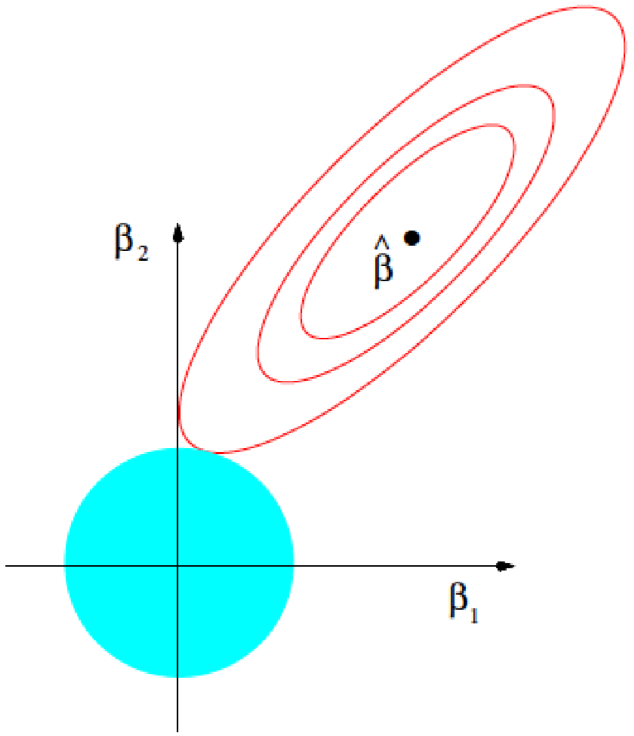
- l_1 regularization
- $\hat{w}_{LASSO} = \arg\min_{w} \sum_i \left(y_i - \left(w_0 + \sum_{k=1}^{K} w_k h_k\,(x_i)\right)\right)^2 + \lambda \sum_{k=1}^{K} |w_k|$

- Linear penalty pushes more weights to zero
- Allows for a type of feature selection
- But, not differentiable and no closed form solution….

Geometric Intuition



Lasso



Ridge