CS480/680: Introduction to Machine Learning

Lec 02: Linear Regression

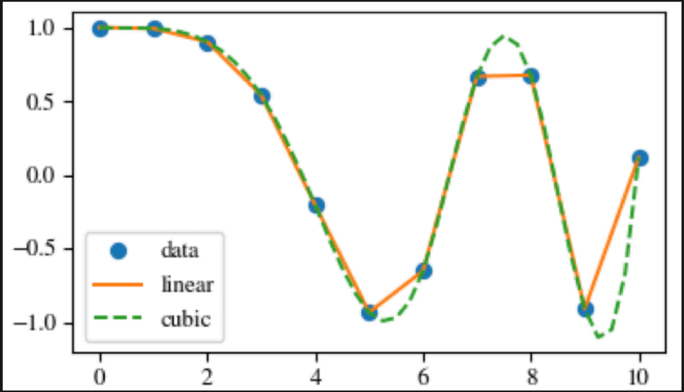
Yaoliang Yu

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

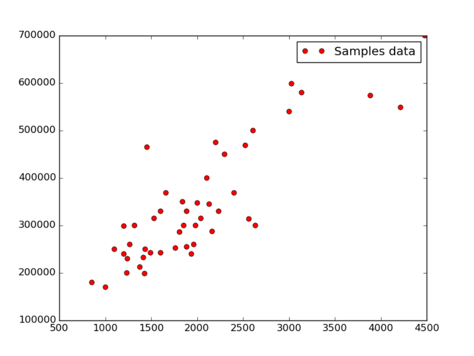
Jan 14, 2025

Regression

- Given training data $\wr(\mathbf{x}_i, \mathbf{y}_i)\int$, find $f : \mathcal{X} \to \mathcal{Y}$ such that $f(\mathbf{x}_i) \approx y_i$
  - $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$: feature vector for the i-th training example
  - $\mathbf{y}_i \in \mathcal{Y} \subseteq \mathbb{R}^t$: t responses, e.g. $t = 1$ or even $t = \infty$

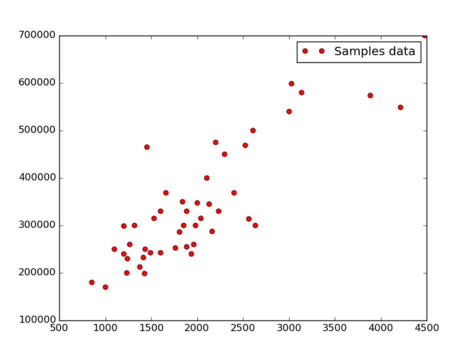


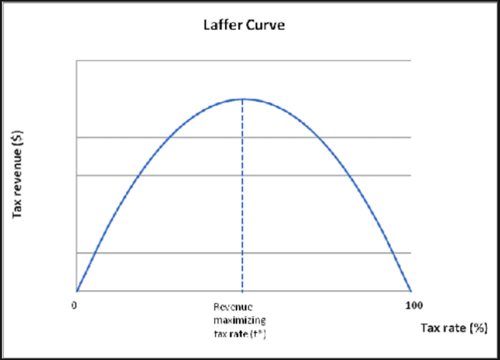
Some Examples



- Prior knowledge on the functional form of f
- Linear vs. nonlinear

Some Examples





- Prior knowledge on the functional form of f
- Linear vs. nonlinear

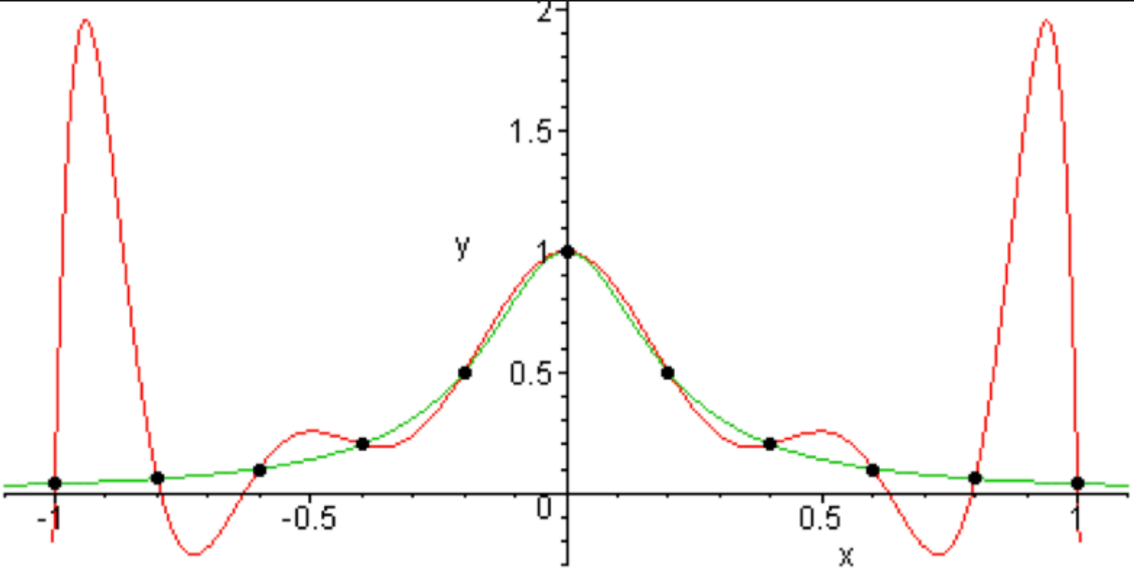
The Difficulty

Theorem: Exact interpolation is always possible

For any* finite training data $\{(\mathbf{x}_i, \mathbf{y}_i) : i = 1, \ldots, n\}$, there exist infinitely many functions f such that for all i,

$$f(\mathbf{x}_i) = \mathbf{y}_i.$$

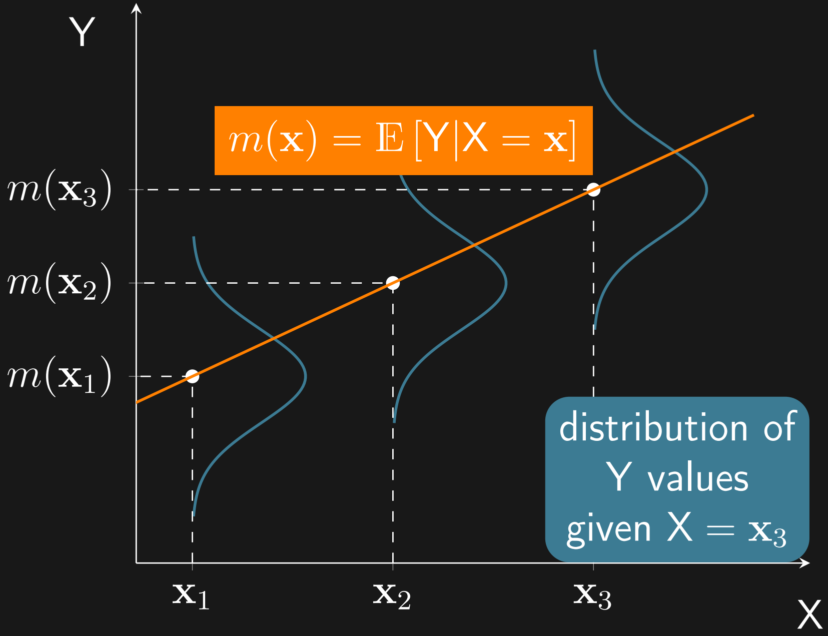
- No amount of training data is enough to decide on a unique f!
- On new data $\mathbf{x}$, our prediction $\hat{\mathbf{y}} = f(\mathbf{x})$ can vary wildly!
- This is where prior knowledge of f comes into play
- Occam's razor: "the simplest explanation is usually the correct one"



2
1.5
y
1
0.5
0
-1
-0.5
0.5
1
x

Statistical Learning

- Training and test data are both iid samples from the same unknown distribution $\mathbb{P}$
  - $(X_i, Y_i) \sim \mathbb{P}$ and $(X, Y) \sim \mathbb{P}$
- Least squares regression: $\min_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}\|f(X) - Y\|_2^2$
- Regression function: $m(\mathbf{x}) = \mathbb{E}[Y|X = \mathbf{x}]$
- Need to know the distribution $\mathbb{P}$, i.e., all pairs (X, Y)!
- Changing the square loss changes the regression function accordingly



Y
$m(\mathbf{x}) = \mathbb{E}[\mathsf{Y}|\mathsf{X} = \mathbf{x}]$
$m(\mathbf{x}_3)$
$m(\mathbf{x}_2)$
$m(\mathbf{x}_1)$
distribution of Y values given $\mathsf{X} = \mathbf{x}_3$
$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{x}_3$
X

Bias-Variance Decomposition

$$
\begin{aligned}
\mathbb{E}\|f(\mathsf{X}) - \mathsf{Y}\|_2^2 &= \mathbb{E}\|f(\mathsf{X}) - m(\mathsf{X}) + m(\mathsf{X}) - \mathsf{Y}\|_2^2 \\
&= \mathbb{E}\|f(\mathsf{X}) - m(\mathsf{X})\|_2^2 + \mathbb{E}\|m(\mathsf{X}) - \mathsf{Y}\|_2^2 \\
&\qquad \cancel{+2\mathbb{E}\left\langle f(\mathsf{X}) - m(\mathsf{X}), m(\mathsf{X}) - \mathsf{Y}\right\rangle} \\
&= \underbrace{\mathbb{E}\|f(\mathsf{X}) - m(\mathsf{X})\|_2^2}_{\text{bias}^2} + \underbrace{\mathbb{E}\|m(\mathsf{X}) - \mathsf{Y}\|_2^2}_{\text{noise variance}}
\end{aligned}
$$

- The noise variance does not depend on our choice of f!
  - it is an inherent measure of the difficulty of our problem
- We aim to choose $f \approx m$ to minimize bias hence squared error

Sampling $\to$ Training

$$\min_{f:\mathcal{X}\to\mathcal{Y}} \hat{\mathbb{E}}\|f(\mathsf{X}) - \mathsf{Y}\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\|f(\mathsf{X}_i) - \mathsf{Y}_i\|_2^2$$

- Replace expectation with sample average: $(\mathsf{X}_i, \mathsf{Y}_i) \overset{i.i.d.}{\sim} P$
- Finite training set $\to$ exact interpolation paradox!
- Need to restrict the form of f, using prior knowledge
- (Uniform) law of large numbers: as training data size $n \to \infty$,

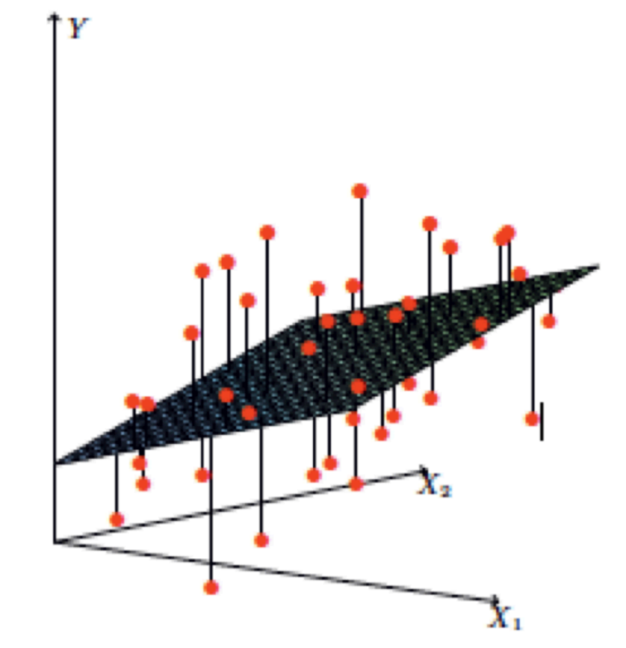
$$\hat{\mathbb{E}} \to \mathbb{E} \text{ and (hopefully) } \operatorname{argmin} \hat{\mathbb{E}} \to \operatorname{argmin} \mathbb{E}$$

Linear Least Squares Regression

- Affine function: $f(\mathbf{x}) = W\mathbf{x} + \mathbf{b}$ with $W \in \mathbb{R}^{t\times d}$ and $\mathbf{b} \in \mathbb{R}^t$
- Padding: $\mathbf{x} \leftarrow \binom{\mathbf{x}}{1}$, $\mathbf{W} \leftarrow [W, \mathbf{b}]$, hence $f(\mathbf{x}) = \mathbf{W}\mathbf{x}$
- In matrix form: $\frac{1}{n}\sum_i \|f(\mathbf{x}_i) - \mathbf{y}_i\|_2^2 \quad = \quad \frac{1}{n}\|\mathbf{W}\mathbf{X} - \mathbf{Y}\|_{\mathsf{F}}^2$
  - $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{(d+1)\times n}$, $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^{t\times n}$
  - $\|A\|_{\mathsf{F}} = \sqrt{\sum_{ij} a_{ij}^2}$

$$\boxed{\min_{\mathbf{W}\in\mathbb{R}^{t\times(d+1)}} \quad \frac{1}{n}\|\mathbf{W}\mathbf{X} - \mathbf{Y}\|_{\mathsf{F}}^2}$$

S. M. Stigler. "Gauss and the Invention of Least Squares". *The Annals of Statistics*, vol. 9, no. 3 (1981), pp. 465–474.


Y
X₂
X₁

Calculus Detour

- Let $f : \mathbb{R}^p \to \mathbb{R}$ be a smooth real-valued function
- Fix an inner product $\langle \cdot, \cdot \rangle$
- Define the gradient $\nabla f : \mathbb{R}^p \to \mathbb{R}^p$ as

$$\frac{\mathrm{d} f(\mathbf{w} + t\mathbf{z})}{\mathrm{d} t} \restriction_{t=0} \quad = \quad \langle \nabla f(\mathbf{w}), \mathbf{z} \rangle$$

  - LHS is the usual (scalar) derivative of the univariate function $t \mapsto f(\mathbf{w} + t\mathbf{z})$
  - $\mathbf{w}$ and $\mathbf{z}$ are fixed as constants: directional derivative
  - gradient ∇f is representation of directional derivative over a chosen inner product
- Chain rule still holds

Example: Univariate functions

Consider $f : \mathbb{R} \to \mathbb{R}$ (i.e., $p = 1$) and the standard inner product $\langle a, b \rangle := ab$. By chain rule:

$$\frac{\mathrm{d}f(w + tz)}{\mathrm{d}t} \restriction_{t=0} \;=\; f'(w + tz)z \restriction_{t=0} = f'(w)z = \langle f'(w), z \rangle \,,$$

i.e., $\nabla f(w) = f'(w)$. What is the gradient if we choose $\langle a, b \rangle := 2ab$?

Example: Partial derivatives

Consider $f : \mathbb{R}^p \to \mathbb{R}$ and the standard inner product $\langle \mathbf{w}, \mathbf{x} \rangle := \sum_j w_j x_j$. Choose the direction $\mathbf{z} = \mathbf{e}_j$ (i.e., 1 at the j-th entry and 0 elsewhere):

$$\frac{\mathrm{d}f(\mathbf{w} + t\mathbf{e}_j)}{\mathrm{d}t} \restriction_{t=0} \;=\; \partial_j f(\mathbf{w}) = \langle \nabla f(\mathbf{w}), \mathbf{e}_j \rangle = [\nabla f(\mathbf{w})]_j,$$

i.e., $\nabla f(w) = [\partial_1 f(\mathbf{w}), \ldots, \partial_p f(\mathbf{w})]$.

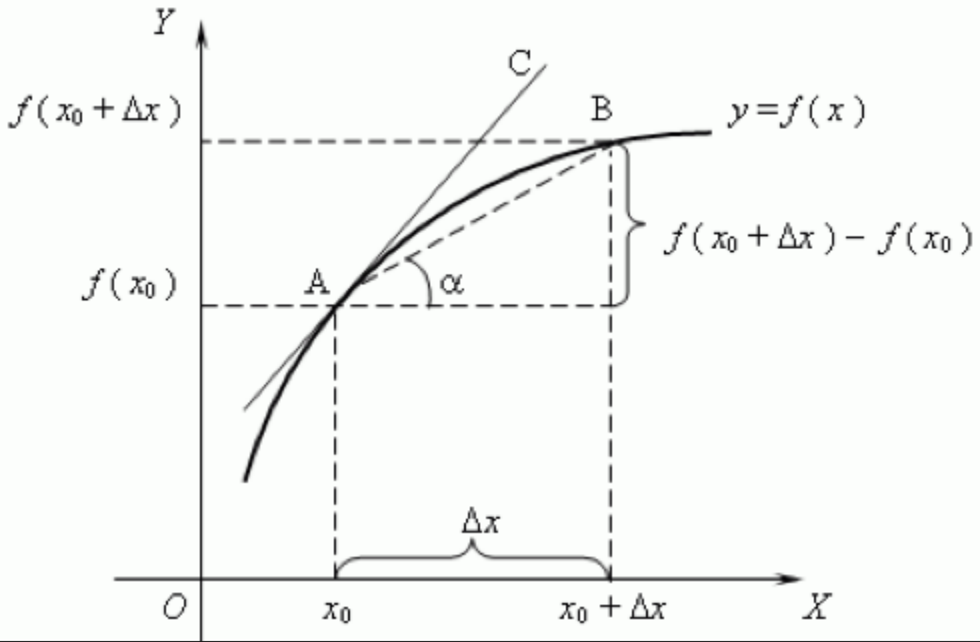
Example: Quadratic function

Consider the quadratic function $f(\mathbf{w}) = \langle \mathbf{w}, A\mathbf{w} + \mathbf{b} \rangle + c$.

$$\begin{aligned} f(\mathbf{w} + t\mathbf{z}) &= \langle \mathbf{w} + t\mathbf{z}, A(\mathbf{w} + t\mathbf{z}) + \mathbf{b} \rangle + c \\ &= t^2 \langle \mathbf{z}, A\mathbf{z} \rangle + t \langle \mathbf{w}, A\mathbf{z} \rangle + t \langle \mathbf{z}, A\mathbf{w} + \mathbf{b} \rangle + \langle \mathbf{w}, A\mathbf{w} + \mathbf{b} \rangle + c \\ \frac{\mathrm{d} f(\mathbf{w} + t\mathbf{z})}{\mathrm{d}t} \restriction_{t=0} &= \langle \mathbf{w}, A\mathbf{z} \rangle + \langle \mathbf{z}, A\mathbf{w} + \mathbf{b} \rangle = \left\langle A^\top \mathbf{w} + A\mathbf{w} + \mathbf{b}, \mathbf{z} \right\rangle, \end{aligned}$$

i.e., $\boxed{\nabla f(\mathbf{w}) = (A^\top + A)\mathbf{w} + \mathbf{b}}$.

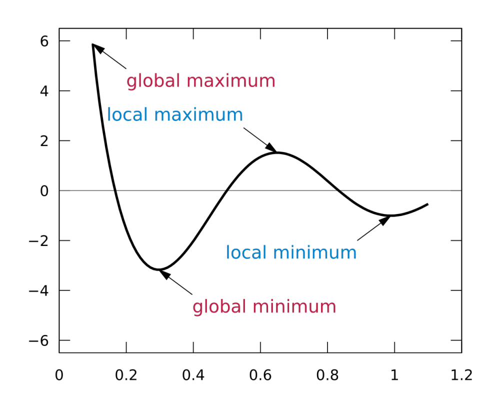
- $\langle \mathbf{a} + \mathbf{b}, \mathbf{x} + \mathbf{y} \rangle = \langle \mathbf{a}, \mathbf{x} \rangle + \langle \mathbf{a}, \mathbf{y} \rangle + \langle \mathbf{b}, \mathbf{x} \rangle + \langle \mathbf{b}, \mathbf{y} \rangle$
- $\langle \mathbf{a}, t\mathbf{b} \rangle = \langle t\mathbf{a}, \mathbf{b} \rangle = t \langle \mathbf{a}, \mathbf{b} \rangle$
- $\langle \mathbf{w}, A\mathbf{z} \rangle = \left\langle A^\top \mathbf{w}, \mathbf{z} \right\rangle$, $\langle A\mathbf{w}, \mathbf{z} \rangle = \left\langle \mathbf{w}, A^\top \mathbf{z} \right\rangle$

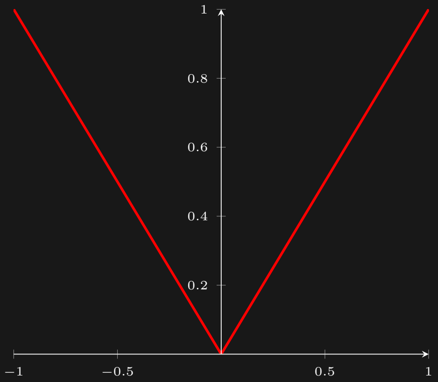

Y
C
f (x0 + Δx)
B
y = f (x)
f (x0 + Δx) − f (x0)
A
α
f (x0)
Δx
O
x0
x0 + Δx
X

Optimality Condition

Theorem: Fermat's necessary condition for extremity

If $\mathbf{w}$ is a minimizer (or maximizer) of a differentiable function f over an open set, then $f'(\mathbf{w}) = \mathbf{0}$.





Solving Linear Regression

$$\|\mathbf{WX} - \mathbf{Y}\|_{\mathsf{F}}^2 = \langle \mathbf{WX} - \mathbf{Y}, \mathbf{WX} - \mathbf{Y} \rangle$$
$$= \langle \mathbf{W}, \mathbf{WXX}^\top - 2\mathbf{YX}^\top \rangle + \langle \mathbf{Y}, \mathbf{Y} \rangle$$

- Taking derivative w.r.t. $\mathbf{W}$ and setting to zero:

Normal equation $\boxed{\mathbf{WXX}^\top = \mathbf{YX}^\top} \implies \mathbf{W} = \mathbf{YX}^\top(\mathbf{XX}^\top)^{-1} =: \mathbf{YX}^\dagger$

- $\mathbf{X} \in \mathbb{R}^{(d+1)\times n}$ hence $\mathbf{XX}^\top \in \mathbb{R}^{(d+1)\times(d+1)}$: may not be invertible if $n \leq d+1$, but a solution always exists
- Even when invertible, never compute the inverse directly!
- Instead, solve the linear system or apply iterative gradient algorithm

Prediction

- Once solved $\mathbf{W}$ on the training set $(\mathbf{X}, \mathbf{Y})$, can predict on unseen data $\mathbf{X}_{\text{test}}$:

$$\hat{\mathbf{Y}}_{\text{test}} = \mathbf{W}\mathbf{X}_{\text{test}}$$

- We may evaluate our test error if true labels were available:

$$\frac{1}{n_{\text{test}}} \|\mathbf{Y}_{\text{test}} - \hat{\mathbf{Y}}_{\text{test}}\|_{\mathsf{F}}^2$$

- We may compare to the training error:

$$\frac{1}{n} \|\mathbf{Y} - \hat{\mathbf{Y}}\|_{\mathsf{F}}^2, \quad \text{where} \quad \hat{\mathbf{Y}} := \mathbf{W}\mathbf{X}$$

- Minimizing the training error as a means to reduce the test error
- Sometimes we even evaluate the test error using a different loss $\mathbb{L}(\mathbf{Y}_{\text{test}}, \hat{\mathbf{Y}}_{\text{test}})$
  - leads to a beautiful theory of loss calibration

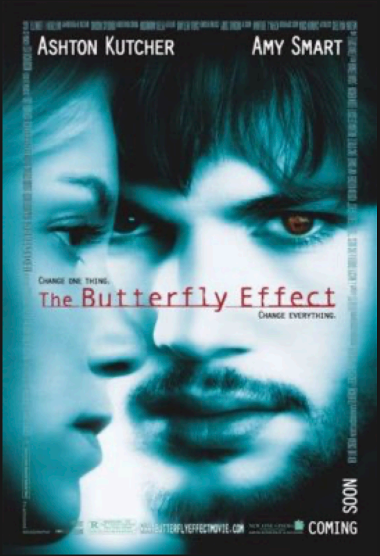
Ill-conditioning

$$\mathbf{X} = \begin{bmatrix} 0 & \epsilon \\ 1 & 1 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} 1 & -1 \end{bmatrix}$$

- Solving linear least squares regression:

$$\mathbf{w} = \mathbf{y}\mathbf{X}^{-1} = \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} -1/\epsilon & 1 \\ 1/\epsilon & 0 \end{bmatrix} = \begin{bmatrix} -2/\epsilon & 1 \end{bmatrix}$$

- Slight perturbation leads to chaotic behaviour!
- Happens whenever $\mathbf{X}$ is ill-conditioned, i.e., (close to) rank deficient



Tikhonov Regularization, a.k.a. Ridge Regression

$$\min_{\mathbf{W}} \ \frac{1}{n}\|\mathbf{W}\mathbf{X} - \mathbf{Y}\|_{\mathsf{F}}^2 + \boxed{\lambda\|\mathbf{W}\|_{\mathsf{F}}^2}$$

- Normal equation: $\mathbf{W}(\mathbf{X}\mathbf{X}^\top + \lambda I) = \mathbf{Y}\mathbf{X}^\top$
- Regularization const. λ controls trade-off
  - $\lambda = 0$ reduces to ordinary linear regression
  - $\lambda = \infty$ reduces to $\mathbf{W} \equiv \mathbf{0}$
  - intermediate λ restricts output to be $\frac{1}{\lambda}$ proportional to input
- May choose to not regularize offset $\mathbf{b}$

A. N. Tikhonov. "Solution of incorrectly formulated problems and the regularization method". *Soviet Mathematics*, vol. 4, no. 4 (1963), pp. 1035–1038, A. E. Hoerl and R. W. Kennard. "Ridge regression: Biased estimation for nonorthogonal problems". *Technometrics*, vol. 12, no. 1 (1970), pp. 55–67.

Data Augmentation

$$\frac{1}{n}\|\mathbf{W}\mathbf{X}-\mathbf{Y}\|_{\mathsf{F}}^2+\boxed{\lambda\|\mathbf{W}\|_{\mathsf{F}}^2}=\frac{1}{n}\|\mathbf{W}\underbrace{\begin{bmatrix}\mathbf{X} & \sqrt{n\lambda}I\end{bmatrix}}_{\tilde{\mathbf{X}}}-\underbrace{\begin{bmatrix}\mathbf{Y} & \mathbf{0}\end{bmatrix}}_{\tilde{\mathbf{Y}}}\|_{\mathsf{F}}^2$$

- Augment **X** with $\sqrt{n\lambda}I$, i.e. p data points $\mathbf{x}_j=\sqrt{n\lambda}\mathbf{e}_j, j=1,\ldots,p$
- Augment **Y** with zero
- Shrinks **W** towards origin

$$\boxed{\text{regularization} = \text{data augmentation}}$$

Sparsity

- Regularization $\iff$ constraint:

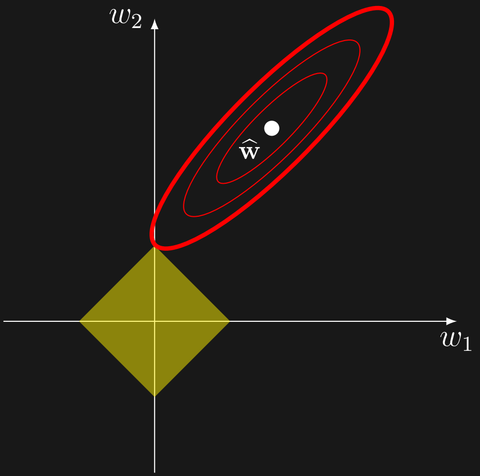
$$\min_{\|\mathbf{W}\|_\mathsf{F} \le \gamma} \frac{1}{n}\|\mathbf{WX} - \mathbf{Y}\|_\mathsf{F}^2$$

- Ridge regression $\to$ dense $\mathbf{W}$
  - more computation / communication
  - harder to interpret
- Lasso (Tibshirani, 1996):

$$\min_{\|\mathbf{W}\|_1 \le \gamma} \frac{1}{n}\|\mathbf{WX} - \mathbf{Y}\|_\mathsf{F}^2$$

- Regularization $\iff$ constraint:

$$\min_{\mathbf{W}} \frac{1}{n}\|\mathbf{WX} - \mathbf{Y}\|_\mathsf{F}^2 + \lambda\|\mathbf{W}\|_1$$



R. Tibshirani. "Regression Shrinkage and Selection via the Lasso". *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1 (1996), pp. 267–288.

Task Regularization

$$\min_{\mathbf{W}} \frac{1}{n}\|\mathbf{W}\mathbf{X}-\mathbf{Y}\|_{\mathsf{F}}^2+\lambda\|\mathbf{W}\|_{\mathsf{F}}^2 \quad \equiv \quad \min_{\mathbf{w}_\tau} \frac{1}{n}\|\mathbf{w}_\tau\mathbf{X}-\mathbf{y}_\tau\|_{\mathsf{F}}^2+\lambda\|\mathbf{w}_\tau\|_2^2,\ \forall\tau=1,\ldots,t$$

- In other words, the tasks are independent and can be solved separately
- Sometimes lumping tasks together (LHS) is computationally more efficient
- If tasks are related, can consider a kind of low-rank regularization:

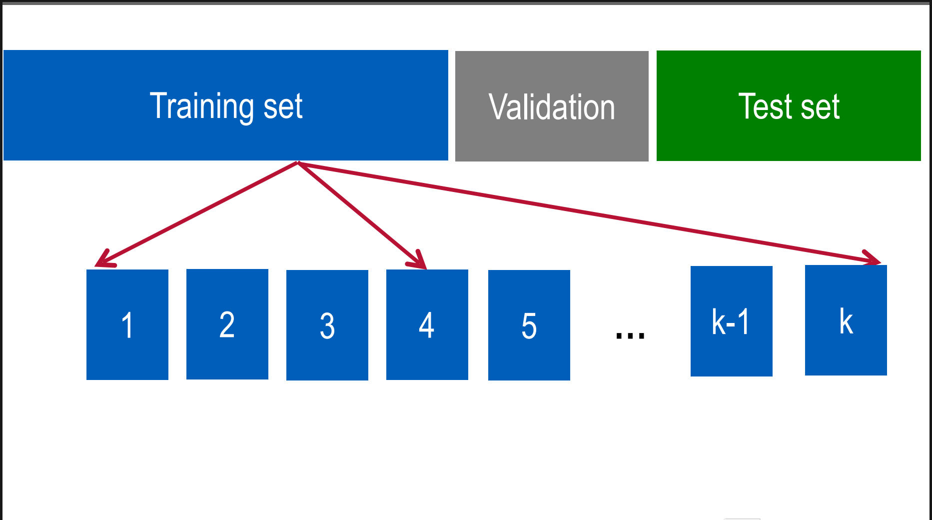
$$\min_{\mathbf{W}} \frac{1}{n}\|\mathbf{W}\mathbf{X}-\mathbf{Y}\|_{\mathsf{F}}^2+\lambda\|\mathbf{W}\|_{\mathrm{tr}},$$

where $\|A\|_{\mathrm{tr}}$ is the sum of singular values (i.e., the trace norm).

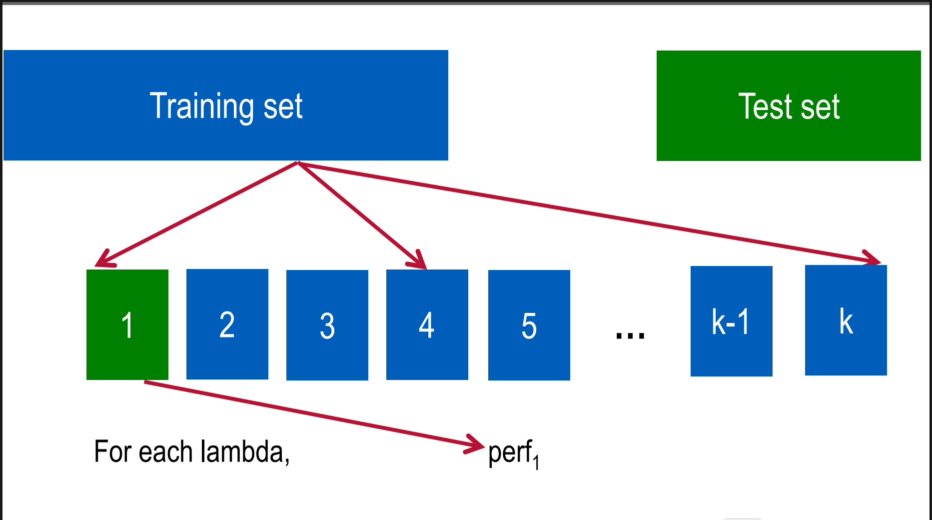
R. Caruana. "Multitask Learning". *Machine Learning*, vol. 28 (1997), pp. 41–75, A. Argyriou, T. Evgeniou, and M. Pontil. "Convex multi-task feature learning". *Machine Learning*, vol. 73 (2008), pp. 243–272.

Cross-validation

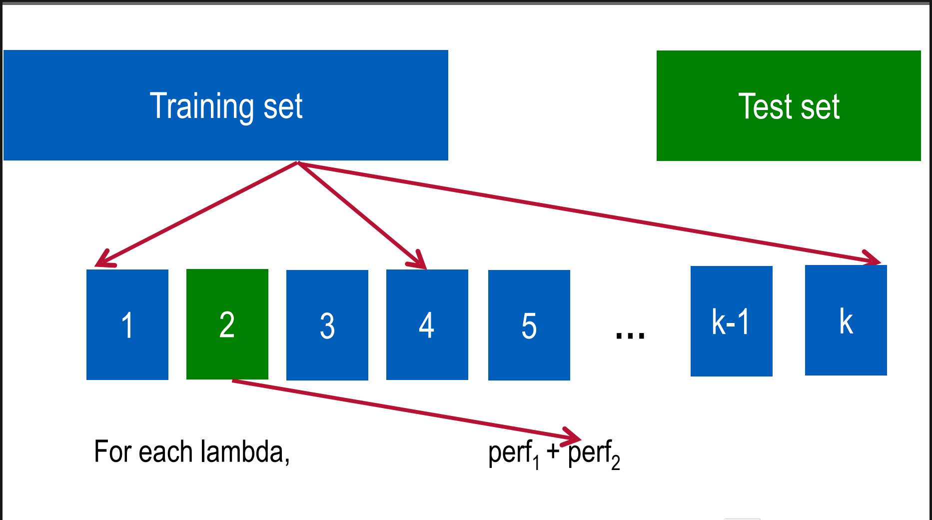
Cross-validation



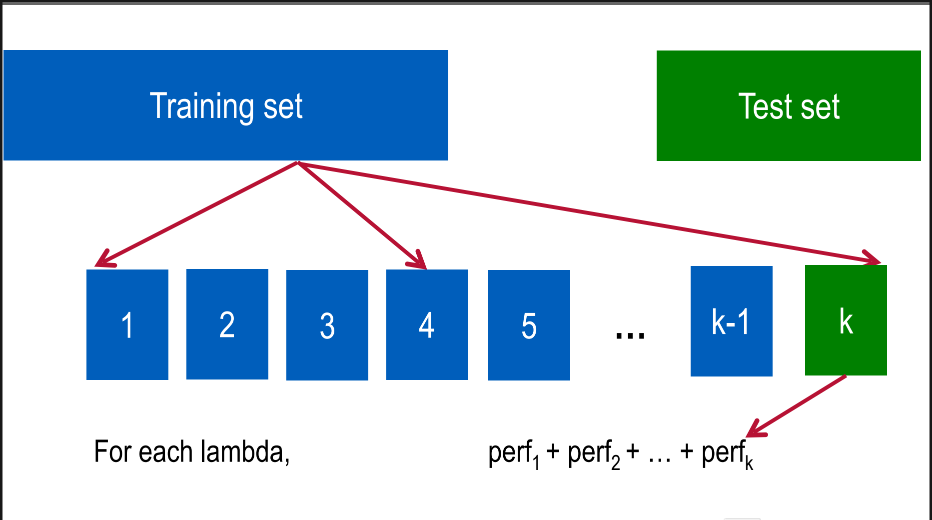
Cross-validation



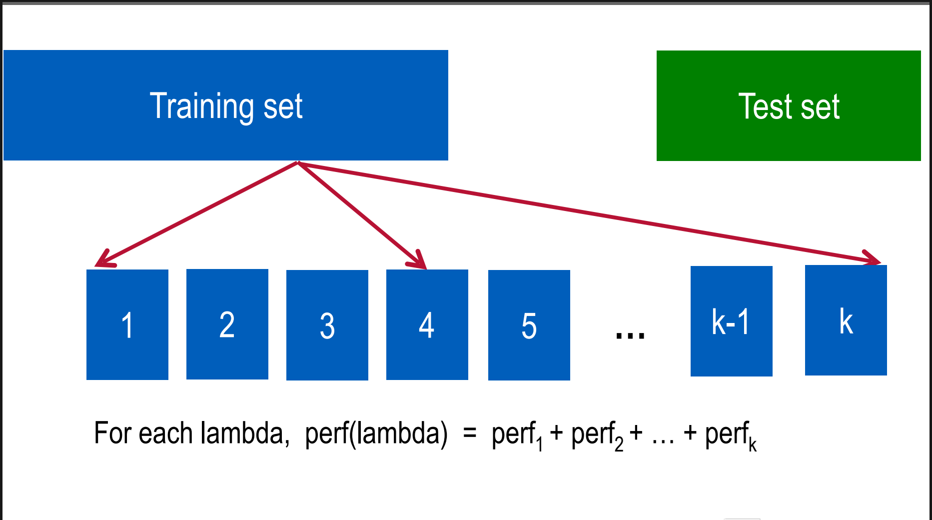
Cross-validation



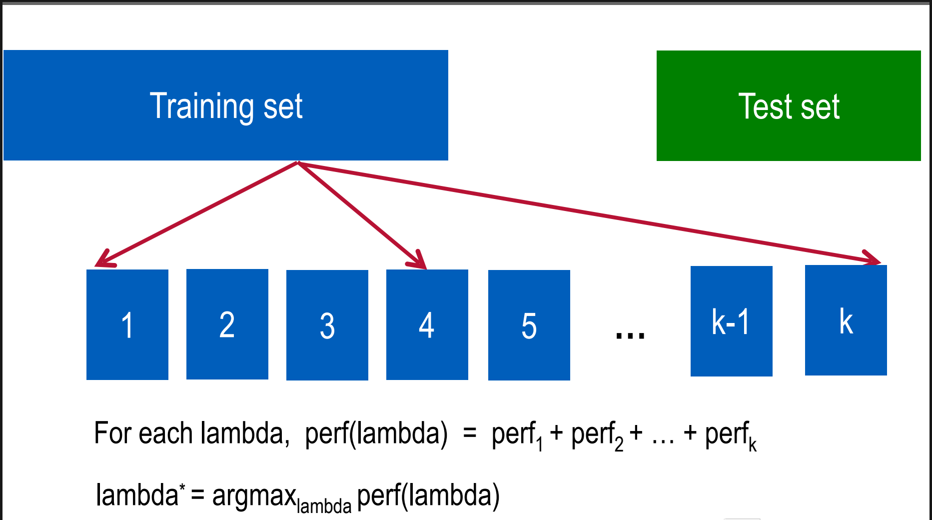
Cross-validation



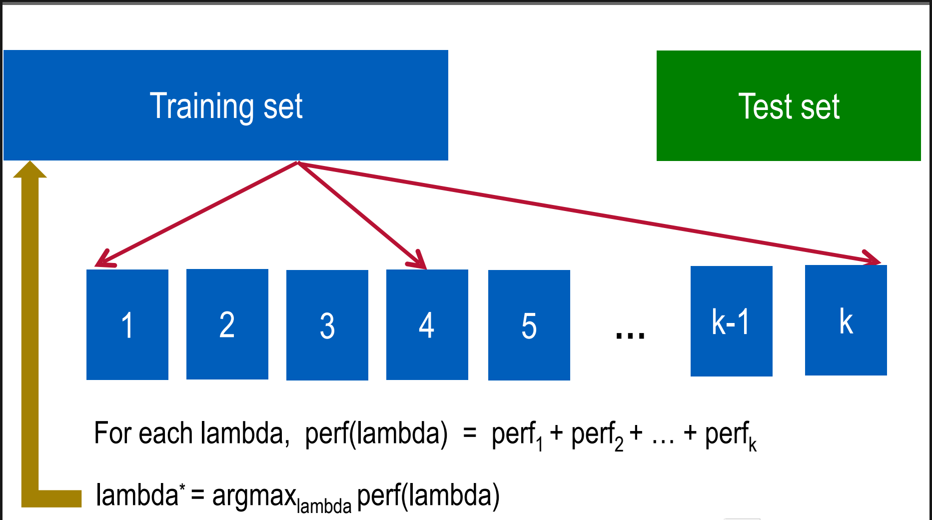
Cross-validation



Cross-validation



Cross-validation



Cross-validation