

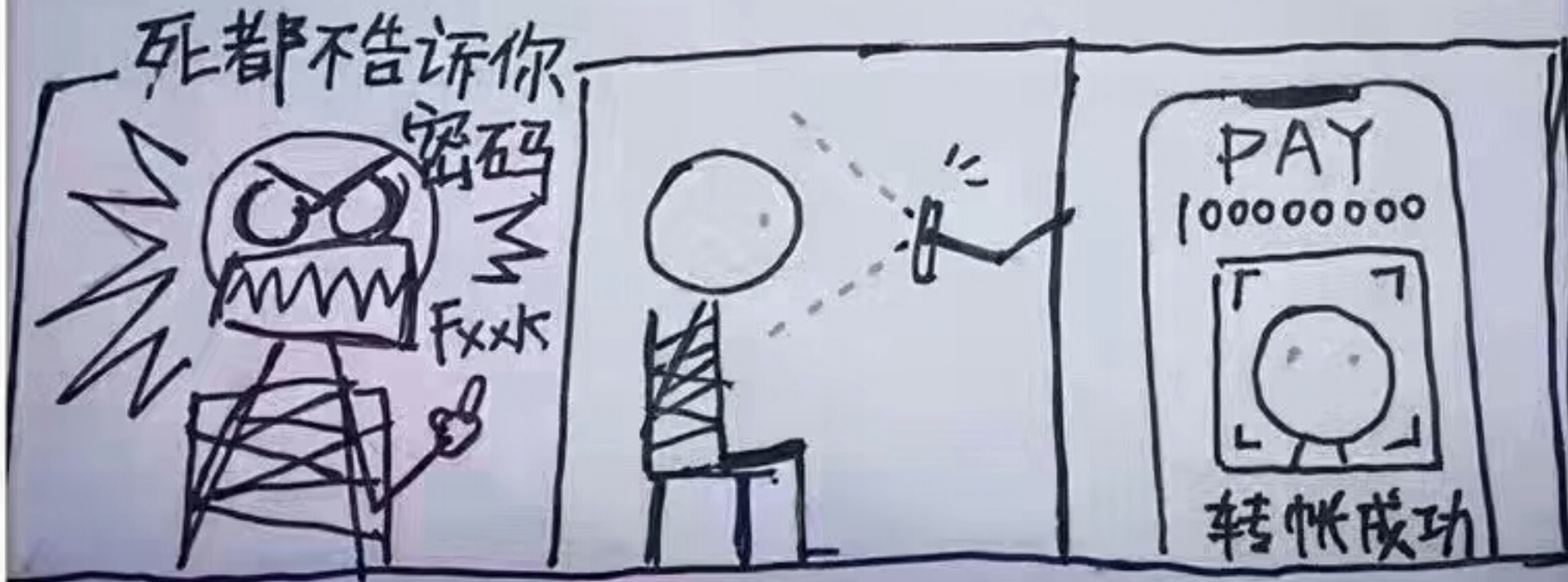
CS489/698: Intro to ML

Lecture 02: Linear Regression



UNIVERSITY OF
WATERLOO

I'd rather die than telling you
my password!



Transfer success!

Outline

- Announcements
- Linear Regression
- Regularization
- Cross-validation

Outline

- Announcements
- Linear Regression
- Regularization
- Cross-validation

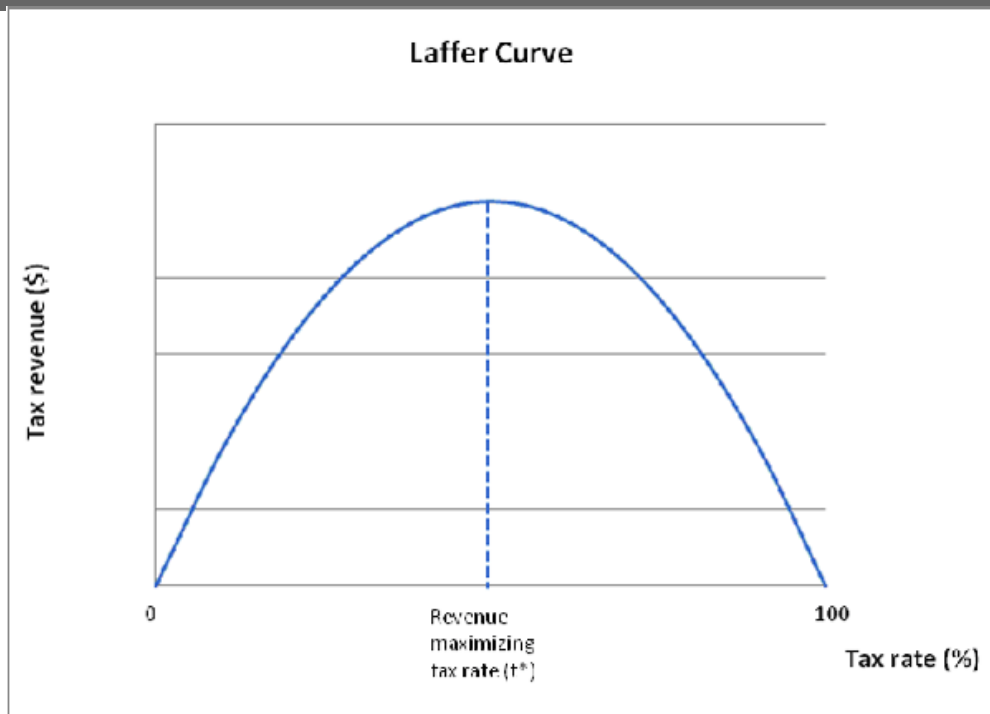
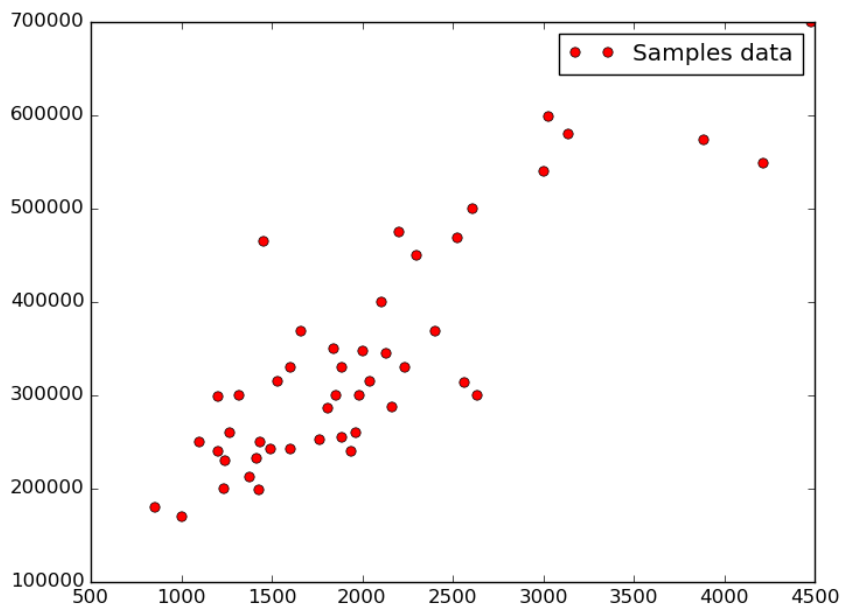
Announcements

- Assignment 1 is out.
 - Due in two weeks
- TA office hour?
- Enrollment
 - CS698: permission numbers sent
 - CS489: ~10 seats available on Quest, ask CS advisors!

Outline

- Announcements
- Linear Regression
- Regularization
- Cross-validation

How much should I bid for?



- Interpolation vs. Extrapolation
- Linear vs. Nonlinear

Regression

- Given a pair (X, Y) , find function f such that

$$f(X) \approx Y$$

- X : feature vector, d -dim real vector
- Y : response, m -dim real vector ($m=1$ say)
- Two problems:
 - (X, Y) is uncertain: samples from an **unknown distribution**
 - How to measure the error: need a **loss** function


Risk minimization

- Minimize the expected loss, aka **risk**

$$\min_{f: \mathcal{X} \rightarrow \mathcal{Y}} \underbrace{\mathbf{E}[L(f(X), Y)]}_{\text{risk}} \quad \begin{aligned} L &: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbf{R}_+ \\ L(y, y) &\equiv 0 \end{aligned}$$

- Which loss to use?
 - Not always clear; convenience dominates for now
- Least squares: $\min_f \mathbf{E} \|f(X) - Y\|_2^2$

The regression function

$$\mathbf{E}\|f(X) - Y\|_2^2 = \mathbf{E}\|f(X) - \mathbf{E}(Y|X)\|_2^2 + \underbrace{\mathbf{E}\|\mathbf{E}(Y|X) - Y\|_2^2}_{\text{Inherent noise variance}}$$


- Regression function

$$f^*(X) = m(X) = \mathbf{E}(Y|X)$$

- Many ways to estimate $m(X)$
- **Simplest:** Let's assume it is linear (affine)!

Linear regression

Assumption: $m(X) = \mathbf{E}(Y|X) = XA + \mathbf{b}$

- Dream: $\min_{A, \mathbf{b}} \mathbf{E} \|XA + \mathbf{b} - Y\|_2^2$
distribution unknown...
- Law of Large Numbers: $\frac{1}{n} \sum_{i=1}^n Z_i \rightarrow \mathbf{E}(Z)$

- Reality: $\min_{A, \mathbf{b}} \frac{1}{n} \sum_{i=1}^n \|X_i A + \mathbf{b} - Y_i\|_2^2$
empirical risk



Simplification, again

$$\min_{A, \mathbf{b}} \frac{1}{n} \sum_{i=1}^n \|X_i A + \mathbf{b} - Y_i\|_2^2$$

$$W \leftarrow \begin{pmatrix} A \\ \mathbf{b} \end{pmatrix}$$
$$X_i \leftarrow (X_i, 1)$$

$$\min_W \frac{1}{n} \sum_{i=1}^n \|X_i W - Y_i\|_2^2$$

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

$$\min_{W \in \mathbf{R}^{(d+1) \times m}} \|\mathbf{X}W - \mathbf{Y}\|_F^2$$

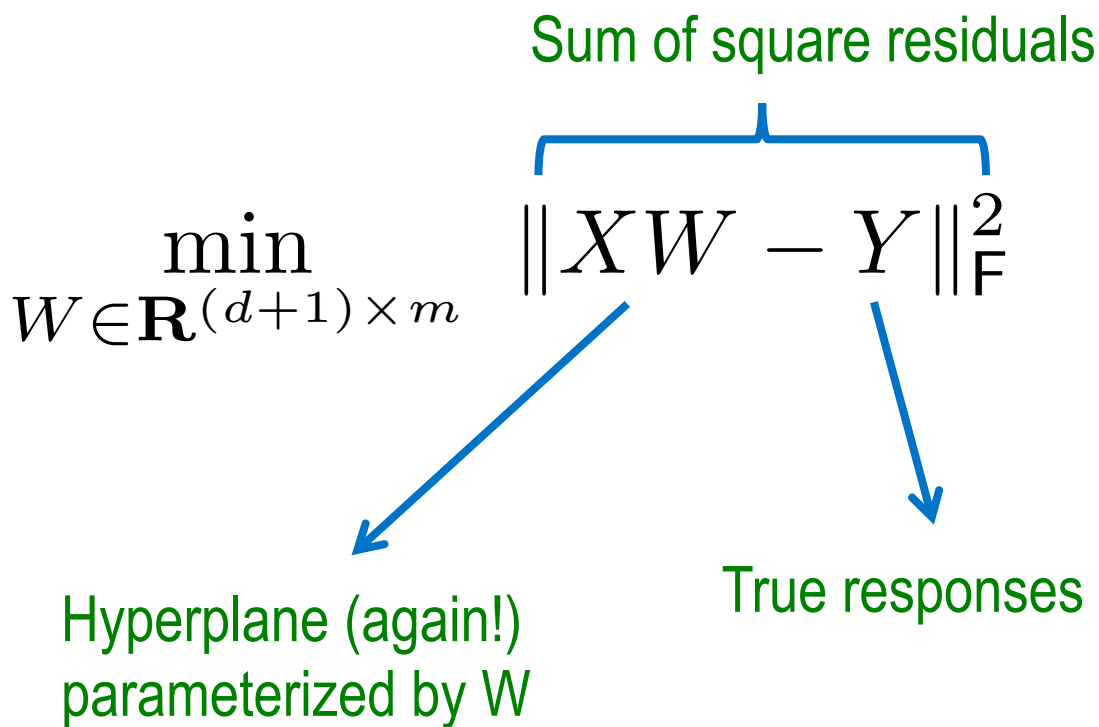
Finally

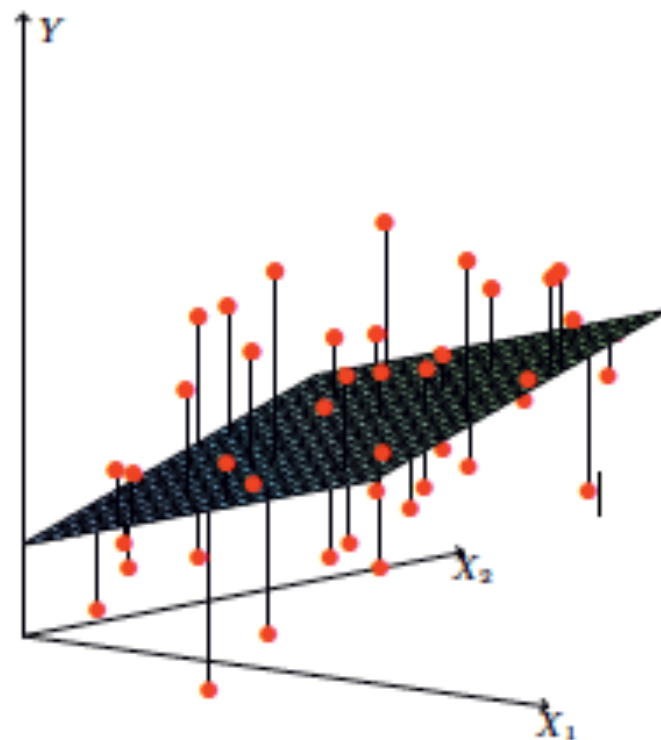
Sum of square residuals

$$\min_{W \in \mathbf{R}^{(d+1) \times m}} \|XW - Y\|_F^2$$

Hyperplane (again!)
parameterized by W

True responses





Why least squares?

$$\min_{W \in \mathbf{R}^{(d+1) \times m}} \|XW - Y\|_F^2$$

Theorem (Sondermann'86; Friedland and Torokhti'07; Yu and Schuurmans'11)

Among all minimizers of $\min_W \|AWB - C\|_F$, $W = A^+CB^+$ is the one that has minimal F-norm.

Pseudo-inverse A^+ is the **unique** matrix G such that

$$AGA = A, \quad GAG = G, \quad (AG)^T = AG, \quad (GA)^T = GA$$

Singular Value Decomposition

$$A = USV^T$$

$$A^+ = VS^{-1}U^T$$

Optimization detour

$$\min_x f(x)$$

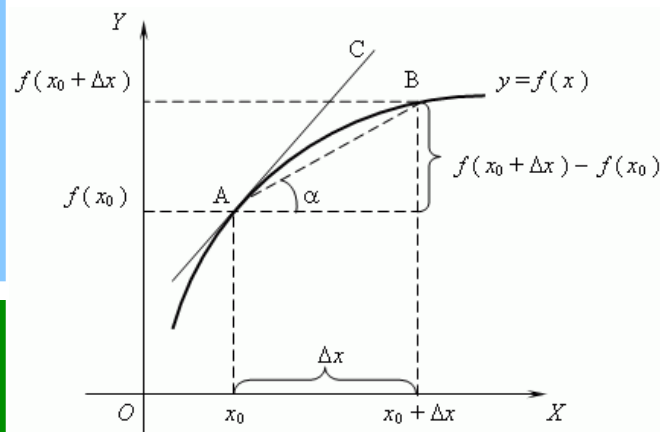
Fermat's Theorem. **Necessarily** $Df(x) = 0$

(Fréchet) Derivative at x .

$$\lim_{\delta \rightarrow 0} \frac{|f(x + \delta) - f(x) - Df(x)\delta|}{|\delta|} = 0$$

Example. $f(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{x}^T \mathbf{b} + c$

$$Df(\mathbf{x}) = (\mathbf{A} + \mathbf{A}^T)\mathbf{x} + \mathbf{b}$$



Solving least squares

$$\min_{W \in \mathbf{R}^{(d+1) \times m}} \|XW - Y\|_F^2 = W^\top (X^\top X)W - 2W^\top X^\top Y + Y^\top Y$$



$$X^\top XW = X^\top Y$$

Normal
Equation

- $X^\top X$ may not be invertible, but there is always a solution
- Even invertible, **never ever compute $W = (X^\top X)^{-1}X^\top Y$!**
- Instead, solve the linear system

Prediction

- Once have W , can predict

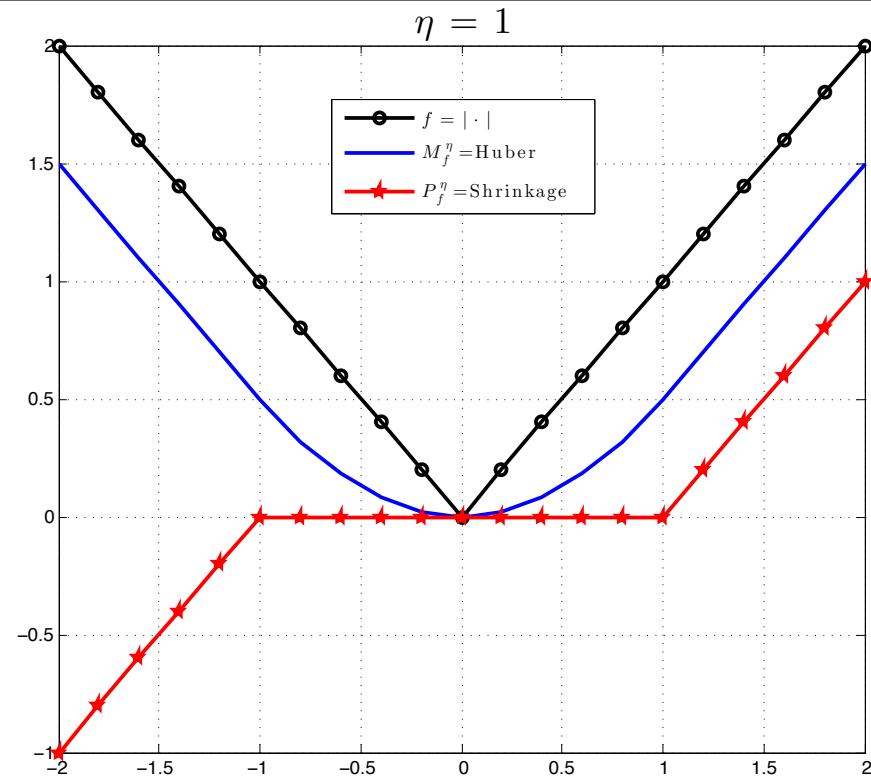
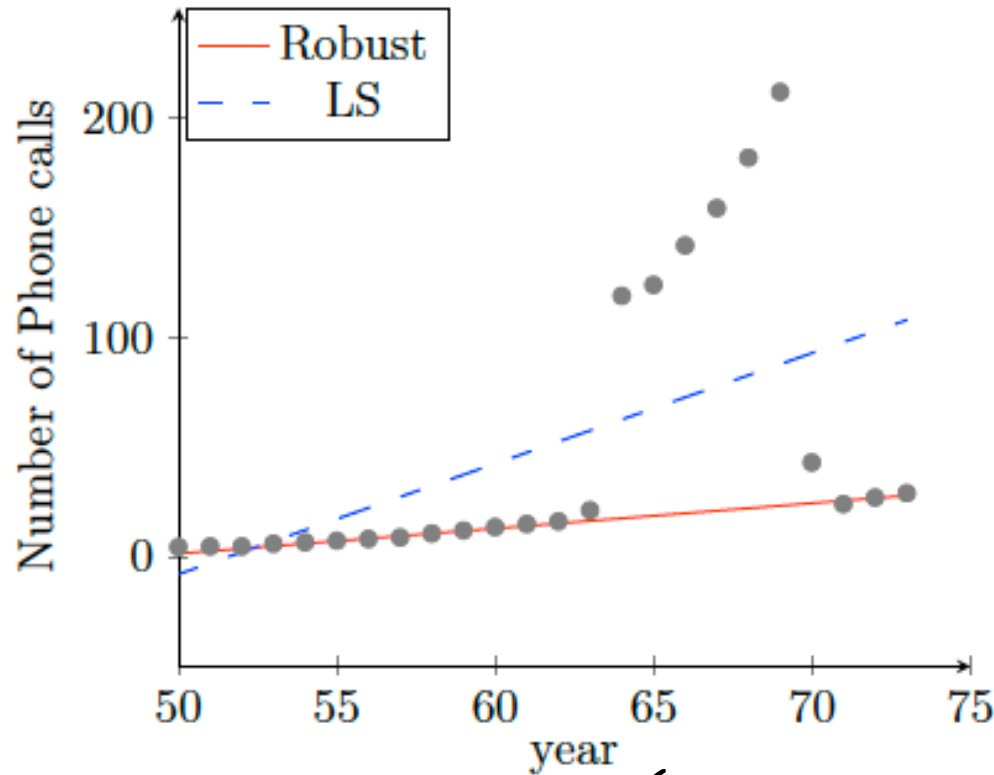
$$\hat{Y} = X_{\text{test}} W$$

- How to evaluate?

$$(Y_{\text{test}} - \hat{Y})^2$$

- Sometimes we evaluate using a different $L(Y_{\text{test}}, \hat{Y})$
 - Leads to a beautiful theory of **calibration**

Robustness



$$H(\hat{y}, y) = \begin{cases} \frac{1}{2}(\hat{y} - y)^2, & |\hat{y} - y| \leq \delta \\ \delta(|\hat{y} - y| - \frac{1}{2}\delta), & |\hat{y} - y| \geq \delta \end{cases}$$

Gauss vs. Laplace

280

S. PORTNOY AND R. KOENKER



FIG. 1. *The Gaussian Hare and the Laplacian Tortoise*: this picture is a slightly “retouched” version of a wood engraving by J. J. Grandville from “*Fables de La Fontaine*” (published in Paris, 1838). The portrait of Gauss is taken from an 1803 portrait by J. C. A. Schwartz. The portrait of Laplace appears in “*Cauchy: Un Mathématicien Légitimiste au XIXe Siècle*,” by Bruno Belhoste (Belin, Paris).

Multi-task learning

$$X^{\top} X W = X^{\top} Y$$

- Everything we've shown still holds if Y is m -dim

- But, can solve each column of Y **independently**

$$X^{\top} X W_{:j} = X^{\top} Y_{:j}$$

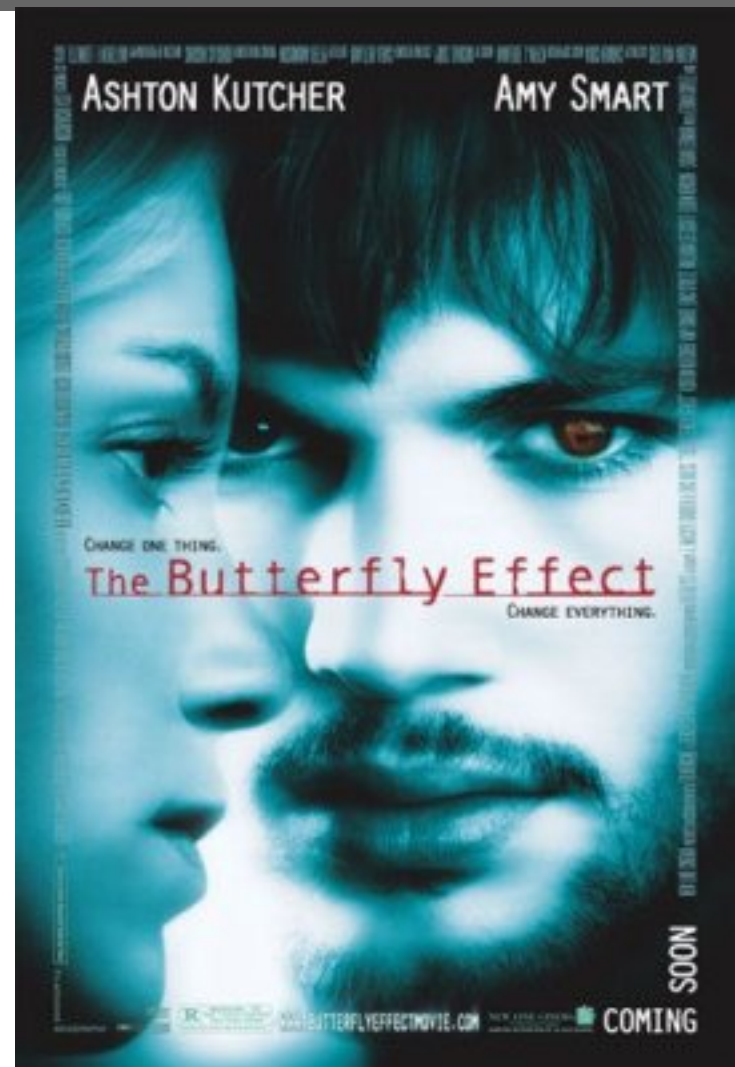
- Things are more interesting if we had **regularization**

Outline

- Announcements
- Linear Regression
- Regularization
- Cross-validation


Ill-posedness

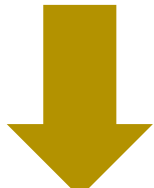

- Let $x_1=0$, $x_2=\epsilon$, $y_1=1$, $y_2=-1$
- $X =$ $y =$
- $w = X^{-1}y =$
- Slight perturbation leads to chaotic behaviour



Tikhonov regularization (Hoerl and Kennard'70)

$$\min_{W \in \mathbf{R}^{(d+1) \times m}} \|XW - Y\|_F^2 + \lambda \|W\|_F^2$$

Ridge regression 

  Reg. constant (hyperparameter)

$$(X^\top X + \lambda I)W = X^\top Y$$

- With positive lambda, slight perturbation in input leads to **proportional (wrt 1/lambda)** perturbation in output

Data augmentation

$$\min_{W \in \mathbf{R}^{(d+1) \times m}} \|XW - Y\|_F^2 + \lambda \|W\|_F^2$$



$$\min_{W \in \mathbf{R}^{(d+1) \times m}} \|\tilde{X}W - \tilde{Y}\|_F^2$$

$$\tilde{X} = \begin{bmatrix} X \\ \sqrt{\lambda}I \end{bmatrix} \quad \tilde{Y} = \begin{bmatrix} Y \\ \mathbf{0} \end{bmatrix}$$

Sparsity

- Ridge regression weight is always dense
 - Computationally heavy
 - Interpretationally cumbersome
- Lasso (Tibshirani'96)

$$\min_{\|W\|_1 \leq C} \|XW - Y\|_F^2$$

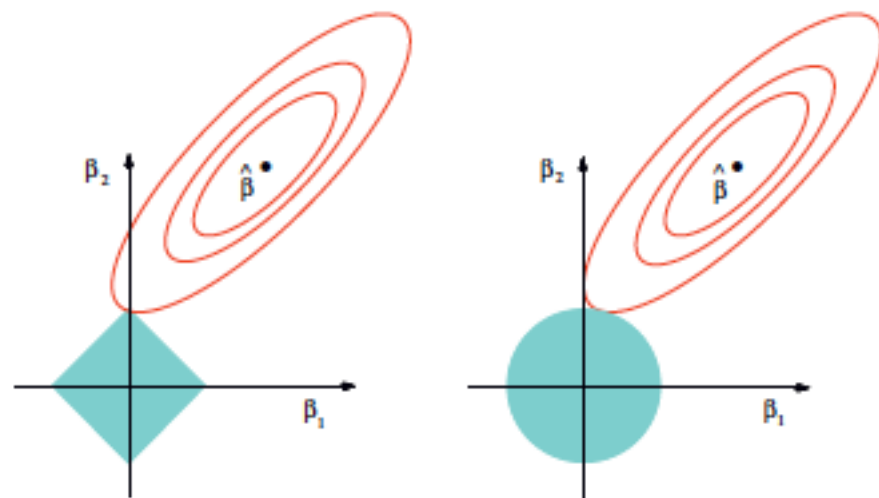
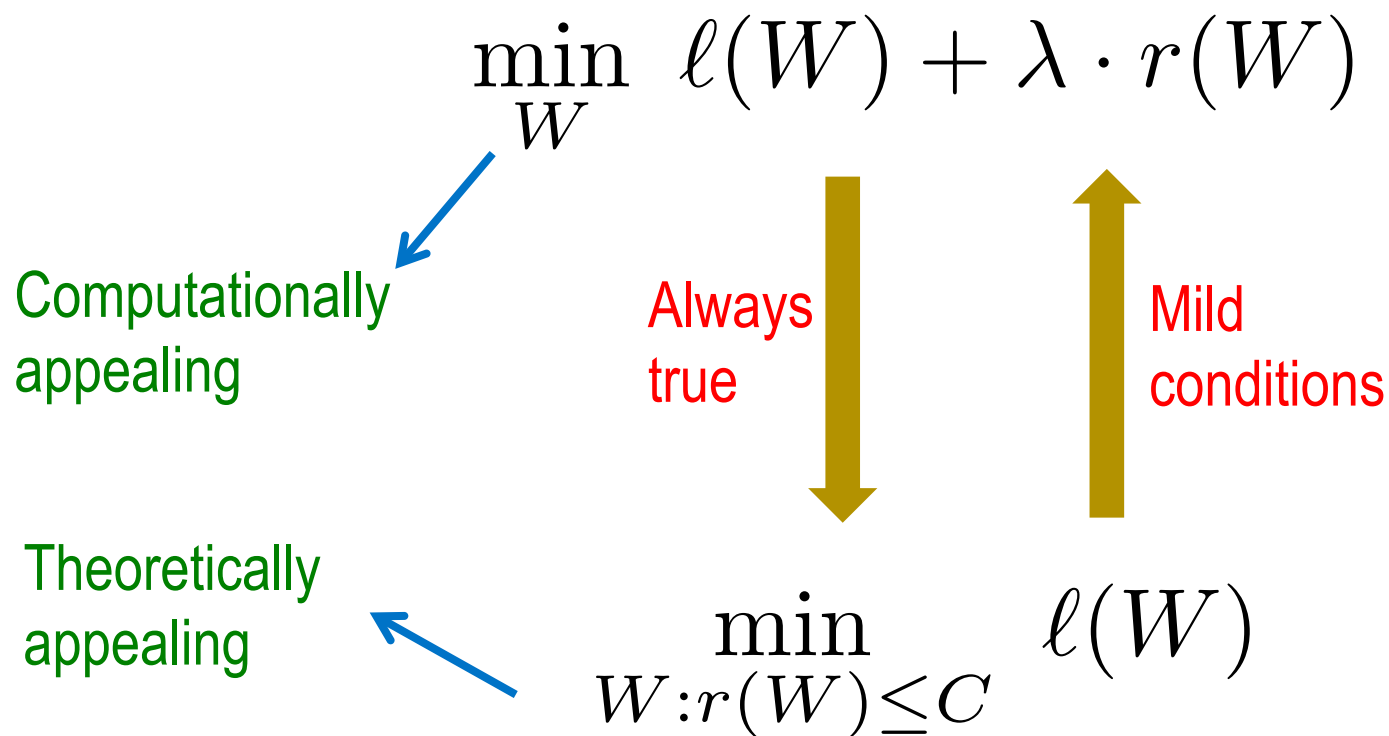


FIGURE 3.11. Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function.

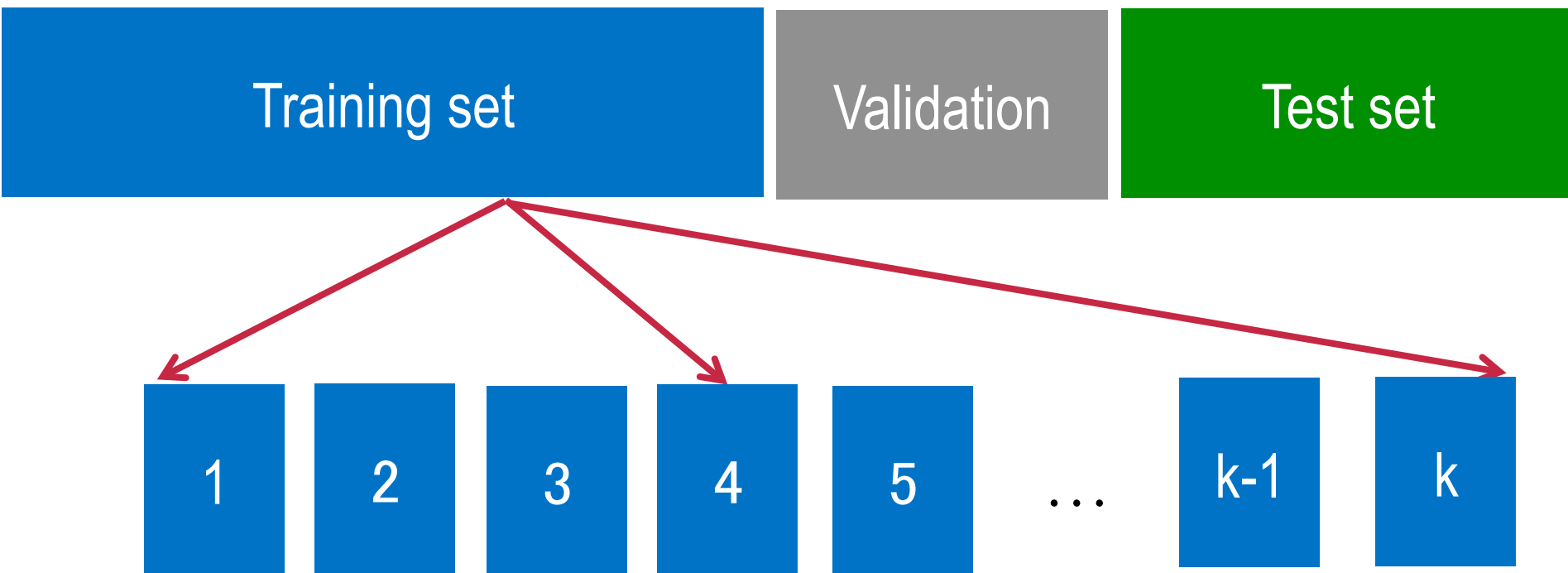
Regularization vs. Constraint



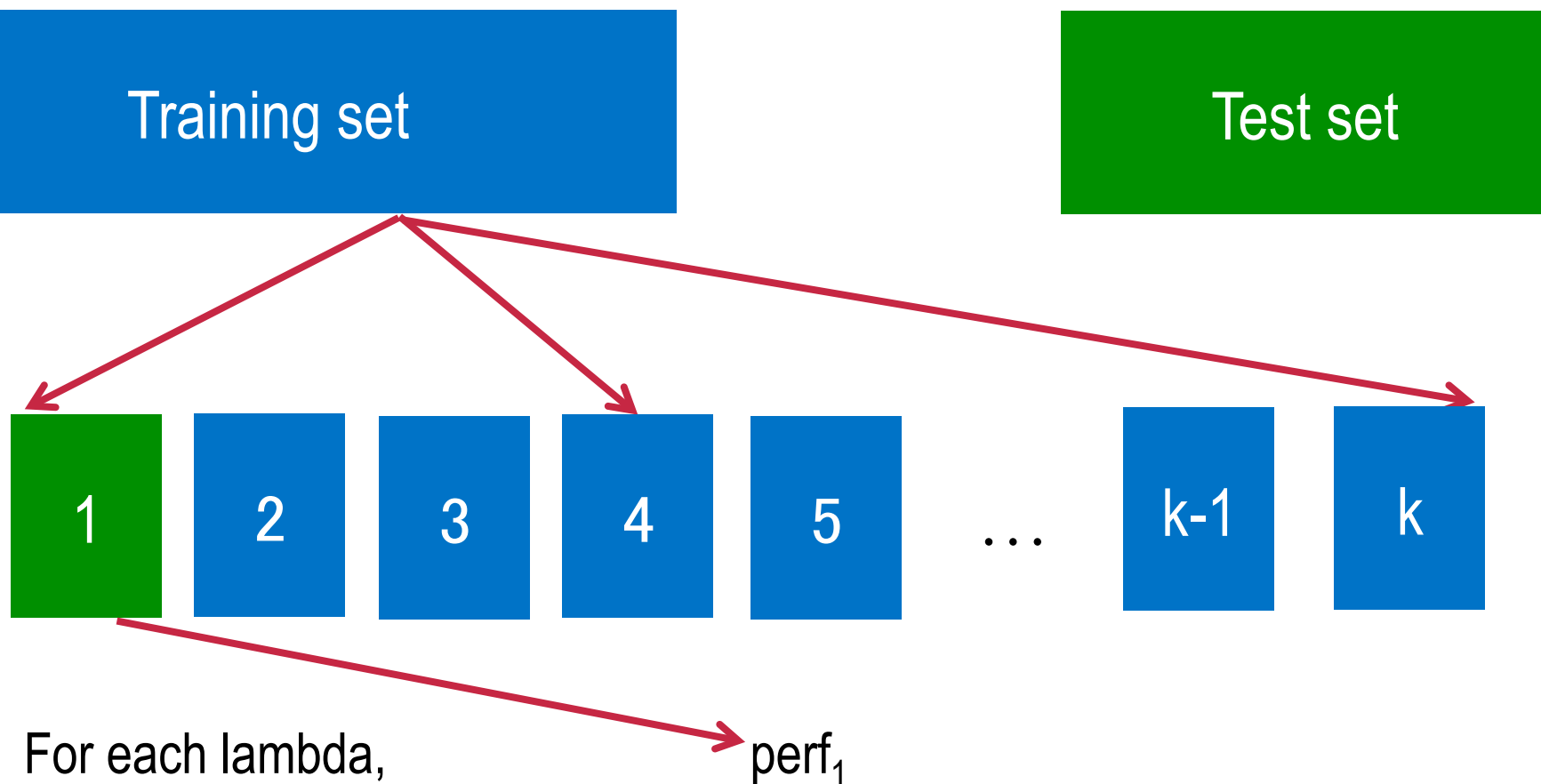
Outline

- Announcements
- Linear Regression
- Regularization
- Cross-validation

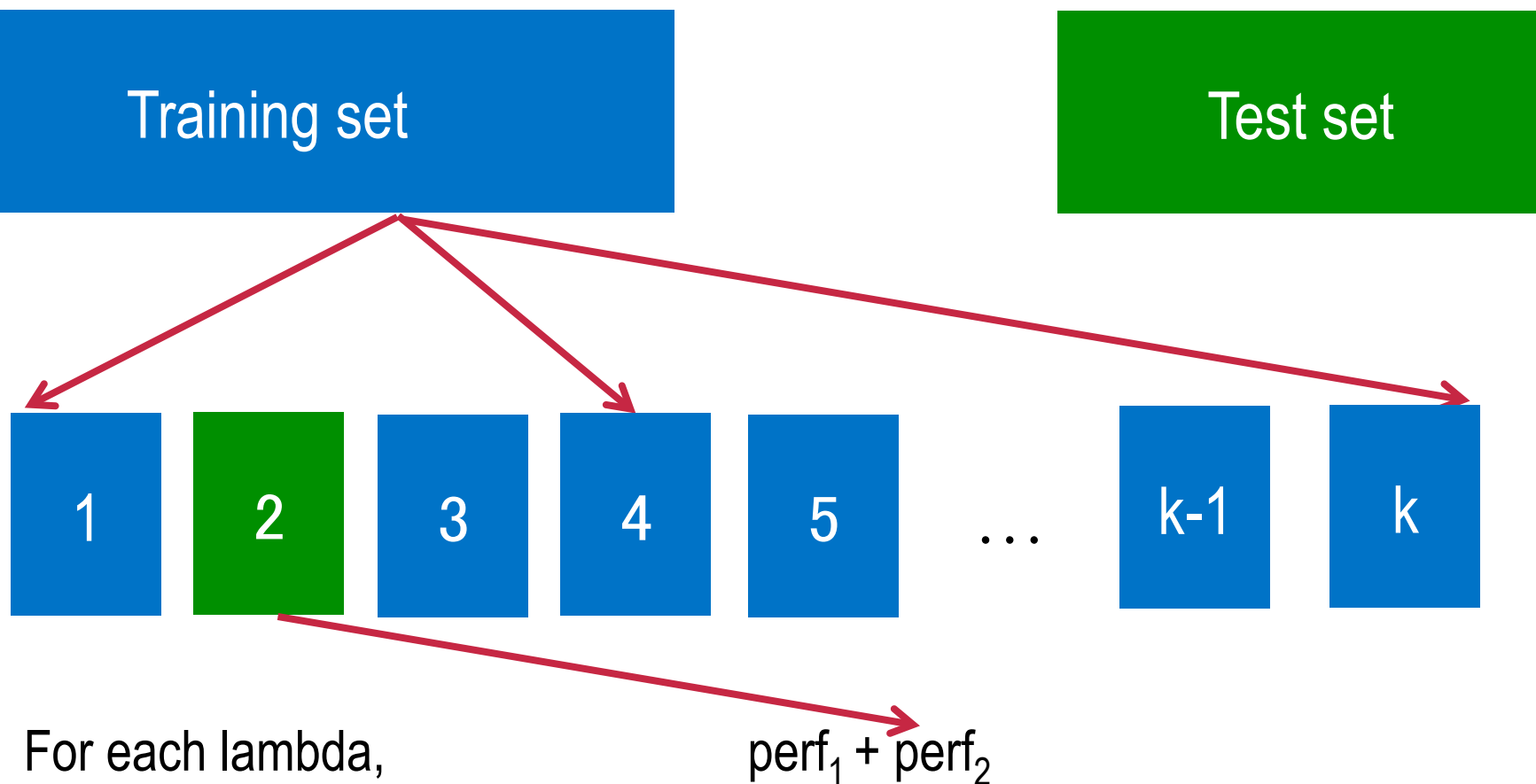
Cross-validation



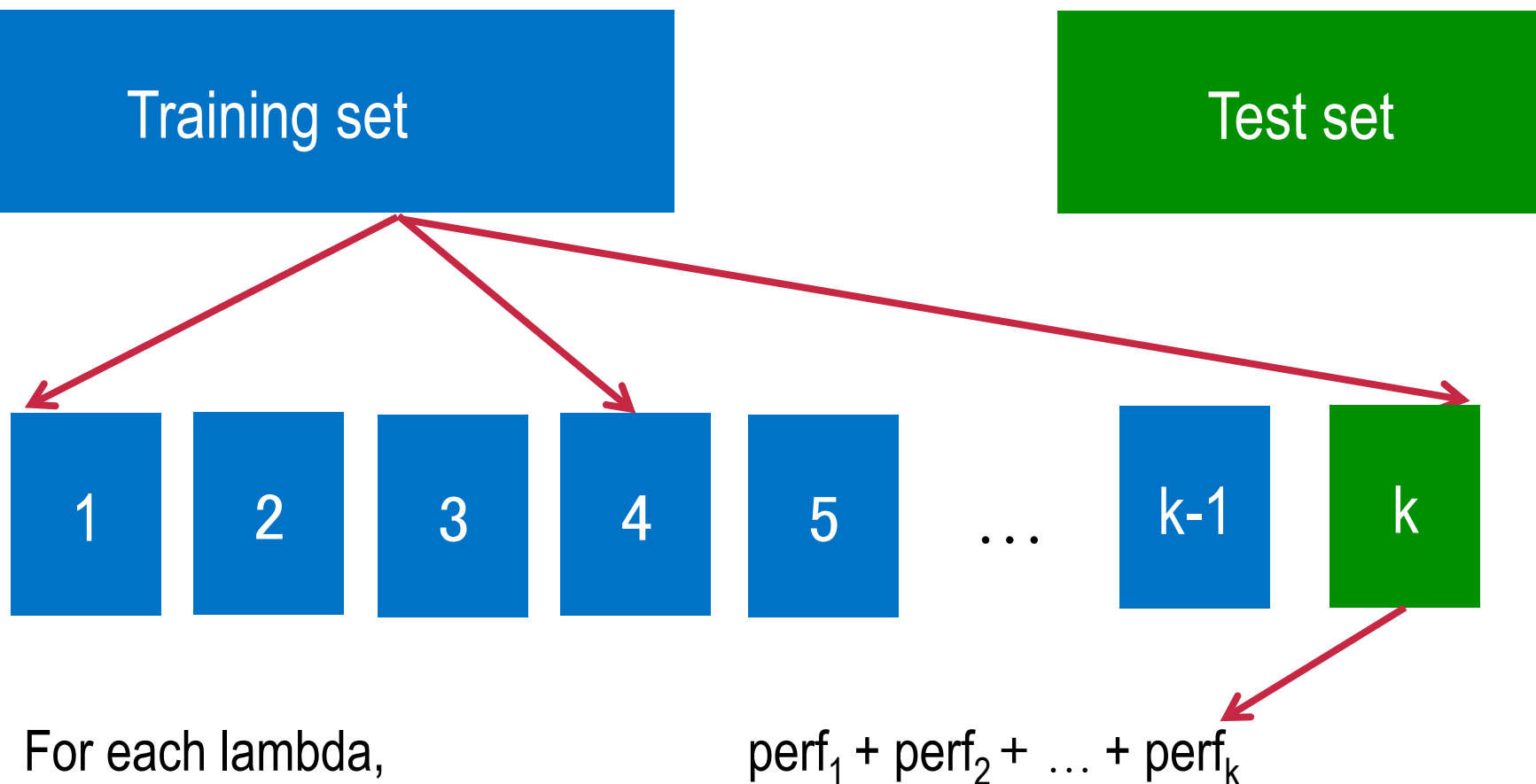
Cross-validation



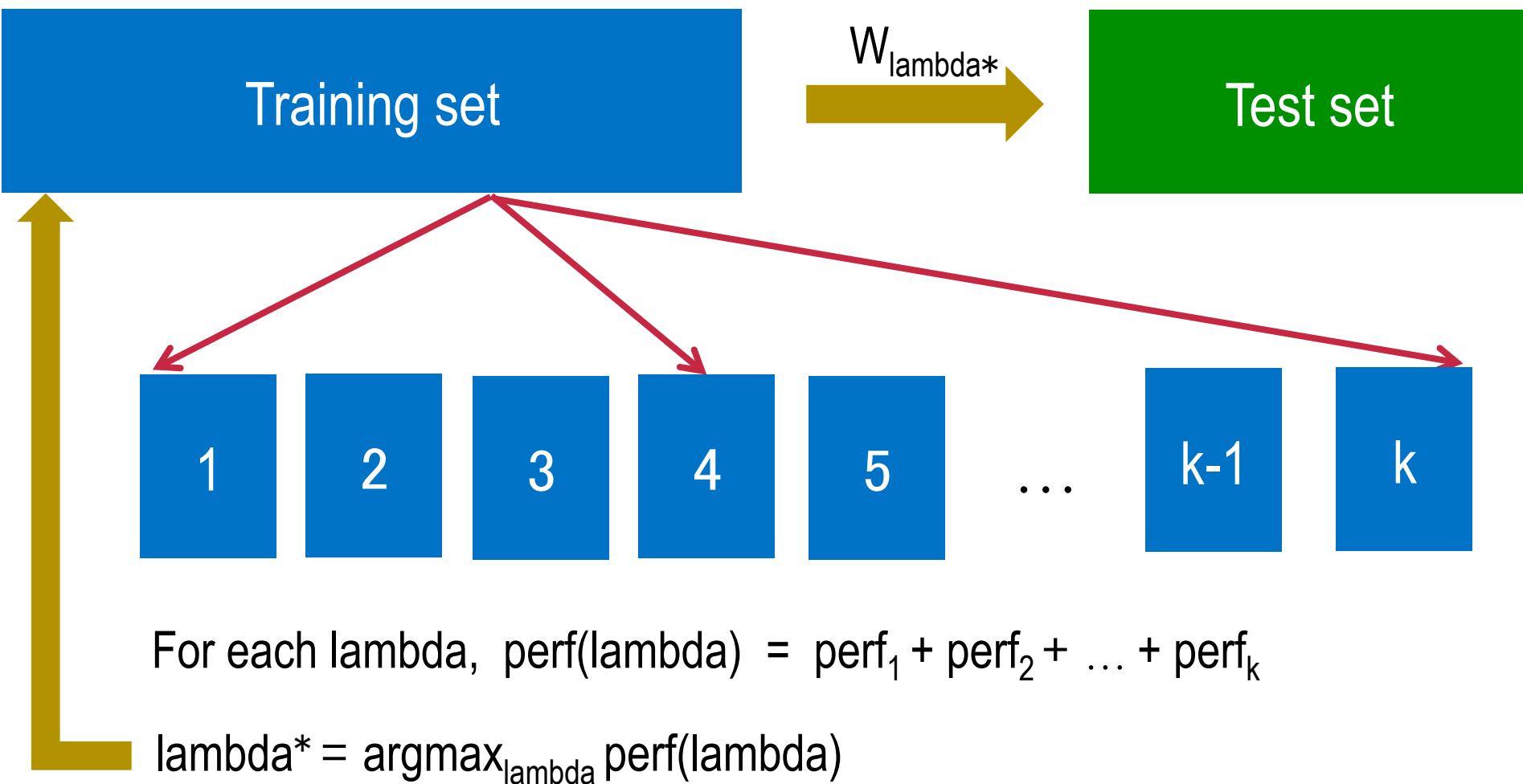
Cross-validation



Cross-validation



Cross-validation



Questions?

