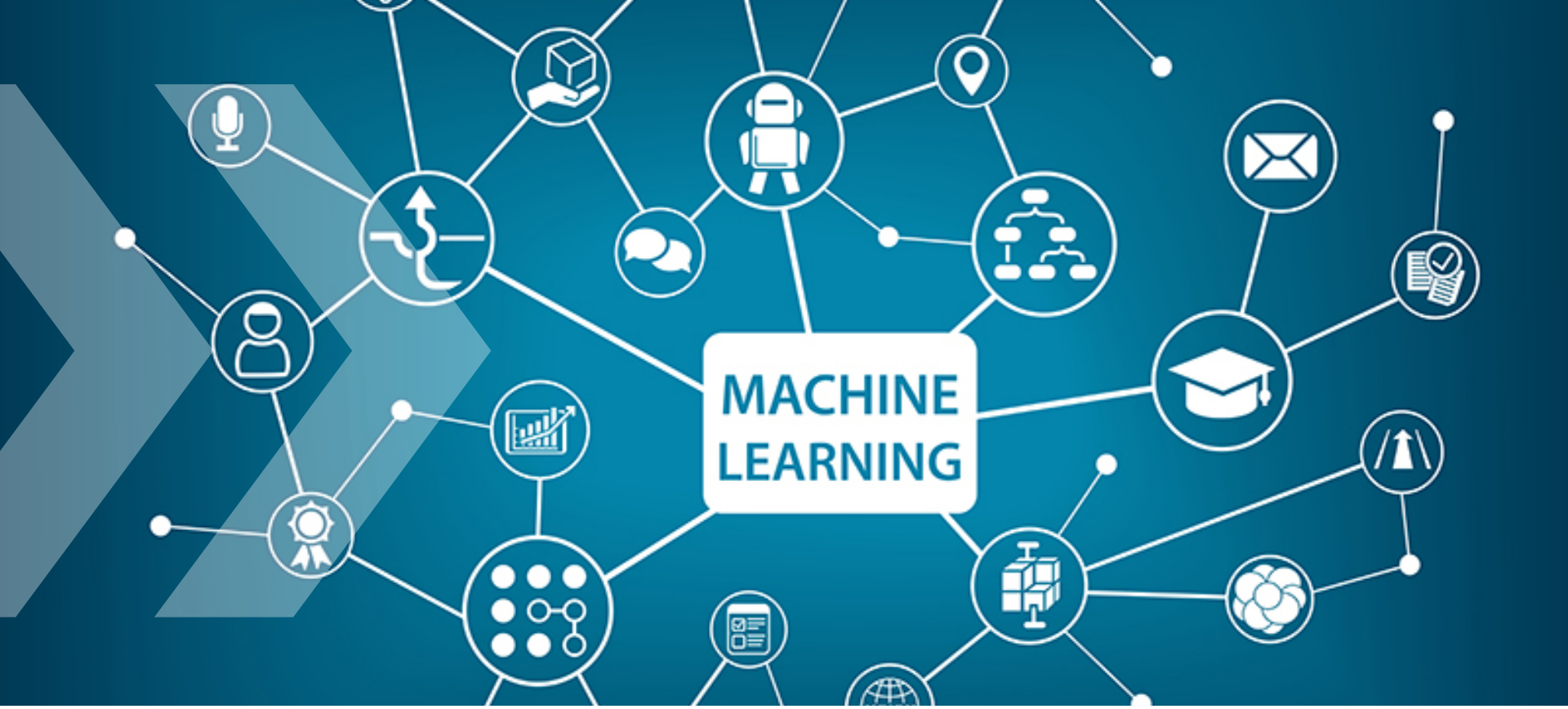


CS480/680: Intro to ML

Lecture 02: Linear Regression

UNIVERSITY OF WATERLOO

Outline

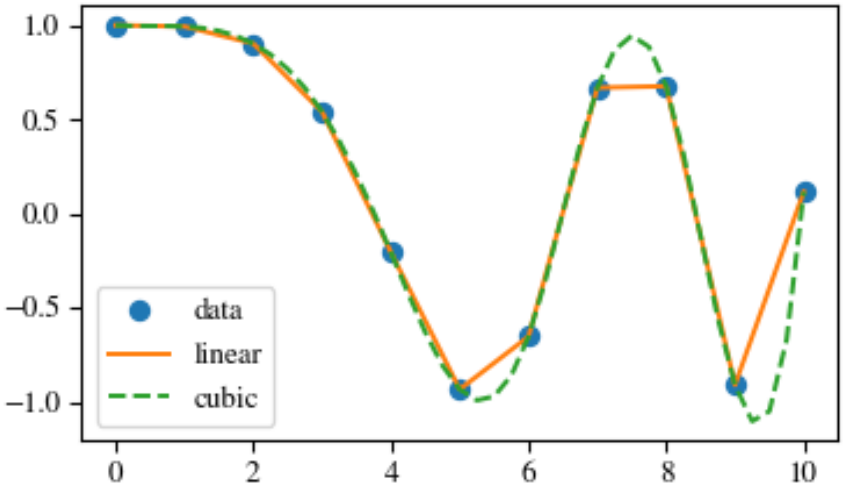
- Linear Regression
- Regularization
- Cross-validation

Regression

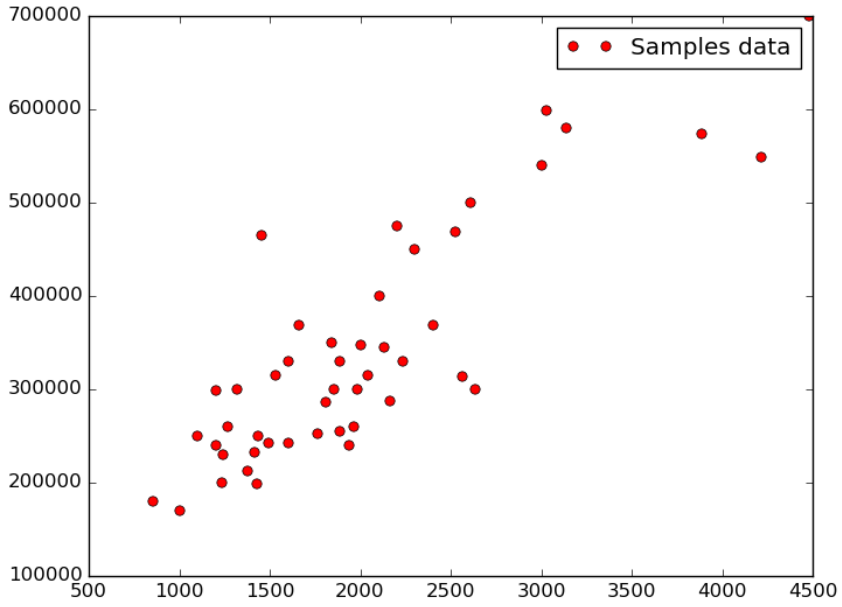
- Given pairs ($\mathbf{x}_i$, $\mathbf{y}_i$), find function f such that

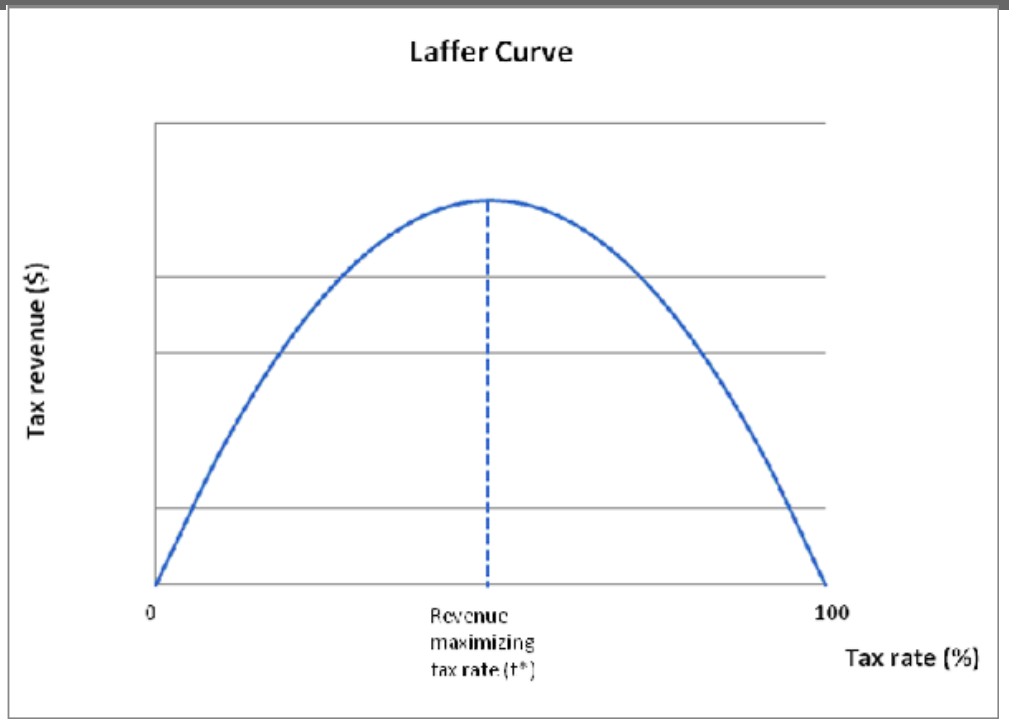
$$f(\boldsymbol{x}_i) \approx \boldsymbol{y}_i$$

  - $\mathbf{x}_i$: feature vector, d-dim real vector
  - $\mathbf{y}_i$: response, m-dim real-valued vector (m=1 say)



How much should I bid for?





- Prior knowledge
- Linear vs. Nonlinear

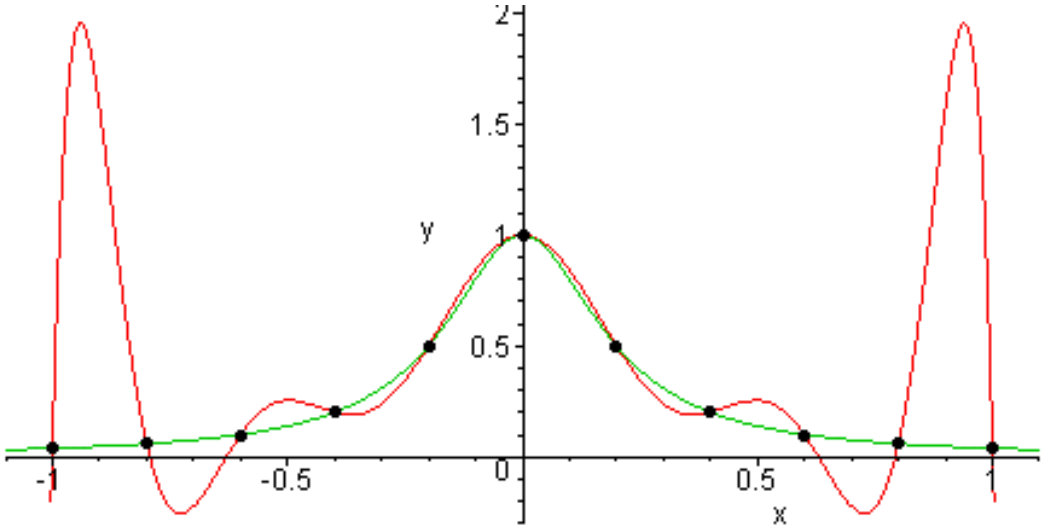
UNIVERSITY OF WATERLOO

Exact interpolation

Theorem. For any **finite** data set $D = \{(\boldsymbol{x}_i, \boldsymbol{y}_i): i = 1, \ldots, n\}$, there exist infinitely many functions f so that for all i,

$$f(\boldsymbol{x}_i) = \boldsymbol{y}_i$$

- On new data $\boldsymbol{x}$, predict
$$\hat{y} = f(x)$$
- Can be wildly different for different f !



UNIVERSITY OF WATERLOO

Statistical Learning

- Assume data (X_i, Y_i) iid samples from an unknown distr.
- Least squares regression:

difficult to evaluate

$$\min_f \underbrace{\mathbf{E}\|f(X) - Y\|_2^2}_{\text{average error}}$$

Law of large numbers

- Role of training

$$\min_f \frac{1}{n} \sum_{i=1}^{n} \|f(X_i) - Y_i\|_2^2$$

UNIVERSITY OF WATERLOO

The regression function

$$\mathbf{E}\|f(X) - Y\|_2^2 = \mathbf{E}\|f(X) - \mathbf{E}(Y|X)\|_2^2 + \mathbf{E}\|\mathbf{E}(Y|X) - Y\|_2^2$$

Inherent noise variance

- Regression function

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

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

Linear Least-squares Regression

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

$$W \leftarrow \begin{pmatrix} A \\ \mathbf{b} \end{pmatrix}$$

$$X_i \leftarrow (X_i, 1)$$

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

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

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

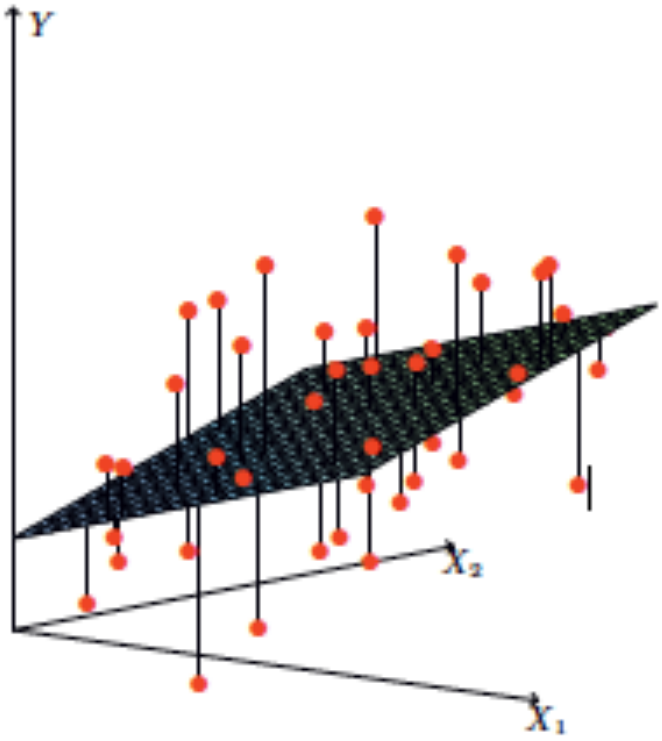
Finally

Sum of square residuals

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

Hyperplane (again!) parameterized by W

True responses



UNIVERSITY OF WATERLOO

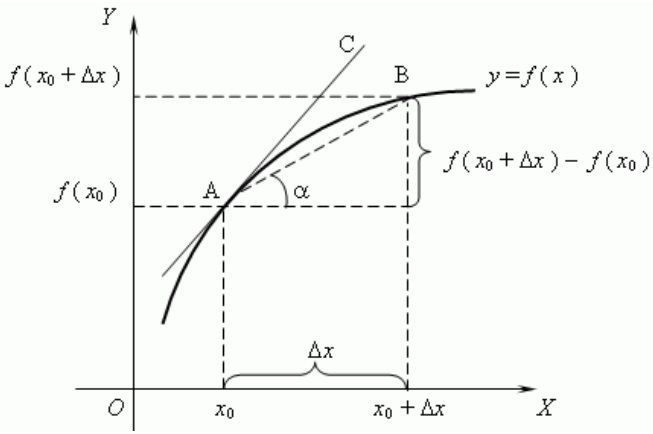
Optimization detour

$$\min_x \ f(x)$$

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

(Fréchet) Derivative at *x*.

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



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

$[Df(\mathbf{x})]^T = \nabla f(x) = (A+A^T)\mathbf{x} + \mathbf{b}$

UNIVERSITY OF WATERLOO

Solving least squares

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

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

Normal Equation

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

Prediction

- Once have W, can predict

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

- How to evaluate?

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

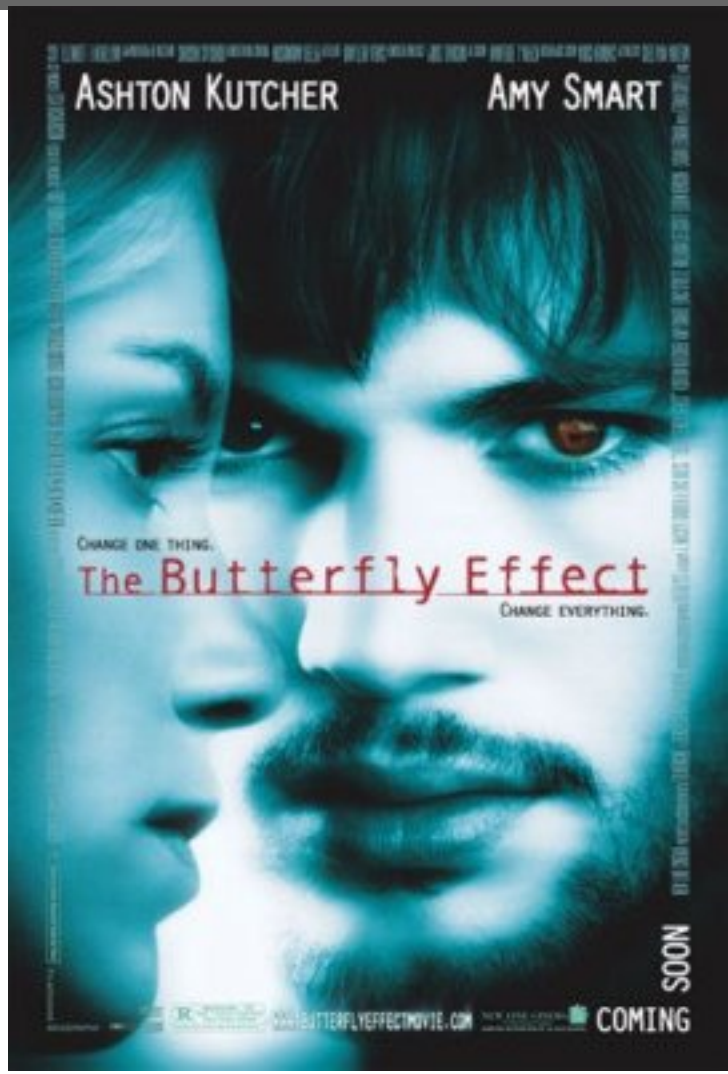
- Sometimes we evaluate using a different $L(Y_{\text{test}}, \hat{Y})$
  - Leads to a beautiful theory of calibration

Outline

- Linear Regression
- Regularization
- Cross-validation

Ill-posedness

- Let $x_1=(0, 1)$, $x_2=(\varepsilon, 1)$, $y_1=1$, $y_2=-1$
- $X = \begin{pmatrix} 0 & 1 \\ \varepsilon & 1 \end{pmatrix}, y = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$
- $w = X^{-1}y = \begin{pmatrix} -2/\varepsilon \\ 1 \end{pmatrix}$
- Slight perturbation leads to chaotic behaviour



UNIVERSITY OF WATERLOO

Tiknohov regularization (Hoerl and Kennard'70)

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

Reg. constant (hyperparameter)

Ridge regression

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

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

Data augmentation

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

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

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

Sparsity

- Ridge regression weight is always dense
  - Computationally heavy
  - Interpretation-wise cumbersome

- Lasso (Tibshirani'96)

$$\min_{\|W\|_1 \le C} \|XW - Y\|_{\mathrm{F}}^2$$

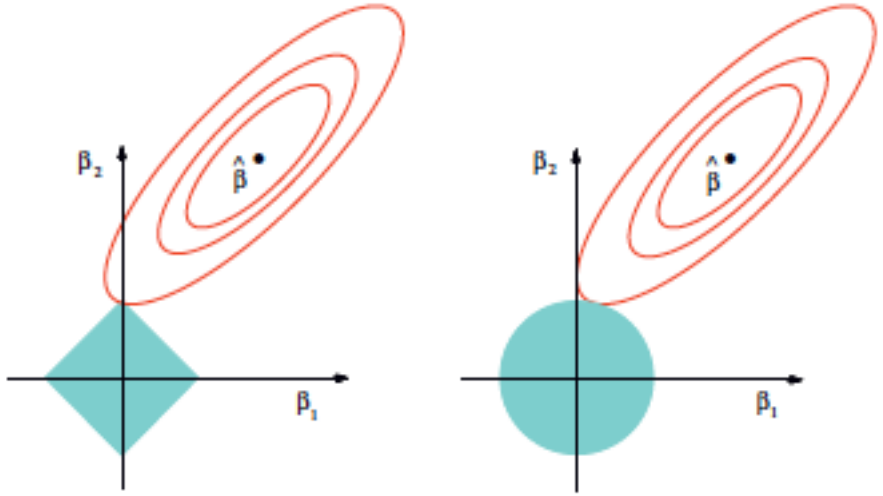
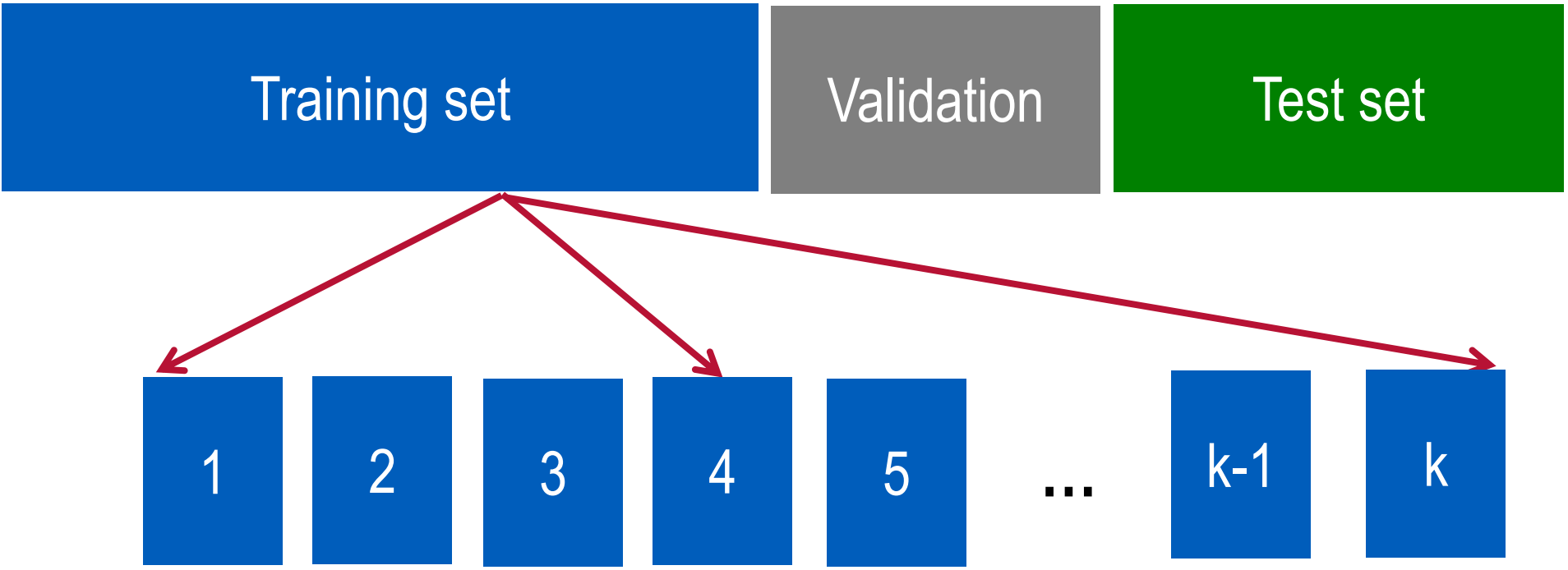


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \le t$ *and* $\beta_1^2 + \beta_2^2 \le t^2$*, respectively, while the red ellipses are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

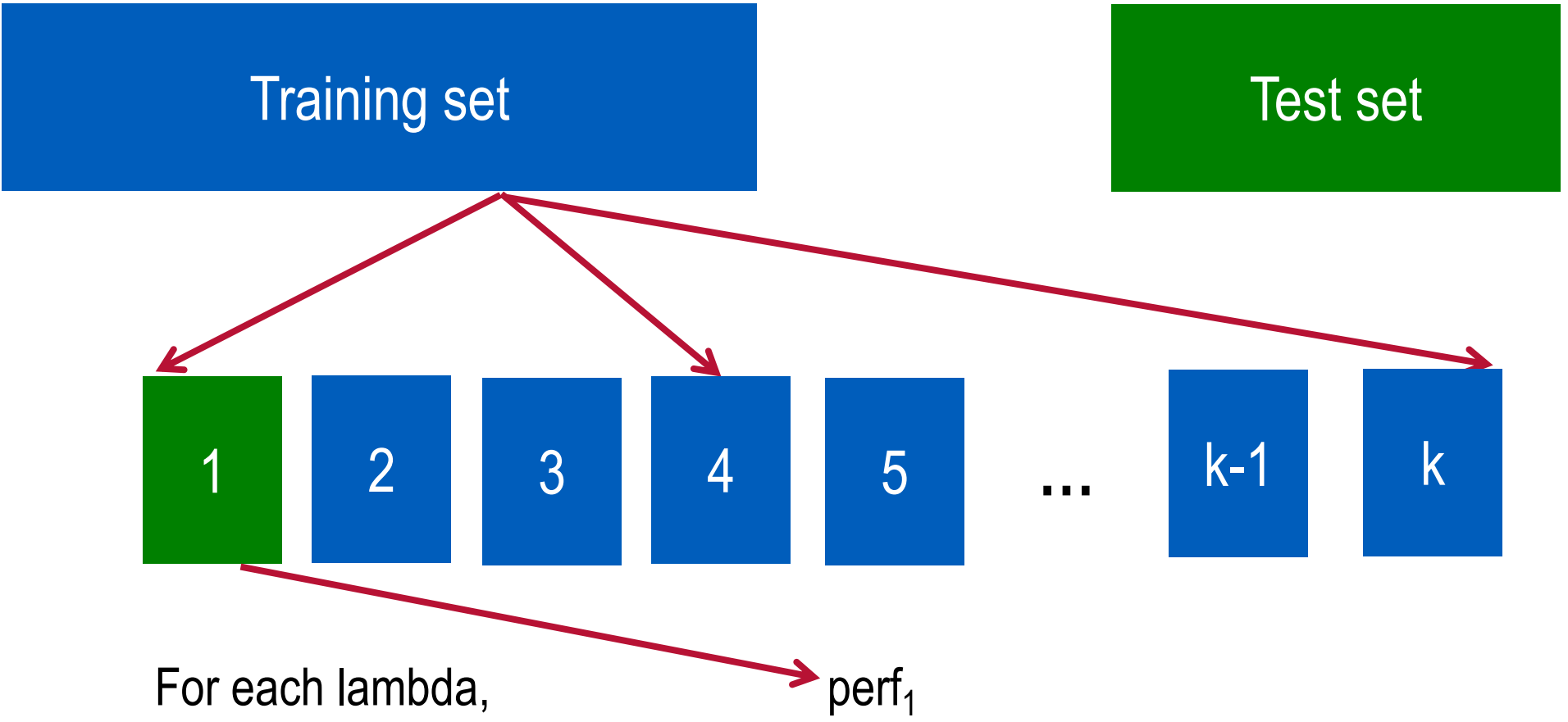
Outline

- Linear Regression
- Regularization
- Cross-validation

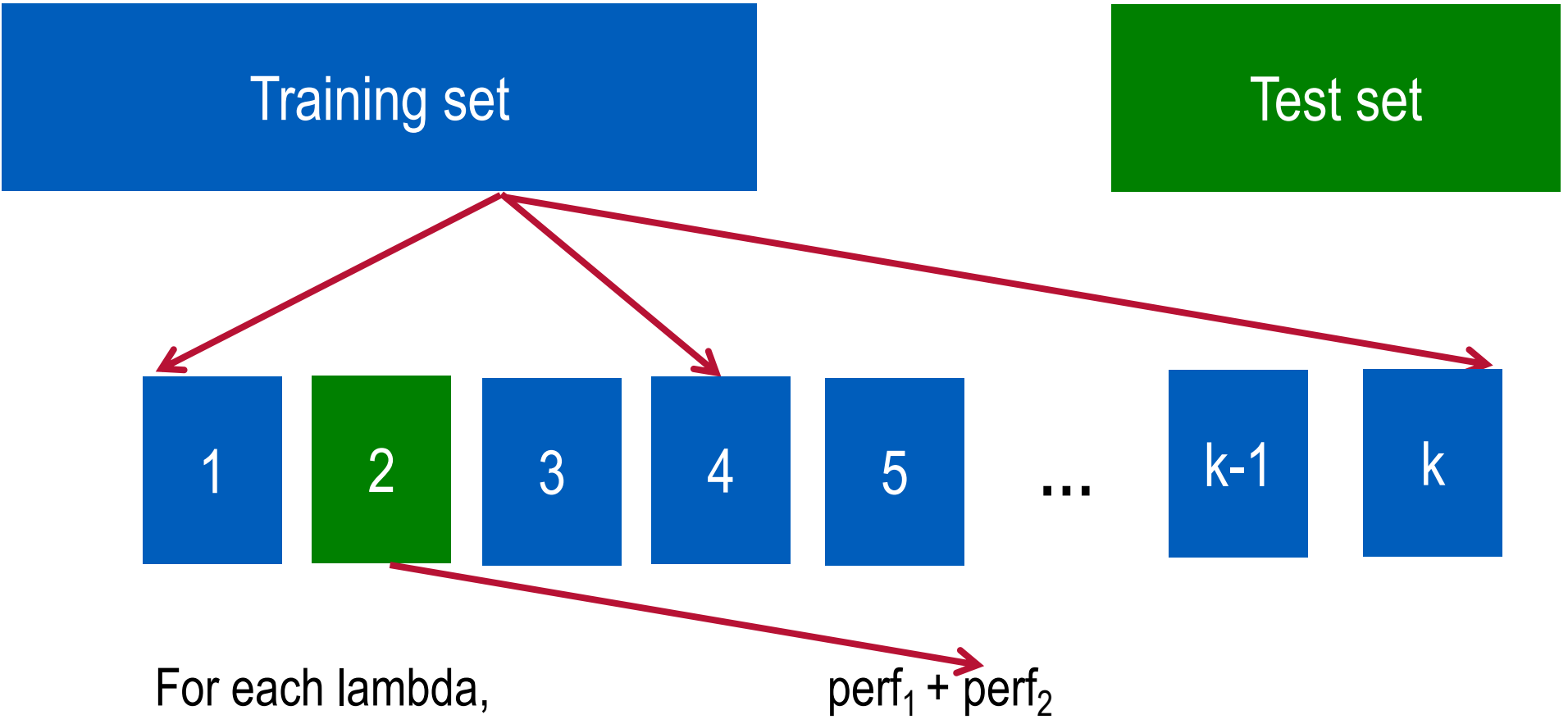
Cross-validation



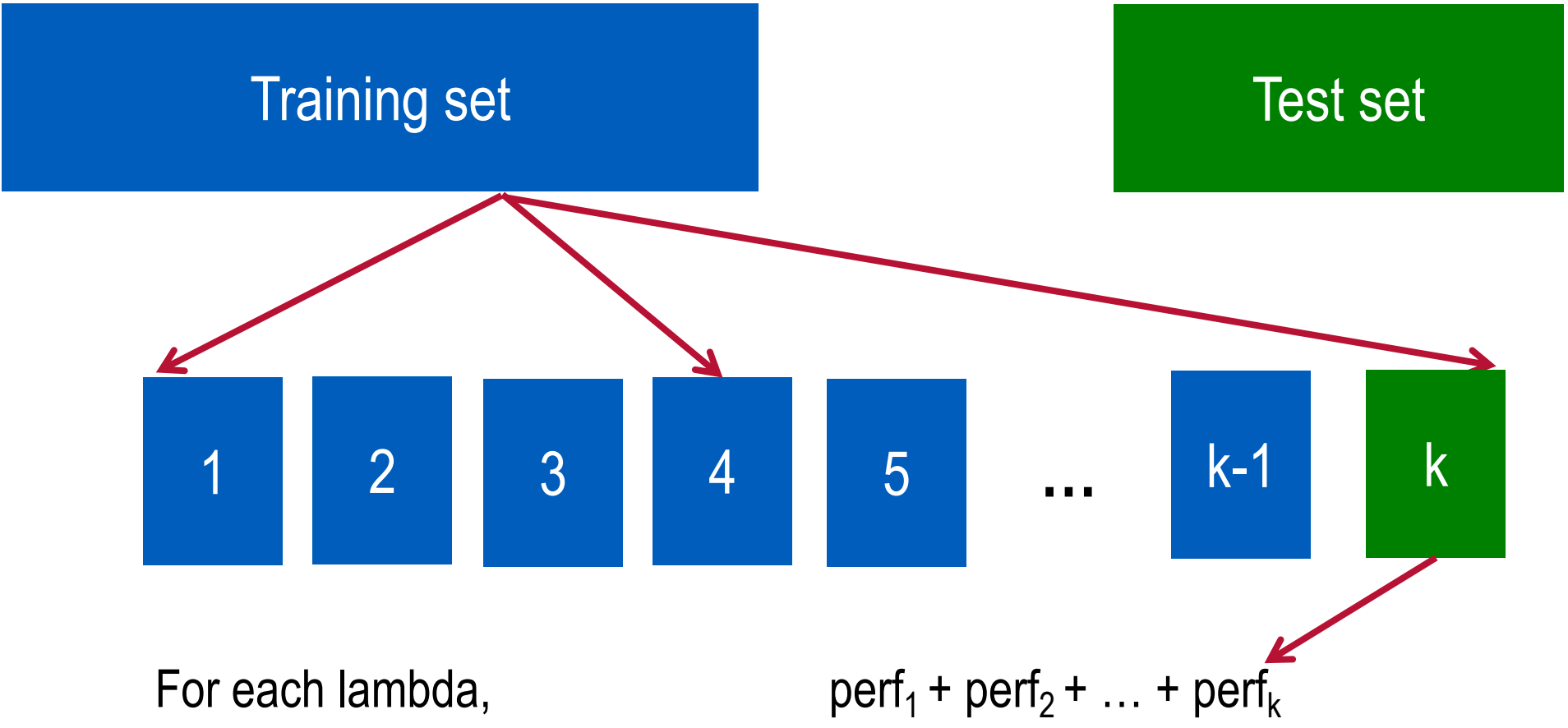
Cross-validation



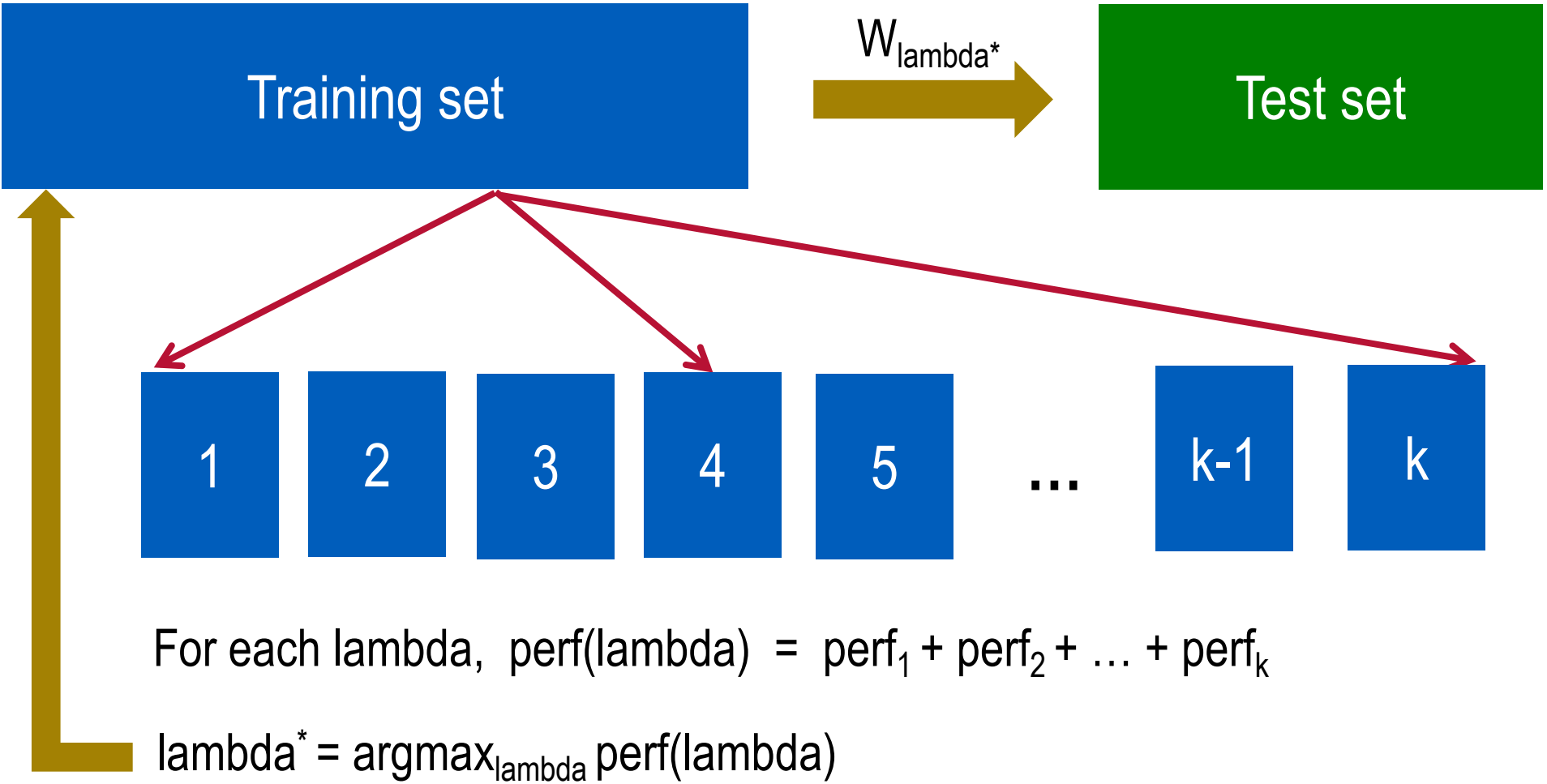
Cross-validation



Cross-validation



Cross-validation



For each lambda, $perf(lambda) = perf_1 + perf_2 + \ldots + perf_k$

$lambda^* = argmax_{lambda}\, perf(lambda)$

Questions?