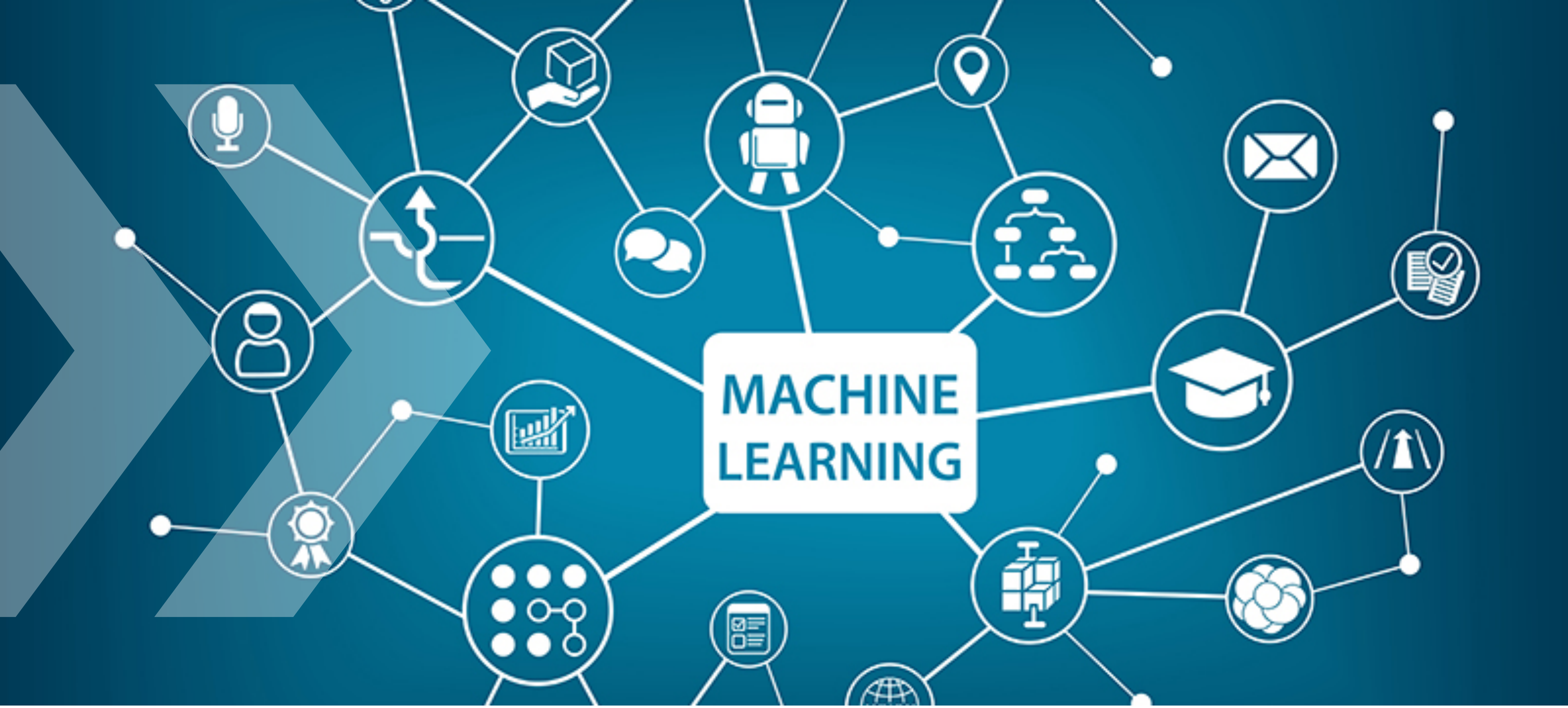


CS489/698: Intro to ML

Lecture 02: Linear Regression

UNIVERSITY OF WATERLOO

Outline

- Announcements
- Linear Regression
- Regularization
- Cross-validation

Outline

- Announcements
- Linear Regression
- Regularization
- Cross-validation

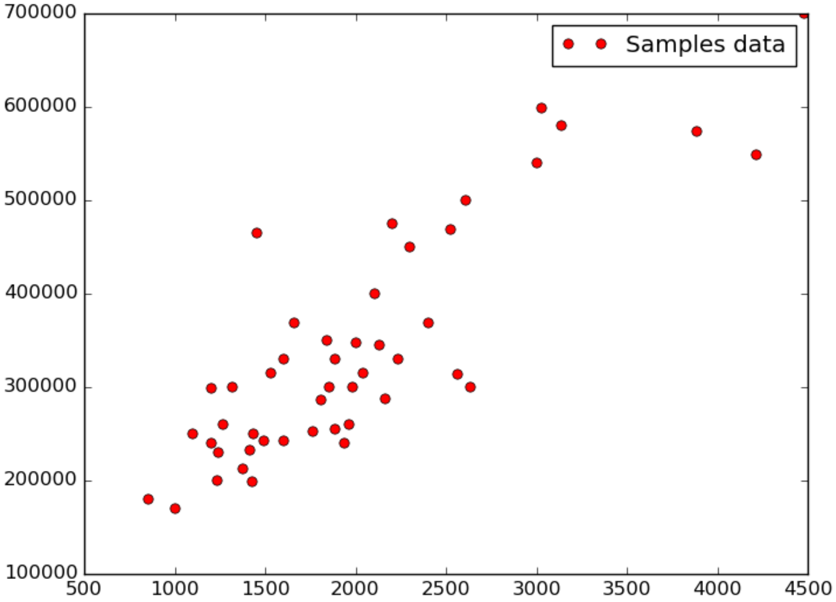
Announcements

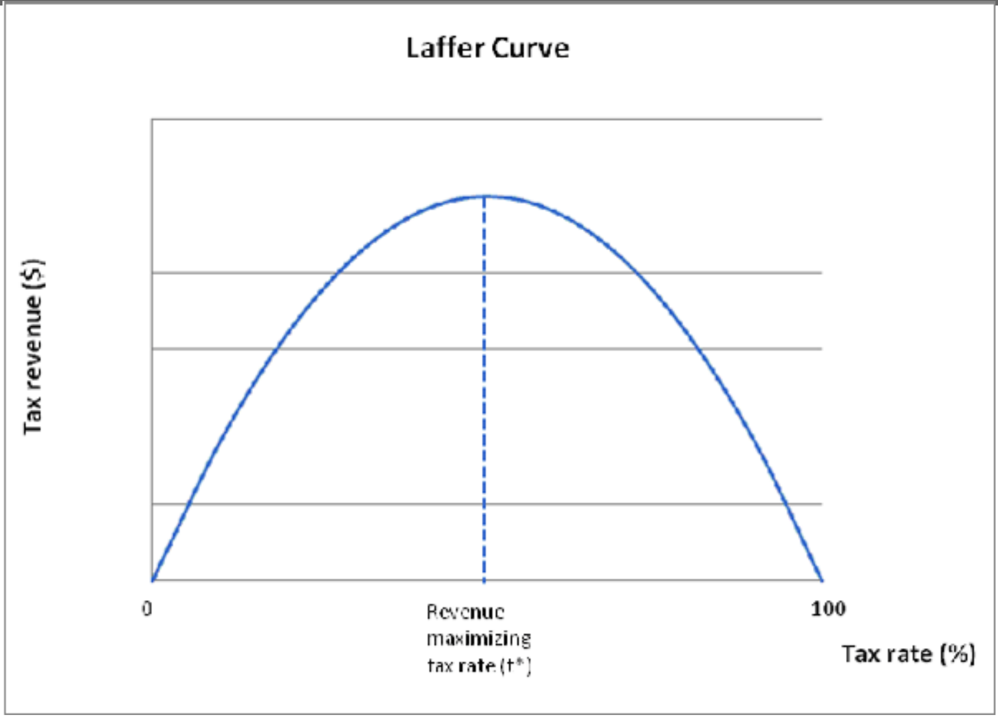
- Assignment 1 is out.
  - Due in two weeks
- TA office hour?
- Enrollment
  - CS698: permission numbers sent
  - CS489: ~10 seats available on Quest, ask CS advisors!

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

$$\min_{f:\mathcal{X}\to\mathcal{Y}} \underbrace{\mathbf{E}[L(f(X),Y)]}$$

$$L:\mathcal{Y}\times\mathcal{Y}\to\mathbf{R}_+$$

$$L(y,y)\equiv 0$$

- 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

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

- Many ways to estimate m(X)

- Simplest: Let’s assume it is linear (affine)!

UNIVERSITY OF WATERLOO

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: $$\frac{1}{n}\sum_{i=1}^{n} Z_i \to \mathbf{E}(Z)$$

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

Simplification, again

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

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

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

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

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

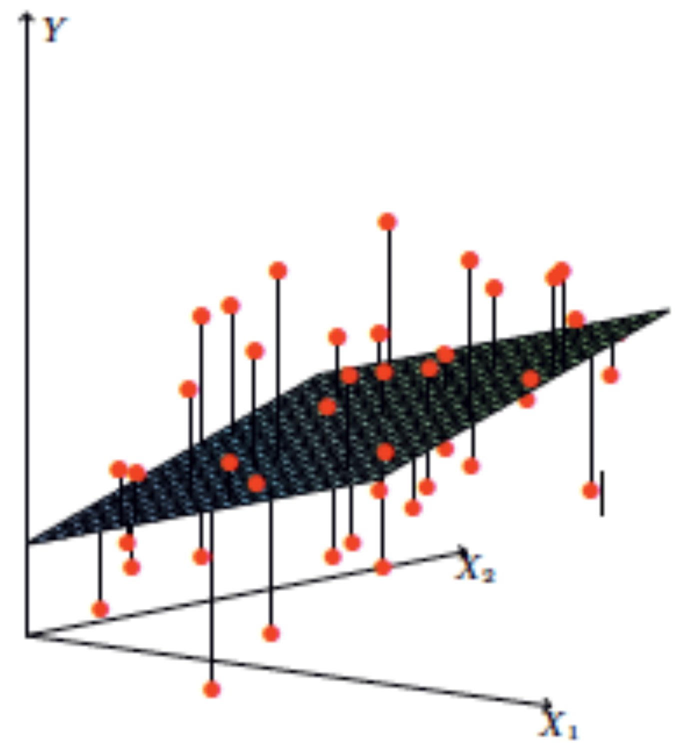
$$\min_{W \in \mathbf{R}^{(d+1)\times m}} \|XW - Y\|_{\mathsf{F}}^2$$

Finally

$$\min_{W \in \mathbf{R}^{(d+1)\times m}} \overbrace{\|XW - Y\|_{\mathsf{F}}^2}^{\text{Sum of square residuals}}$$

Hyperplane (again!) parameterized by W

True responses



UNIVERSITY OF WATERLOO

Why least squares?

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

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

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

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

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

Singular Value Decomposition $A = USV^T$ $A^+ = VS^{-1}U^T$

UNIVERSITY OF WATERLOO

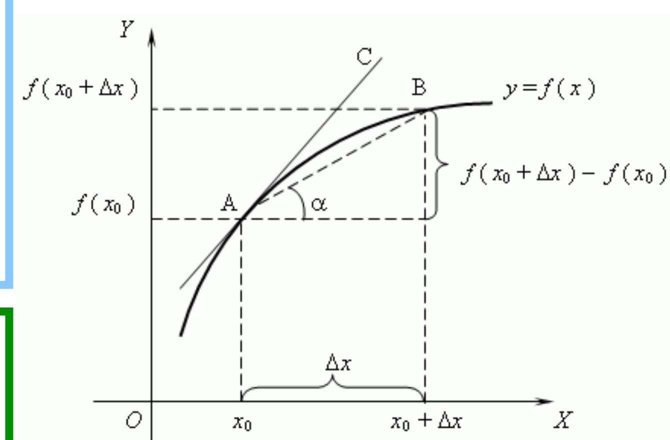
Optimization detour

$$\min_{x} \; f(x)$$

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

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

Solving least squares

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

$$\boxed{X^\top XW = X^\top Y}$$

Normal Equation

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

Prediction

- Once have W, can predict

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

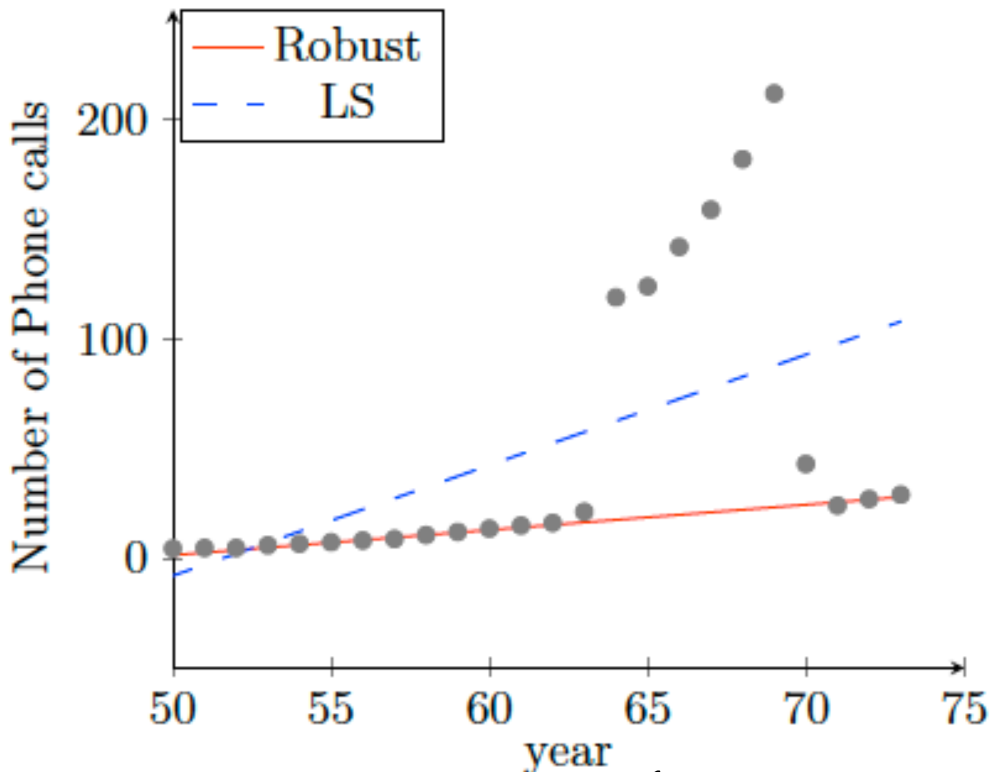
- How to evaluate?

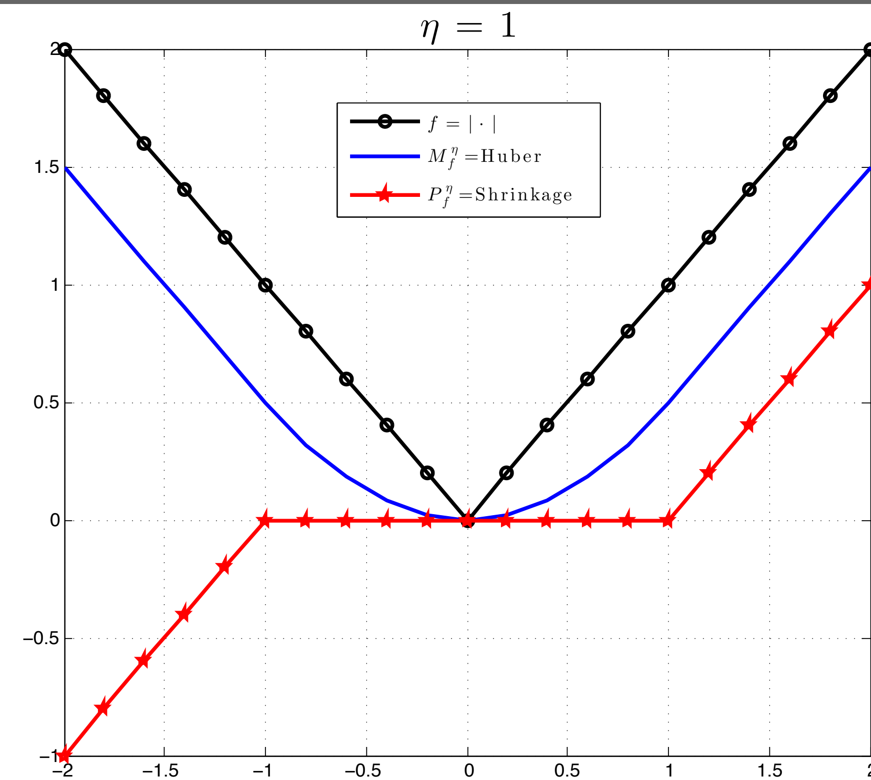
$$(Y_{\text{test}} - \hat{Y})^2$$

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

UNIVERSITY OF WATERLOO

Robustness





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

UNIVERSITY OF WATERLOO

Gauss vs. Laplace

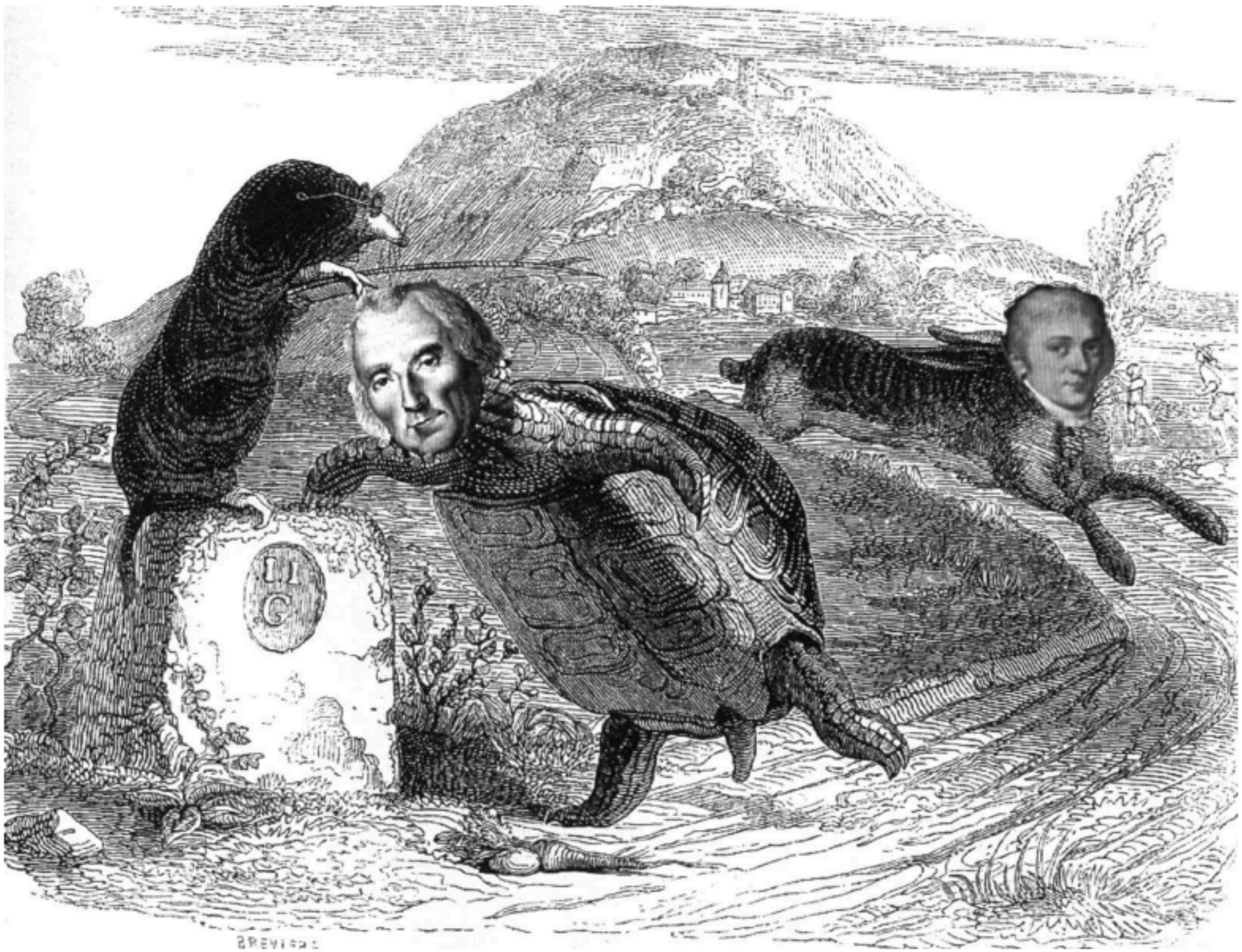


FIG. 1. *The Gaussian Hare and the Laplacian Tortoise: this picture is a slightly "retouched" version of a wood engraving by J. J. Grandville from "Fables de La Fontaine" (published in Paris, 1838). The portrait of Gauss is taken from an 1803 portrait by J. C. A. Schwartz. The portrait of Laplace appears 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

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

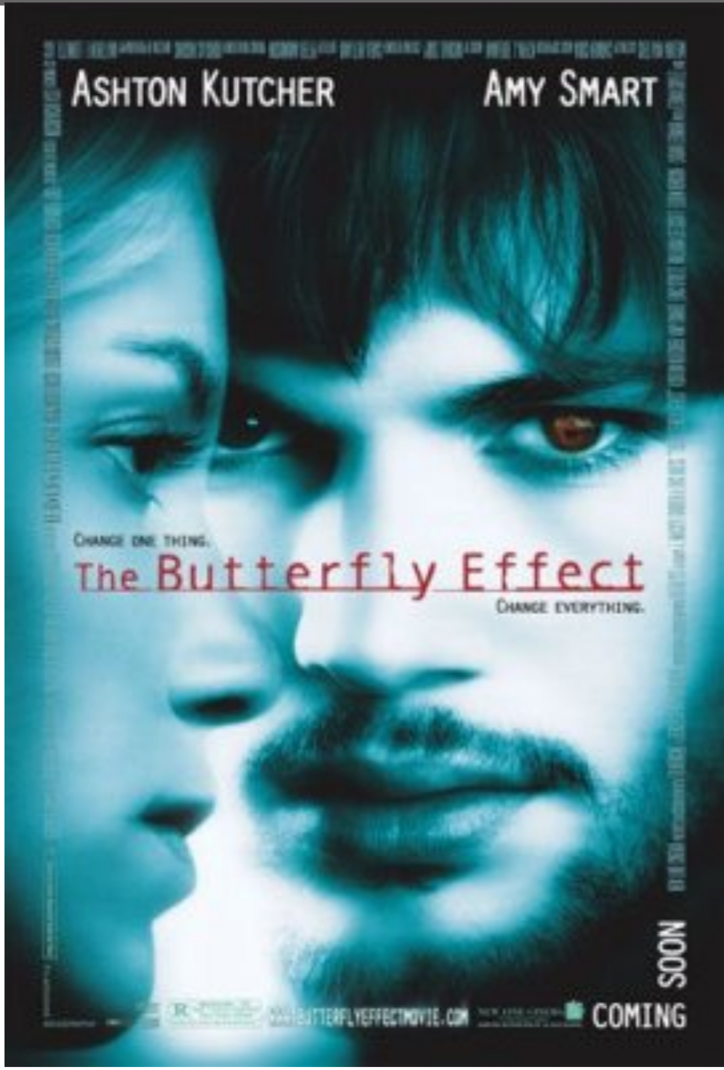
- Things are more interesting if we had regularization

Outline

- Announcements
- Linear Regression
- Regularization
- Cross-validation

Ill-posedness

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



UNIVERSITY OF WATERLOO

Tiknohov regularization (Hoerl and Kennard'70)

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

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

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

$$\Updownarrow$$

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

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

Sparsity

- Ridge regression weight is always dense
  - Computationally heavy
  - Interpretationally cumbersome

- Lasso (Tibshirani'96)

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

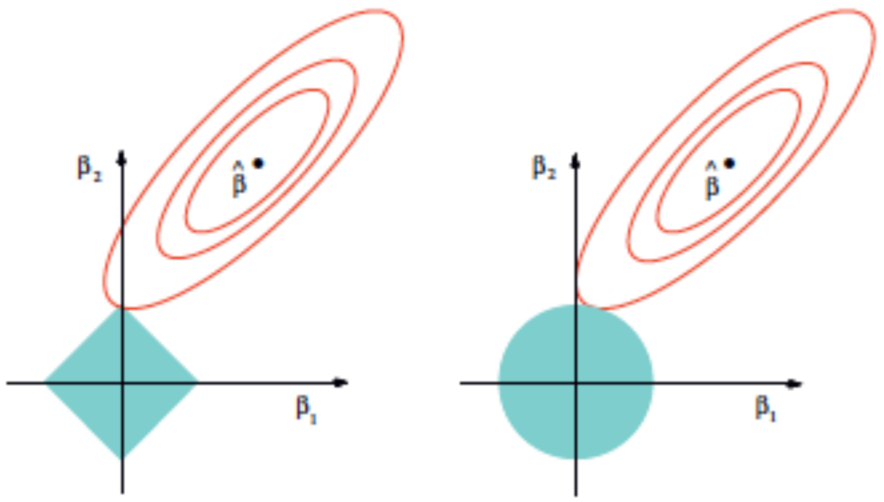


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \le t$ and $\beta_1^2 + \beta_2^2 \le t^2$, respectively, while the red ellipses are the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

UNIVERSITY OF WATERLOO

Regularization vs. Constraint

$$\min_{W} \; \ell(W) + \lambda \cdot r(W)$$

Computationally appealing

Always true

Mild conditions

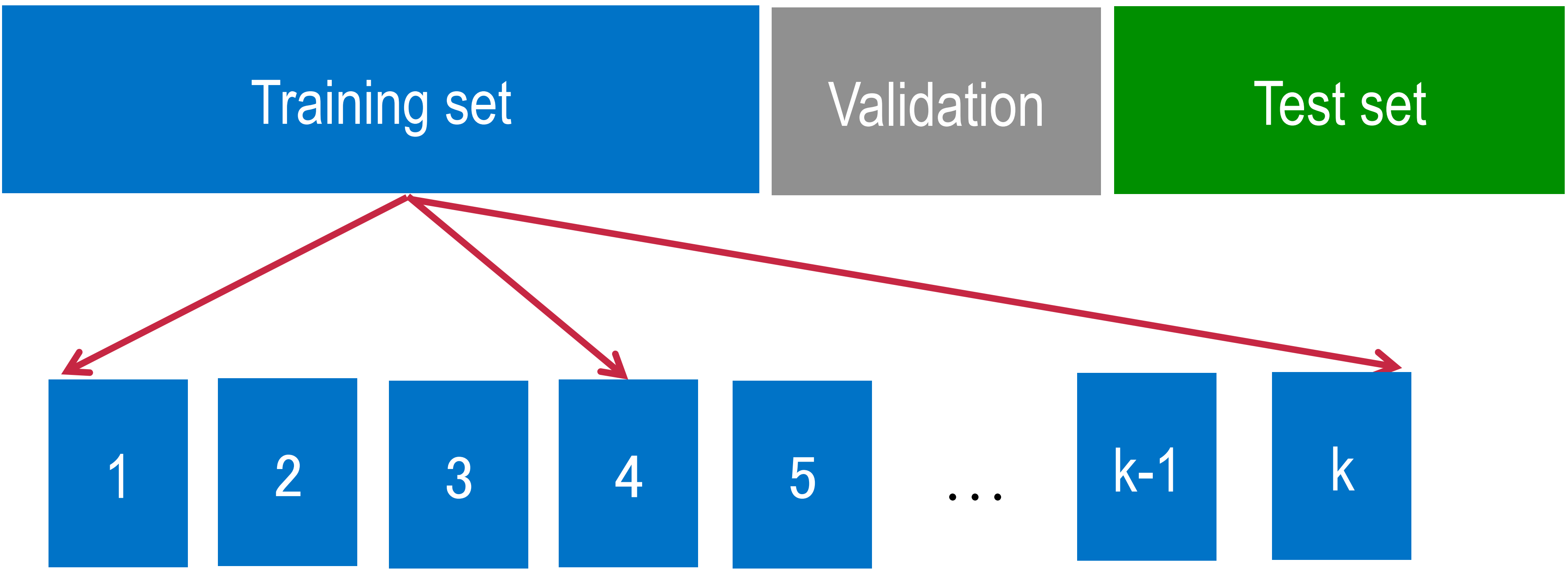
$$\min_{W : r(W) \leq C} \; \ell(W)$$

Theoretically appealing

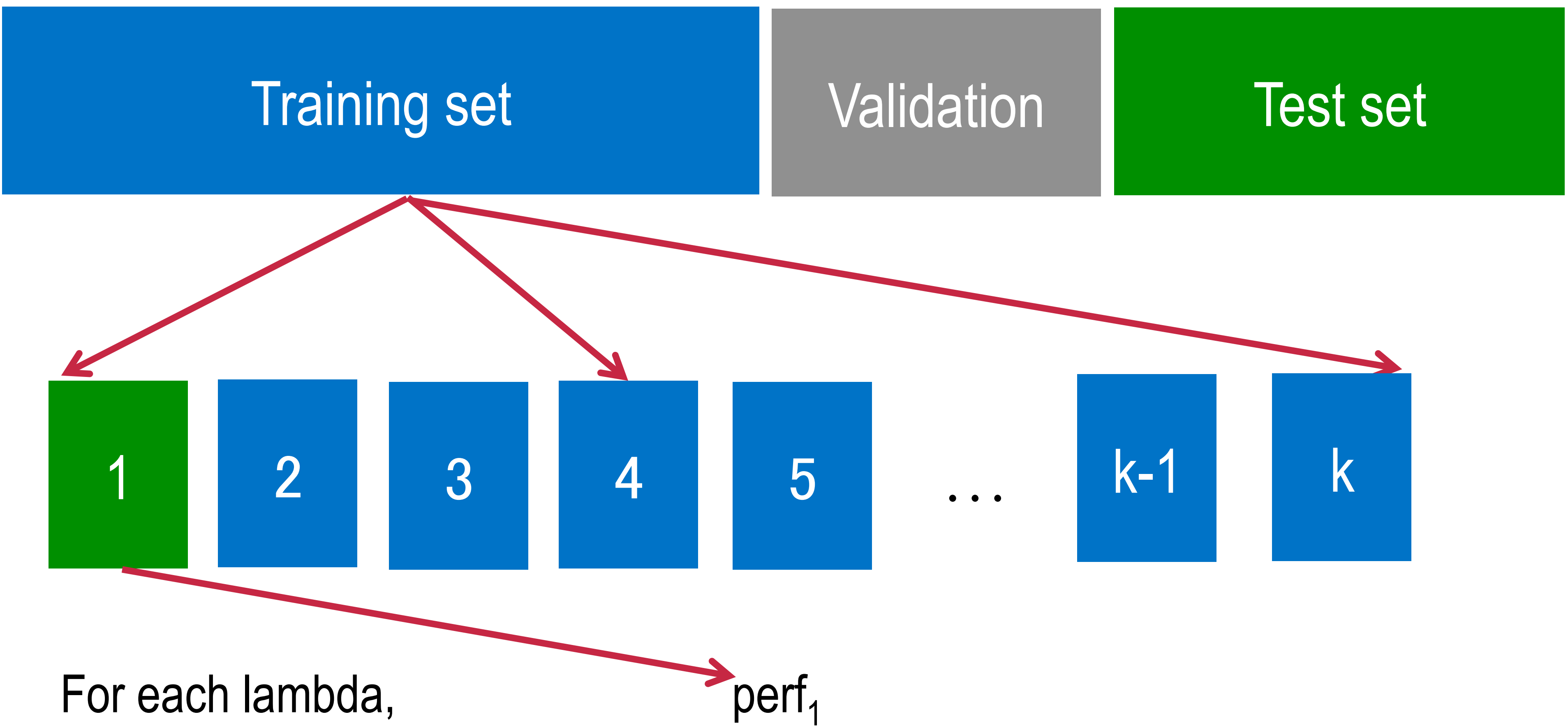
Outline

- Announcements
- Linear Regression
- Regularization
- Cross-validation

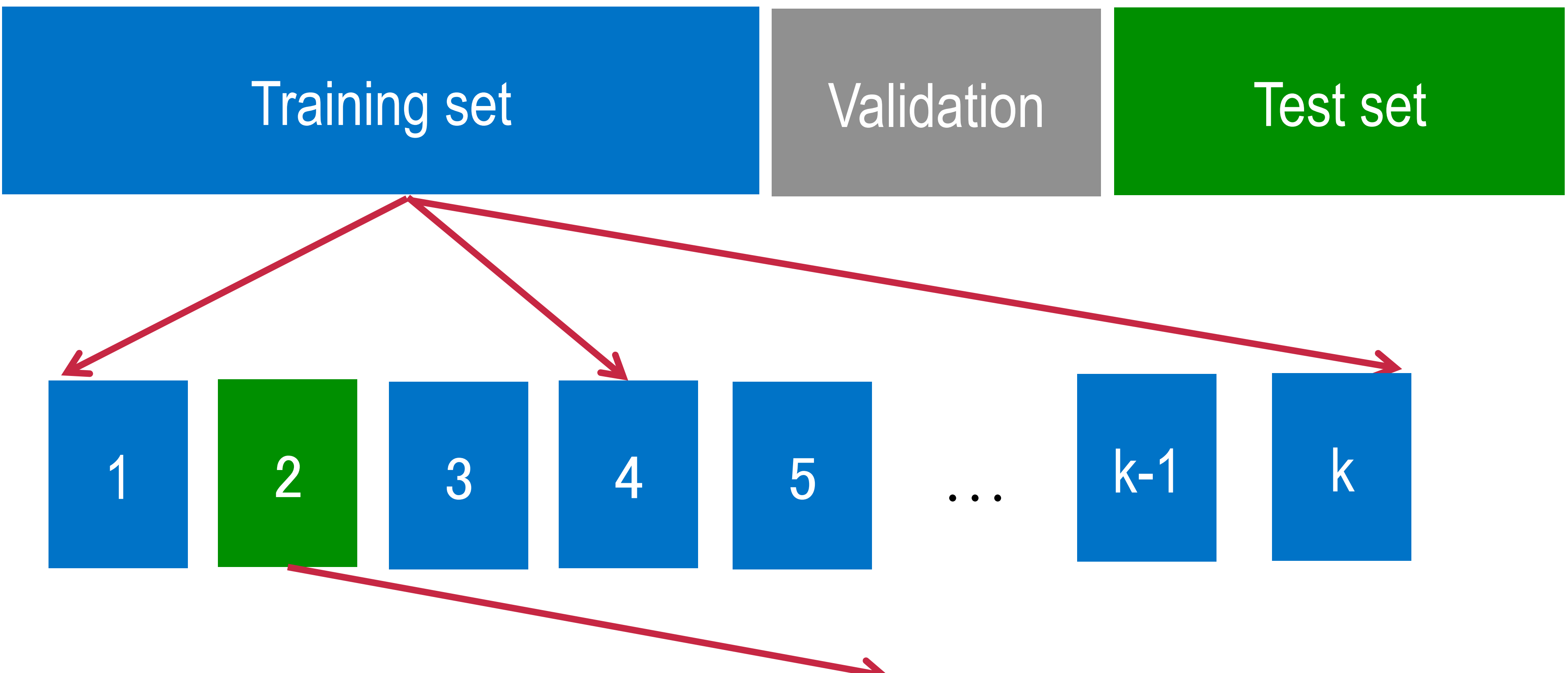
Cross-validation



Cross-validation

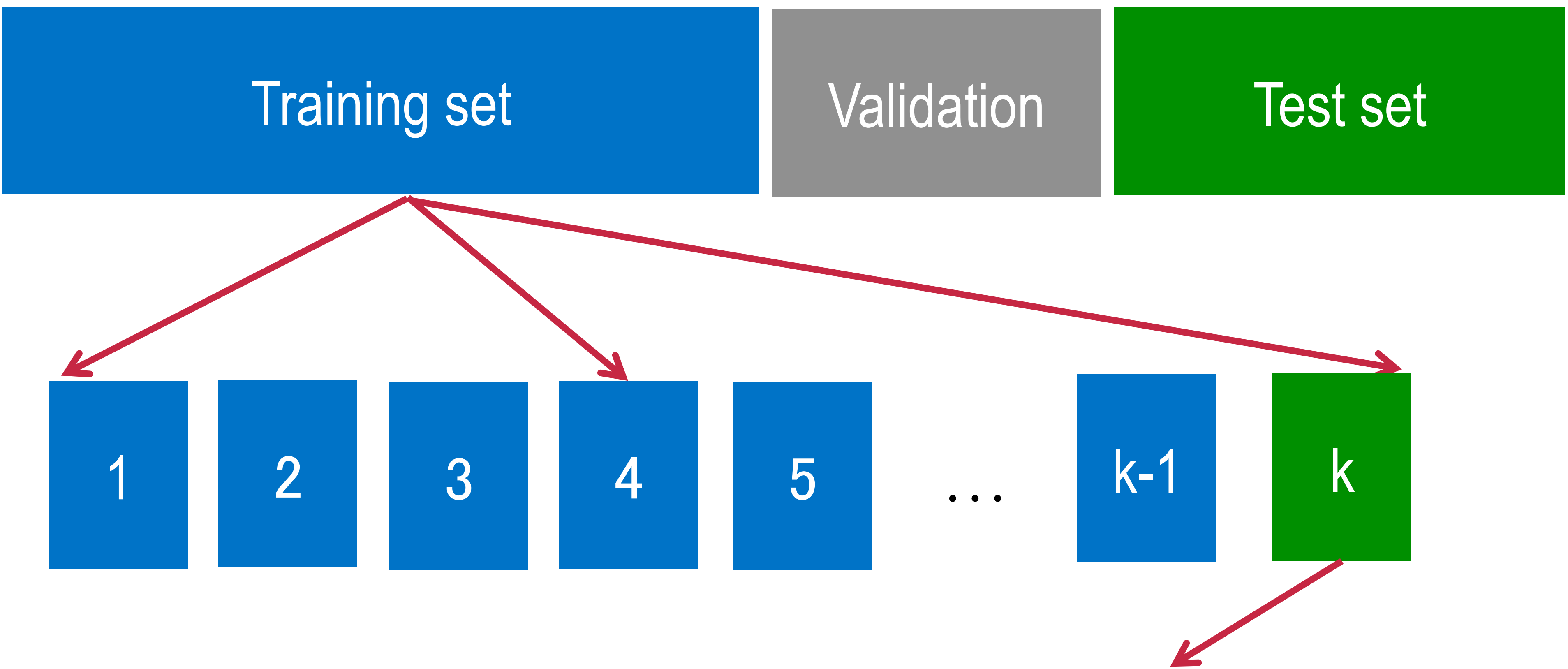


Cross-validation



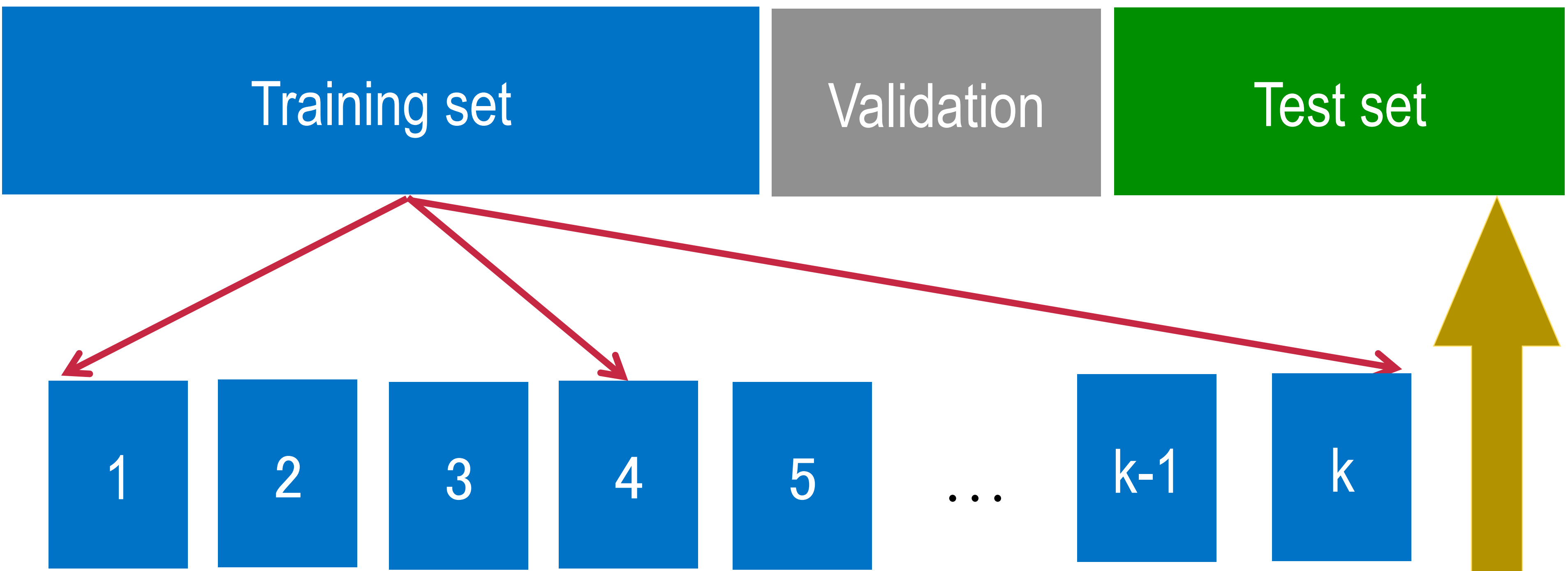
For each lambda, $\text{perf}_1 + \text{perf}_2$

Cross-validation



For each lambda, $\text{perf}_1 + \text{perf}_2 + \ldots + \text{perf}_k$

Cross-validation



For each lambda, perf(lambda) = $\text{perf}_1 + \text{perf}_2 + \ldots + \text{perf}_k$

Lambda* = $\text{argmax}_{\text{lambda}}$ perf(lambda)

Questions?