Introduction to Machine Learning

Benjamin Peherstorfer

Fall 2020

Slides adapted from David Sontag, Andrew Ng

Today

- Last time
  - Least-squares regression
  - Probabilistic view on least-squares regression
  - Logistic regression for classification

- Today
  - Generalized linear models (reading “Part III” of http://cs229.stanford.edu/notes/cs229-notes1.pdf)
  - Multi-class classification
  - Cross validation
  - Finish lab example

- Announcements
  - Homework 1 due on Wed, Sep 30 **before class**

Feedback

https://docs.google.com/forms/d/e/1FAIpQLSc_vxl9s4Y3OZjYWe2dikviHiUc3s8rmDHqJsxnZahlEwpJfw/viewform?usp=sf_link

<u>Recap:</u> MLE regression with Gaussian noise

- Let's revisit our regression problem
- Inputs and targets are related via the equation

$$\boldsymbol{y}^{(i)} = \boldsymbol{\theta}^T \boldsymbol{x}^{(i)} + \boldsymbol{\epsilon}^{(i)}, \qquad i = 1,\ldots,N$$

  - Error term $\boldsymbol{\epsilon}^{(i)}$ captures unmodeled effects and noise
  - Error terms are independent and identically distributed (iid)

$$\boldsymbol{\epsilon}^{(i)} \sim \mathcal{N}(0,\sigma^2)$$

- Just as we can have different loss functions, we model $\boldsymbol{y}^{(i)}$ with different distributions
- Gaussian noise implies

$$p(\boldsymbol{y}^{(i)} \,|\, \boldsymbol{x}^{(i)}; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\boldsymbol{y}^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)})^2}{2\sigma^2} \right)$$

Recap: Logistic regression

- Use principle of maximum likelihood to find $\boldsymbol{\theta}^*$
- Probabilistic assumption: Model $h_{\boldsymbol{\theta}}(x)$ gives the probability that y is 1, i.e.,

$$p(y = 1 \mid x; \boldsymbol{\theta}) = h_{\boldsymbol{\theta}}(x)$$

This also means (remember $h_{\boldsymbol{\theta}}(x) \in [0,1]$)

$$p(y = 0 \mid x; \boldsymbol{\theta}) = 1 - h_{\boldsymbol{\theta}}(x)$$

- Write more compactly as (because $y \in \{0,1\}$)

$$p(y \mid x; \boldsymbol{\theta}) = (h_{\boldsymbol{\theta}}(x))^y (1 - h_{\boldsymbol{\theta}}(x))^{(1-y)}$$

- (Remember that we modeled y with a Gaussian distribution in the earlier regression problem. We now model y with Bernoulli)

Generalized linear models

- Regression example $y \mid x; \theta \sim \mathcal{N}(\mu, \sigma^2)$
- Classification example $y \mid x; \theta \sim \mathrm{Bernoulli}(\phi)$
- Generalized linear models (GLM) generalize the concepts we have seen to the exponential family of distributions

$$p(y; \eta) = b(y) \exp(\eta^T T(y) - a(\eta))$$

- Natural parameter η
- Sufficient statistics $T(y)$
- Log partition function $a(\eta)$
- A fixed choice of b, T, a defines a family with parameter η

Examples of distributions in exponential family

- Bernoulli distribution is an exponential family distribution

$$p(y = 1; \phi) = \phi$$
$$p(y = 0; \phi) = 1 - \phi$$

Transform so that $p(y; \eta) = b(y)\exp(\eta^T T(y) - a(\eta))$

board

Examples of distributions in exponential family (cont'd)

$$
\begin{aligned}
p(y;\phi) &= \phi^y(1-\phi)^{1-y} \\
&= \exp(y\log\phi + (1-y)\log(1-\phi)) \\
&= \exp\left(\left(\log\left(\frac{\phi}{1-\phi}\right)\right)y + \log(1-\phi)\right)
\end{aligned}
$$

Corresponds to $p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$ with

$$
\begin{aligned}
T(y) &= y \\
a(\eta) &= -\log(1-\phi) \\
&= \log(1+e^\eta) \\
b(y) &= 1
\end{aligned}
\qquad
\eta = \log\left(\frac{\phi}{1-\phi}\right)
$$

Gaussian distribution is in the exponential family as well

Constructing GLMs

- Would like to predict value y as function of x
- Assumptions
  1. $y \mid x; \theta \sim \mathrm{ExpFamily}(\eta)$, i.e., given x, θ the distribution of y follows an exponential family with parameter η
  2. Given x the goal is to predict expected value $T(y)$ (sufficient statistics).
     - In most cases we consider $T(y) = y$ and thus
     $$h_\theta(x) = \mathbb{E}[y \mid x]$$
  3. Natural parameter η and inputs x are related linearly $\eta = \theta^T x$

Example: logistic regression as GLM

board

Example: logistic regression as GLM

- Select Bernoulli distribution for $y \,|\, x, \theta \sim \text{Bernoulli}(\phi)$

- Our GLM then gives $h_\theta(x) = \mathbb{E}[T(y) \,|\, x; \theta] = \mathbb{E}[y \,|\, x; \theta]$ because $T(y) = y$ for Bernoulli distribution

- Obtain $\mathbb{E}[y \,|\, x; \theta] = 0p(y = 0 \,|\, x; \theta) + 1p(y = 1 \,|\, x; \theta) = p(y = 1 \,|\, x; \theta) = \phi$

- Have $\phi = 1/(1 + \mathrm{e}^{-\eta})$ and $\eta = \theta^T x$ and thus
$$h_\theta(x) = \mathbb{E}[y \,|\, x; \theta] = \phi = \frac{1}{1 + \mathrm{e}^{-\theta^T x}}$$

Softmax regression

- Construct a GLM for multi-class response variable

$$y \in \mathcal{Y} = \{1,2,\ldots,k\}$$

- Categorical distribution extends Bernoulli distribution to k outcomes

$$p(y = i \,|\, x; \theta) = \phi_i$$

with

$$\sum_{i=1}^{k} \phi_i = 1\,, \qquad \phi_i > 0$$

Derivation of GLM for multi-class classification

Softmax function

- Softmax function

$$\sigma(z) = \frac{1}{\sum_{j=1}^{K} \mathrm{e}^{z_j}} \begin{bmatrix} \mathrm{e}^{z_1} \\ \vdots \\ \mathrm{e}^{z_K} \end{bmatrix}$$

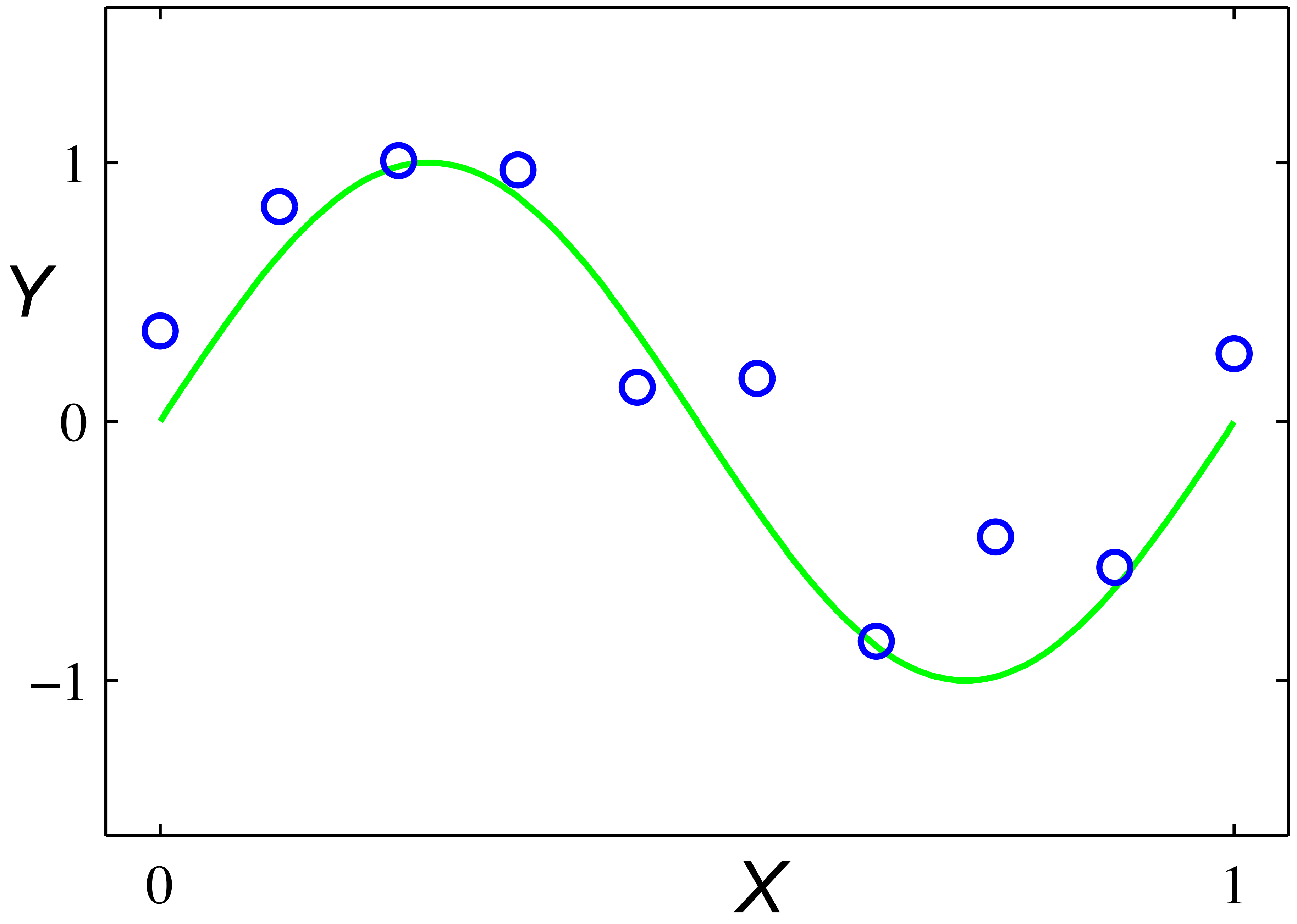
- Components of $\sigma(z)$ sum to 1 ⇒ interpret as "probability"

- Smooth approximation of $\arg\max$
  - $\arg\max(1,2,3) = [0,0,1]$
  - $\sigma([1,2,3]) \approx [0.09,0.24,0.66]$

Model selection and cross validation

Second example: Regression

Dataset: 10 (X,Y) points generated from a sin function, with noise



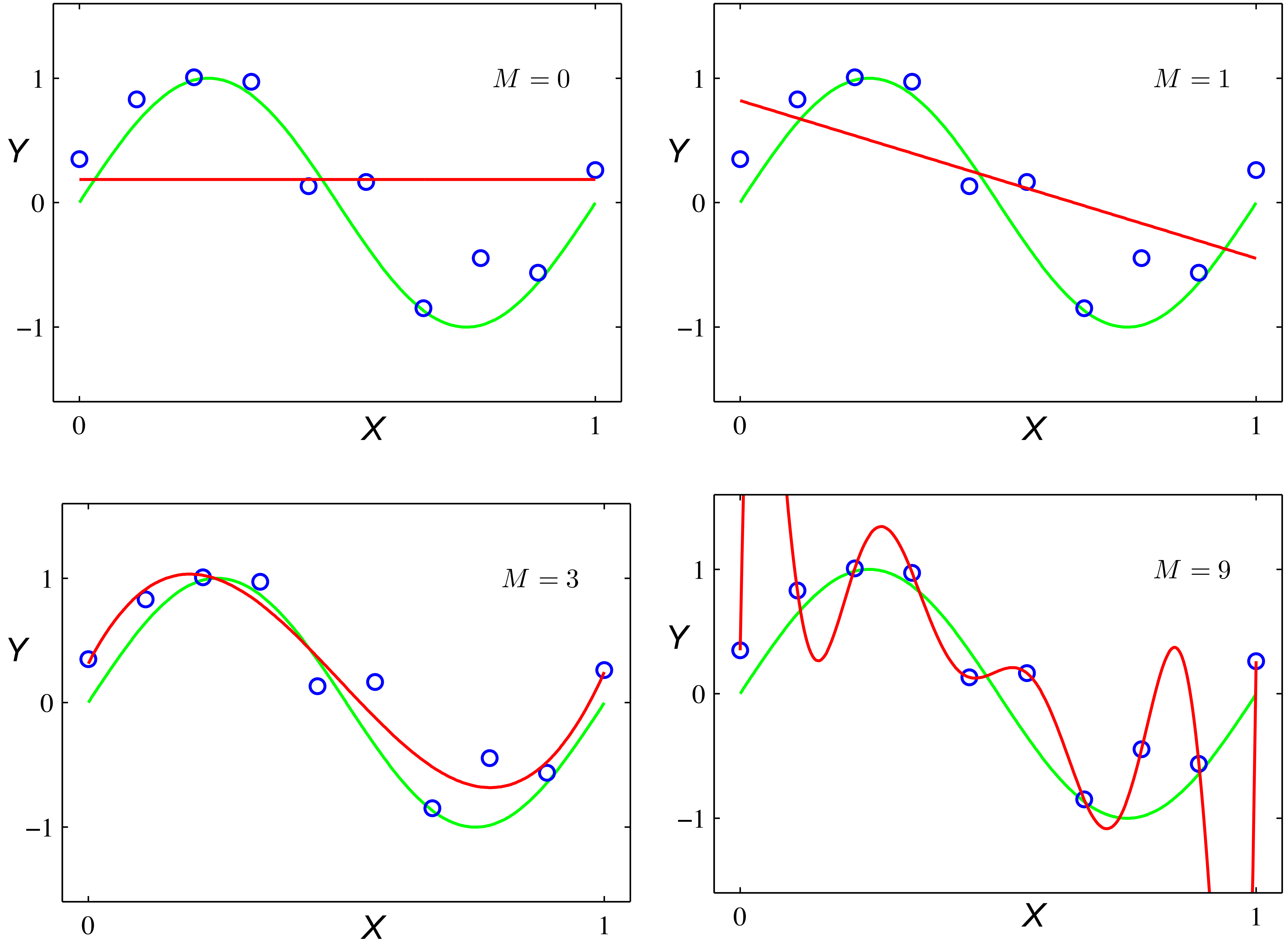
- Regression:
  - $f : X \rightarrow Y$
  - $X = \Re$
  - $Y = \Re$

[Bishop]

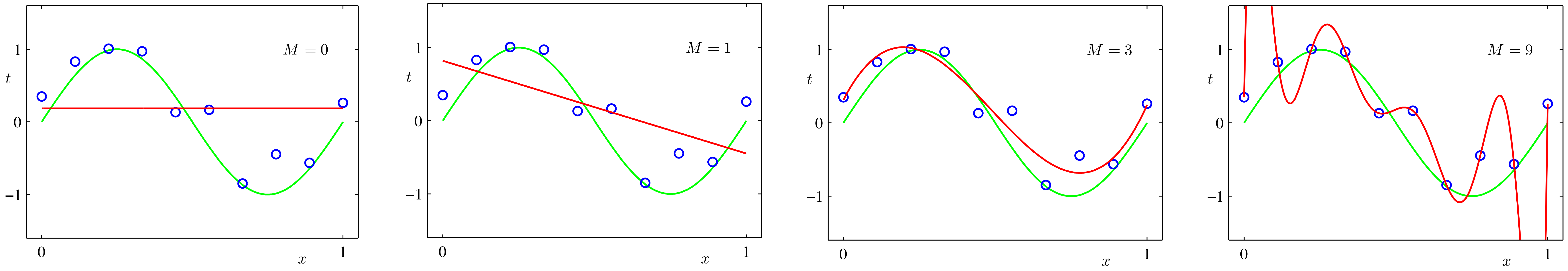
Degree-M Polynomials

How about letting *f* be a degree M polynomial?

- Which one is **best**?



Hypo. Space: Degree-N Polynomials



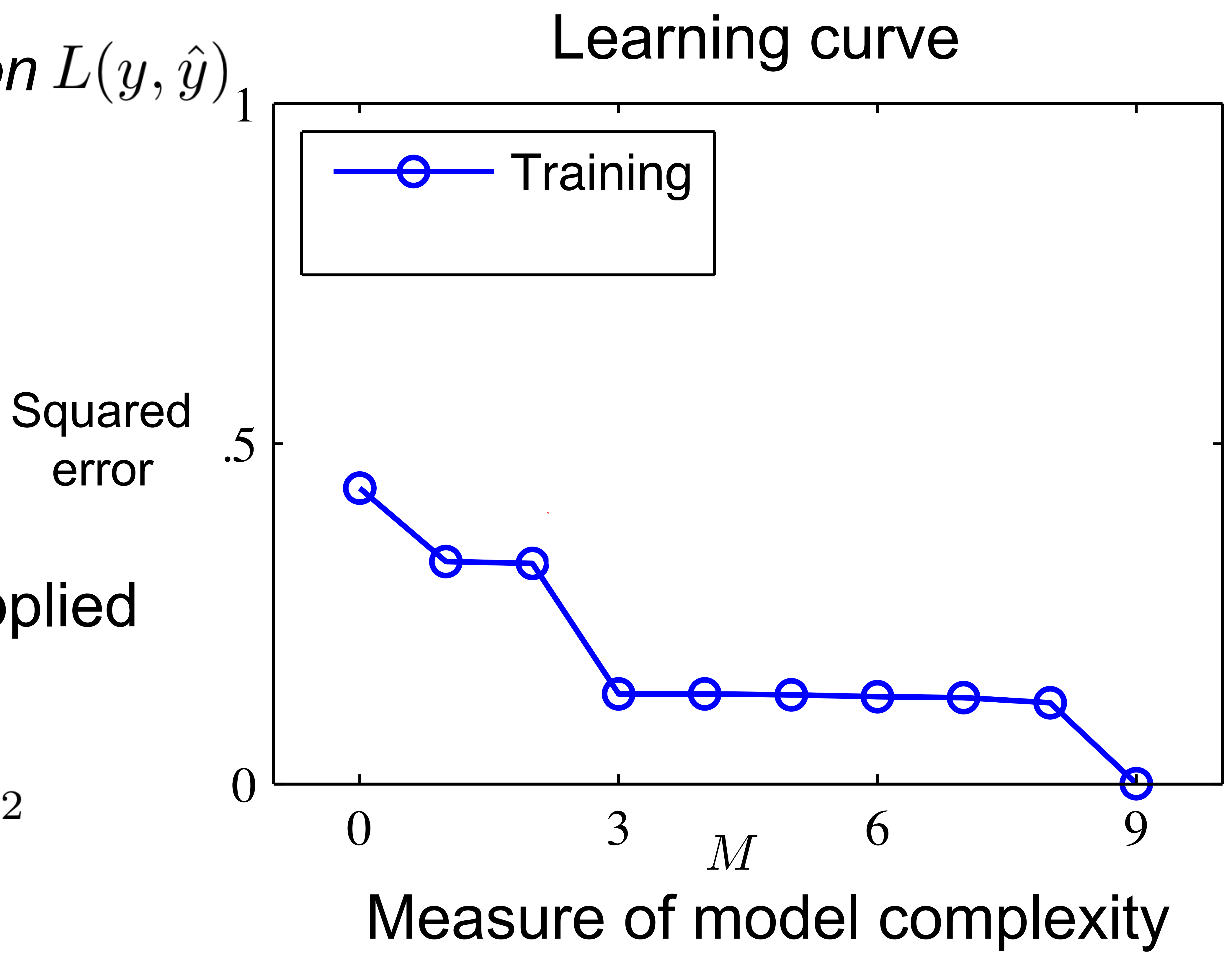
We measure error using a *loss function* $L(y, \hat{y})$

For regression, a common choice is squared loss:

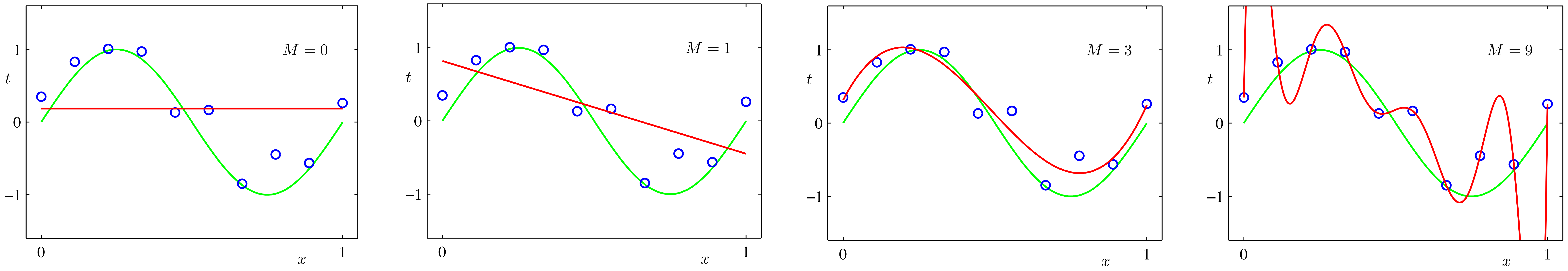
$$L(y_i, f(x_i)) = (y_i - f(x_i))^2$$

The *empirical loss* of the function *f* applied to the training data is then:

$$\frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2$$



Hypo. Space: Degree-N Polynomials



We measure error using a *loss function* $L(y, \hat{y})$

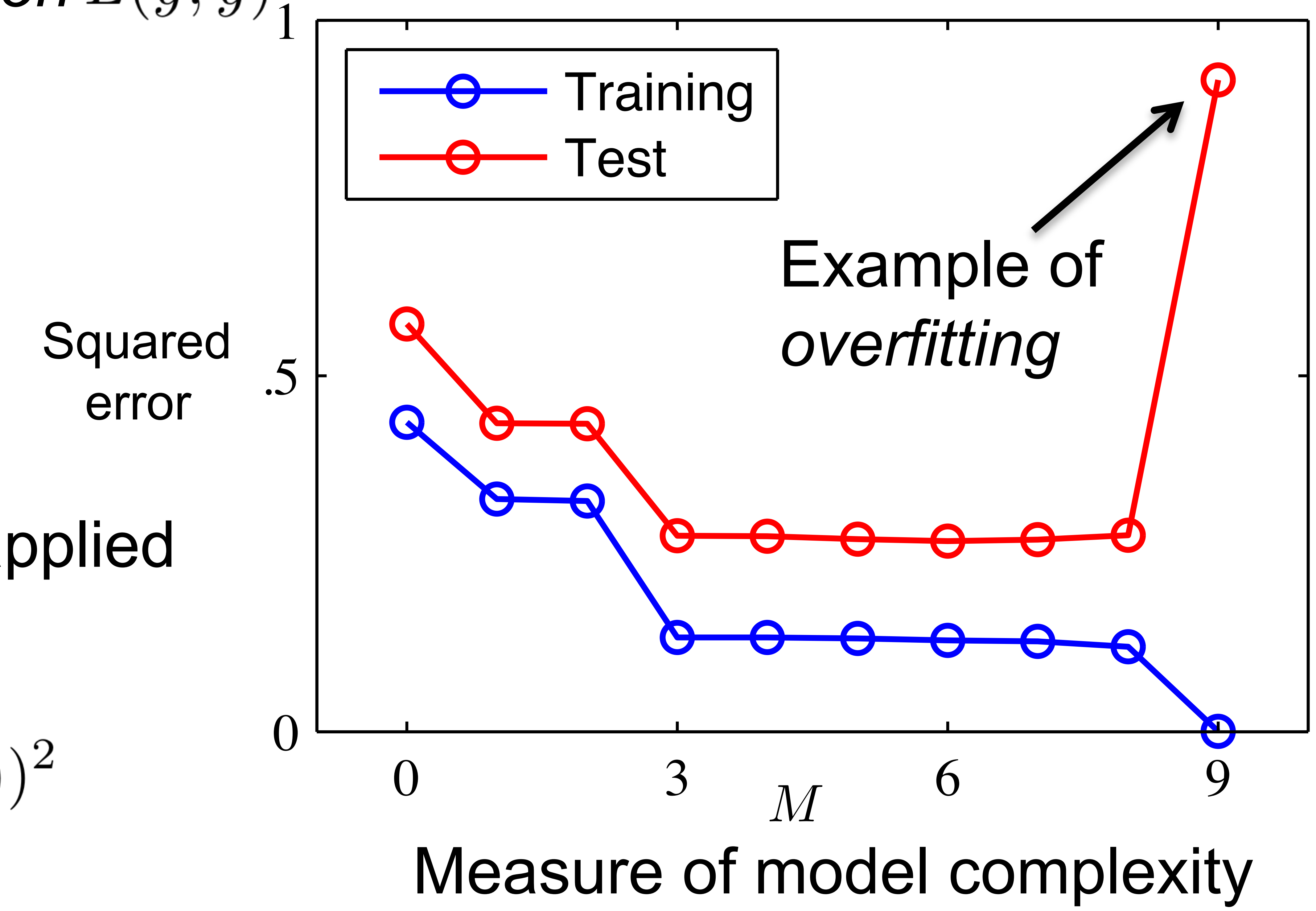
For regression, a common choice is squared loss:

$$L(y_i, f(x_i)) = (y_i - f(x_i))^2$$

The *empirical loss* of the function *f* applied to the training data is then:

$$\frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2$$

Learning curve



Train, validation and test data sets

- Train our models (find parameter $\boldsymbol{\theta}^*$ of hypothesis) on training set
$$\mathscr{D} = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})\}$$

- Cannot assess quality of trained model on training set because might have overfitted model (—> lecture 1)

- Instead, need a test set
$$\mathscr{D}_T = \{(x_{test}^{(1}, y_{test}^{(1)}), \ldots, (x_{test}^{(M)}, y_{test}^{(M)})\}$$
to estimate, e.g., misclassification error, least-squares costs

- Test set **must** be withheld throughout the whole learning process

Validation set

- Validation set
$$\mathcal{D}_V = \{(x_{val}^{(1)}, y_{val}^{(1)}), \ldots, (x_{val}^{(M')}, y_{val}^{(M')})\}$$
is useful if we have to choose between multiple "best" hypothesis $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_k^*$

- Example 1: Found hypothesis $h_{\boldsymbol{\theta}_1^*}$ with logistic regression and $h_{\boldsymbol{\theta}_2^*}$ with least-squares regression
- Example 2: Had different assumptions on $y \,|\, x; \theta \sim \mathcal{P}$ in GLM
- Example 3: Typically find $\boldsymbol{\theta}^*$ via iterative method and take the $\boldsymbol{\theta}^*$ at final iteration. Use early stopping by considering the history $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ of parameters over all iterations and and select $\boldsymbol{\theta}^*$ via validation set

- Critical to have validation set in these examples

Cross validation

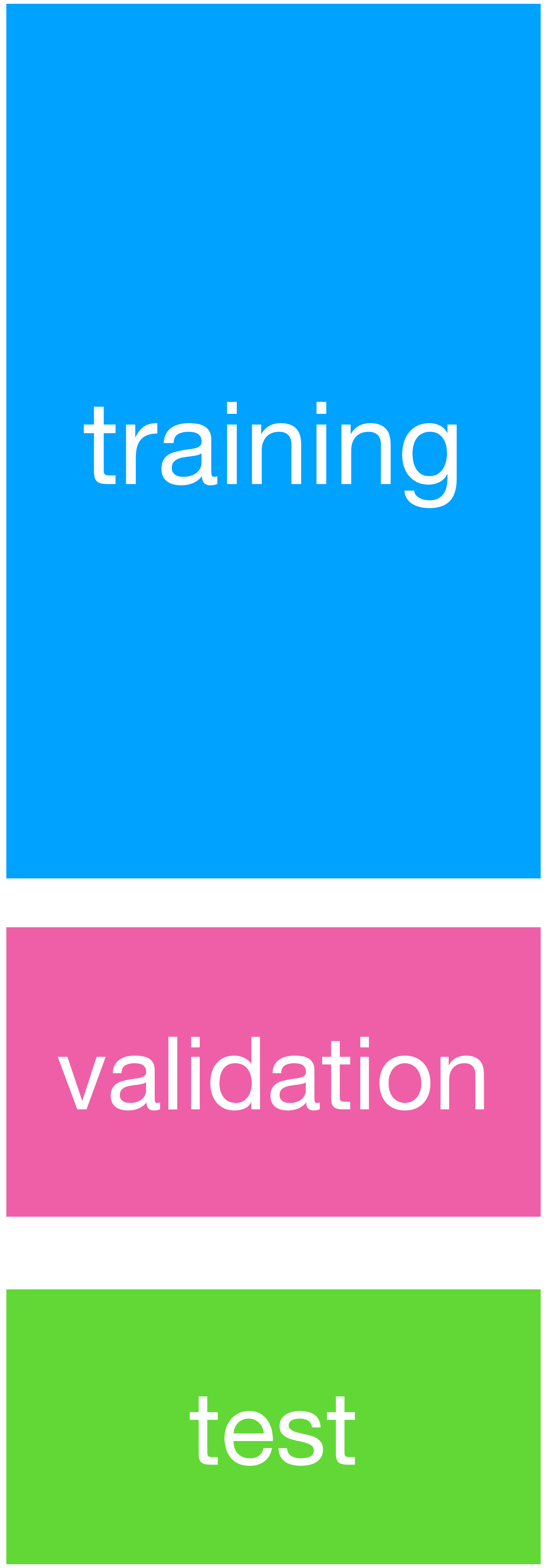
- If not enough data to split into train, val, test
- Split of training data $\mathscr{D}$ into k disjoint subsets

$$\mathscr{D} = \mathscr{D}_1 \cup \ldots \cup \mathscr{D}_k$$

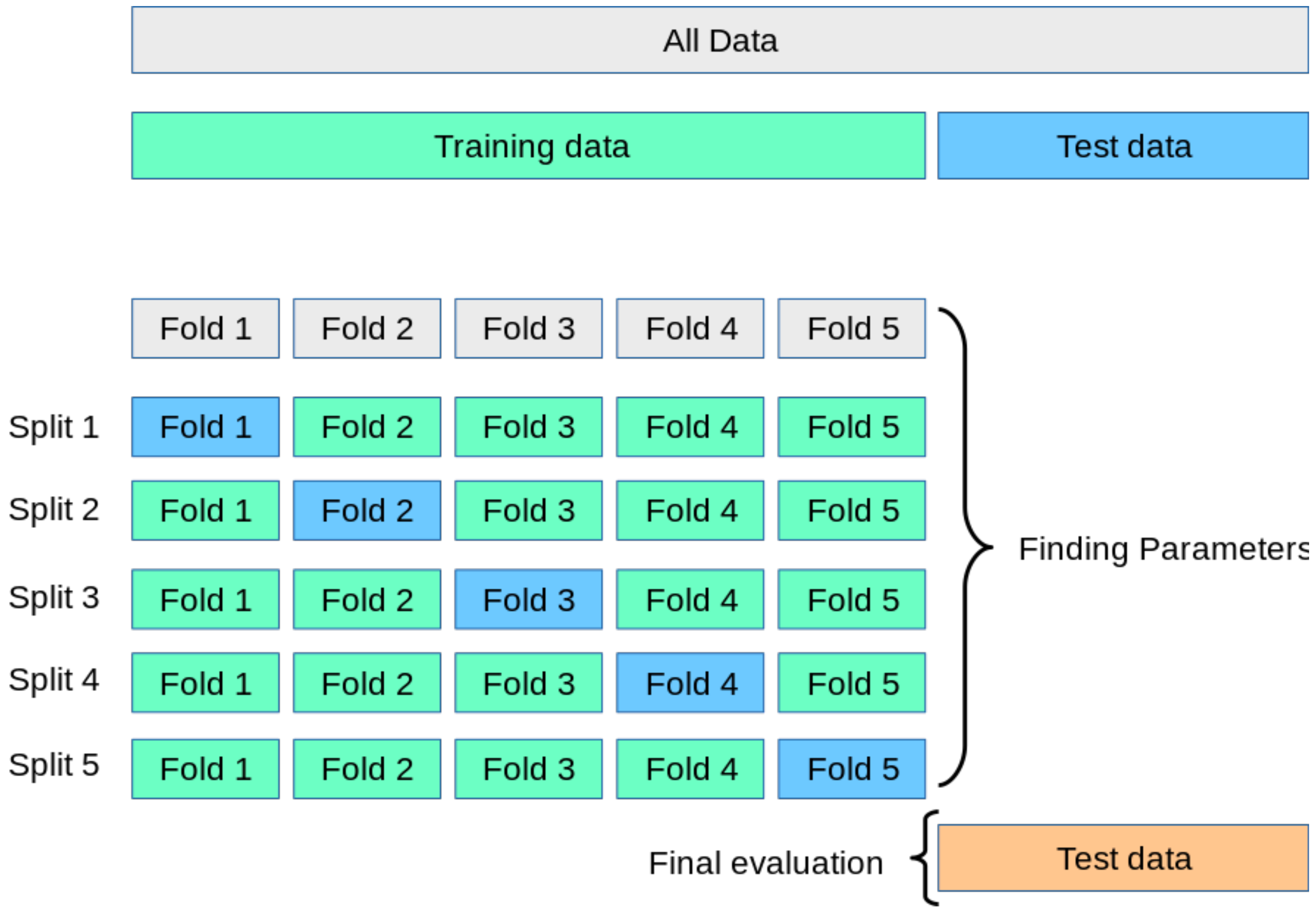
- Ensure that statistics of all partitions are about the same, e.g., label proportion

- Train k hypothesis on each $\mathscr{D} \backslash \mathscr{D}_i$ and validate on $\mathscr{D}_i$

- Take the “best” hypothesis and test on test data



Cross validation (cont'd)



[scikit-learn documentation]

Finish up lab from last time