Large Scale Optimization for Machine Learning

Meisam Razaviyayn

Lecture 11

razaviya@usc.edu

Agenda

- Some optimization problems in ML:
  - SVM, regression, logistic regression, deep learning
- Empirical risk minimization framework
  - Generalization error, tradeoffs, cross validation
- Exploiting structure

Linear Regression

Area	Crime Rate	Age	RAD	PTRATIO	Bedrooms	…	Price (K)
600	1.05	12	2.4	10.1	1	…	500
1000	2.34	10	2.5	20.1	1	…	800
1200	1.45	3	3.1	9.7	3	…	1500
1500	1.56	30	1.7	7.2	2	…	1200
…	…	…	…	…	…	…	…
2700	1.01	20	0.9	4.3	4	…	5000

$\mathbf{x}_1$ y_1

$\mathbf{x}_n$ y_n

$\mathbf{x}_i \in \mathbb{R}^d$

$$\mathbf{y} = \mathbf{w}^T\mathbf{x} + \mathbf{z} \text{ with } \mathbf{z} \sim \mathcal{N}(0, \sigma^2)$$

Maximum Likelihood Estimation

$$\min_{\mathbf{w}} \quad \frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i - y_i)^2$$

$$\text{s.t.} \quad \mathbf{w} \in \mathbb{R}^d$$

What is the loss function here?

Logistic Regression

Radius	Texture	Area	Compactness	Symmetry	…	Rec/non-Rec
1.1	2.3	3.5	2.4	1.4	…	1
0.7	1.2	2.5	1.4	3.2	…	0
1.7	2.4	1.5	3.3	1.3	…	1
…	…	…	…	…	…	…
0.2	3.4	0.7	4.3	2.0	…	1

$\mathbf{x}_1$ y_1 $\mathbf{x}_i \in \mathbb{R}^d$ $\mathbf{x}_n$ y_n

Model: $\log\left(\frac{\mathbb{P}(y=1 \mid \mathbf{w},\mathbf{x})}{\mathbb{P}(y=0 \mid \mathbf{w},\mathbf{x})}\right) = \mathbf{w}^T\mathbf{x}$

Maximum likelihood estimator

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \log\left(1+\exp(\mathbf{w}^T\mathbf{x}_i)\right) - \sum_{\{i:y_i=1\}} \mathbf{w}^T\mathbf{x}_i$$

$$\text{s.t.} \quad \mathbf{w} \in \mathbb{R}^d$$

Logistic Regression

Model: $\log\left(\frac{\mathbb{P}(y=1 \mid \mathbf{w},\mathbf{x})}{\mathbb{P}(y=0 \mid \mathbf{w},\mathbf{x})}\right) = \mathbf{w}^T\mathbf{x}$

Maximum likelihood estimator

$$\min_{\mathbf{w}} \quad \sum_{i=1}^{n} \log\left(1+\exp(\mathbf{w}^T\mathbf{x}_i)\right) - \sum_{\{i:y_i=1\}} \mathbf{w}^T\mathbf{x}_i$$

$$\text{s.t.} \quad \mathbf{w} \in \mathbb{R}^d$$

What is the loss function here?

$$\min_{\mathbf{w}} \quad \sum_{i=1}^{n} \left(\log\left(1+\exp(\mathbf{w}^T\mathbf{x}_i)\right) - y_i\mathbf{w}^T\mathbf{x}_i\right)$$

Support Vector Machines

- Binary classification task

Blood Pressure	Age	Sex	LDL	Glucose	BMI	…	Diabetic
85	29	0	100	75.1	24.2	…	-1
95	50	1	115	90.2	19.2	…	-1
…	…	…	…	…	…	…	…
123	42	1	150	110	25.2	…	1

Samples

Features

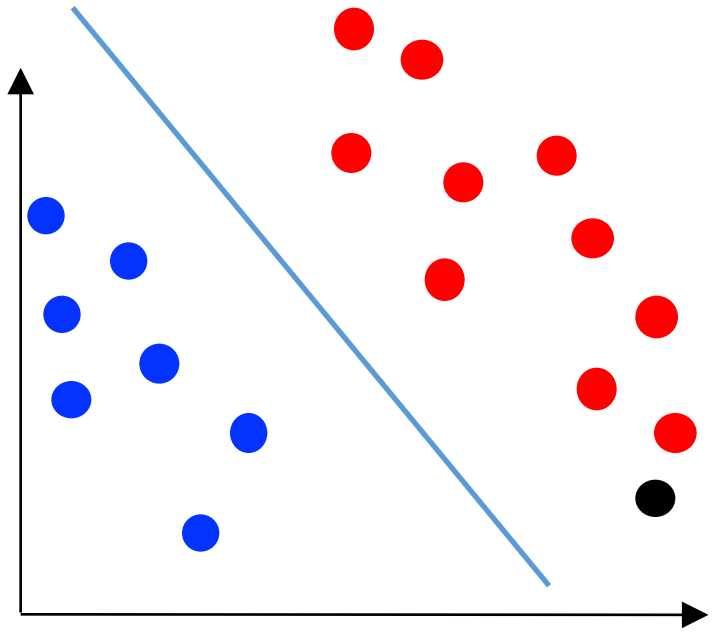
Label

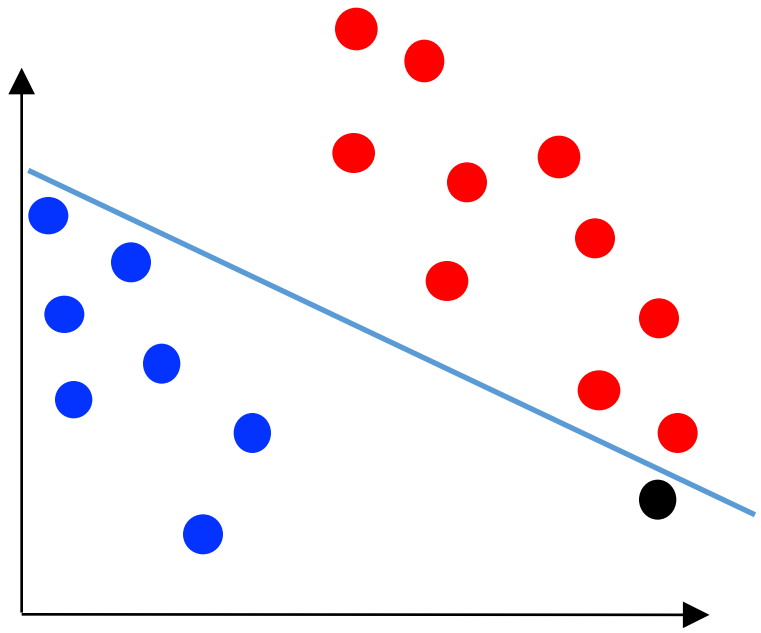
New patient:

119	37	0	120	100	19.3	…	**????**

Task: find a mapping $h : \text{Features} \mapsto \text{Labels}$

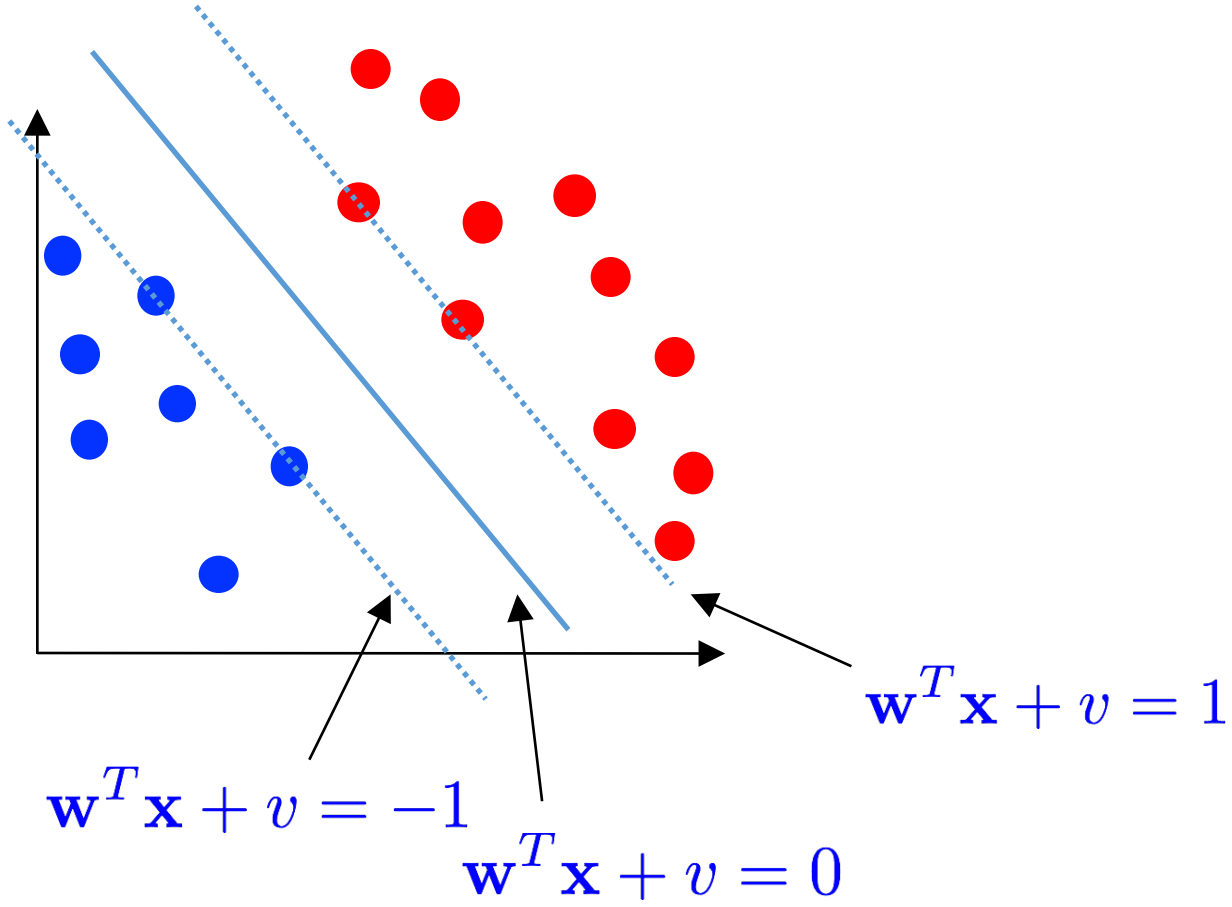
Support Vector Machines





Which one is better? → Maximum margin classifier

Support Vector Machines



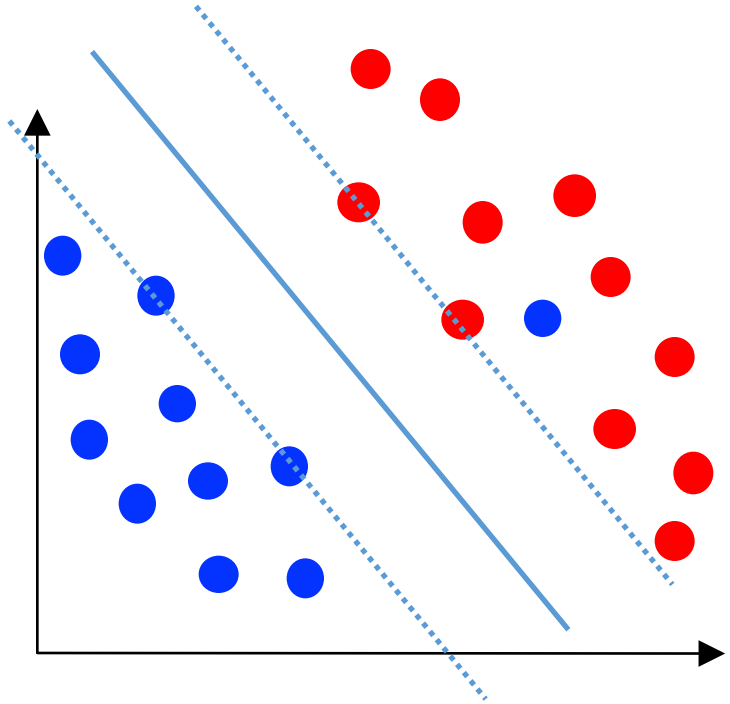
$$\max \quad \frac{2}{\|\mathbf{w}\|}$$
$$\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + v) \geq 1$$

$$\min_{\mathbf{w},v} \quad \|\mathbf{w}\|_2^2$$
$$\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + v) \geq 1$$

Convex optimization

Feasible?

Support Vector Machines: Soft-Margin



Might be infeasible

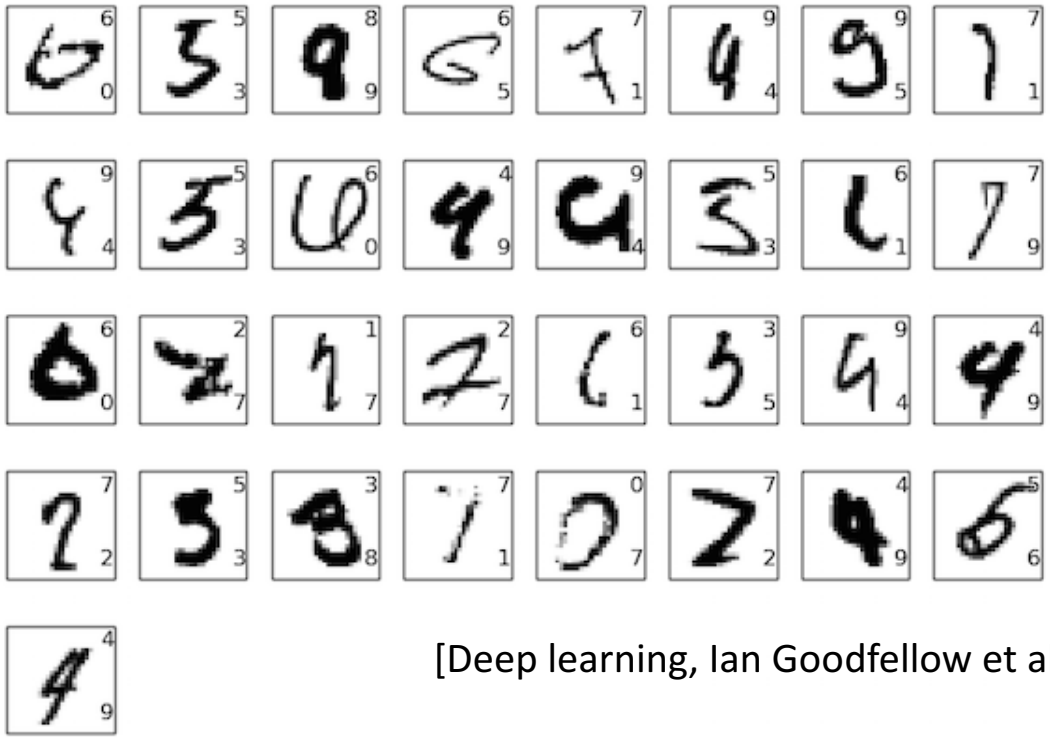
$$\min_{\mathbf{w},v} \quad \|\mathbf{w}\|_2^2$$
$$\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + v) \geq 1$$

Soft-Margin SVM:

$$\min_{\mathbf{w},v} \quad \frac{1}{n}\sum_{i=1}^{n} \max\left\{0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i + v)\right\} + \lambda\|\mathbf{w}\|_2^2$$

What is the loss function here?

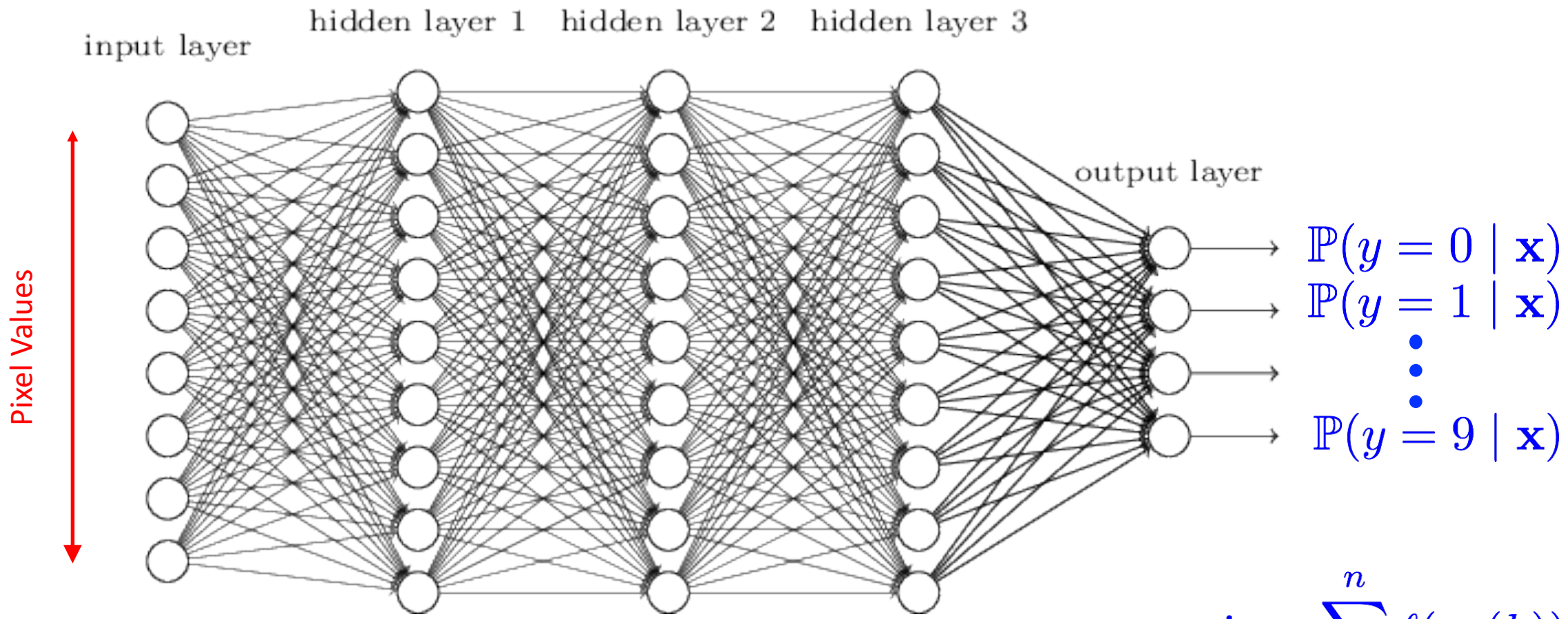
Neural Networks: Digit classification



[Deep learning, Ian Goodfellow et al. 2016]

Should be able to construct complex predictors

Neural Networks / Deep Learning



[Deep learning, Ian Goodfellow et al. 2016]

$$\min_h \sum_{i=1}^{n} \ell(\mathbf{p}_i(h))$$

SVM, regression, logistic regression can be viewed as neural networks with one layer

Empirical Risk Minimization Framework

Predicting an output $y \in \mathcal{Y}$ given an input $\mathbf{x} \in \mathcal{X}$, e.g., $\mathcal{X} = \mathbb{R}^d, \mathcal{Y} \in \{0, 1\}$

Set of hypotheses: $\mathcal{H}$ with $h \in \mathcal{H}$ maps $\mathcal{X}$ to $\mathcal{Y}$

Loss function: $\ell : (\mathcal{X} \times \mathcal{Y}) \times \mathcal{H} \mapsto \mathbb{R}$

Data generating distribution $\mathbb{P}^*$ with $(\mathbf{x}, y) \sim \mathbb{P}^*$

Expected risk/Test error: $L(h) \triangleq \mathbb{E}_{\mathbb{P}^*}\left[\ell((\mathbf{x}, y), h)\right]$  $h^* \in \arg\min_{h \in \mathcal{H}} L(h)$

Best we can hope for

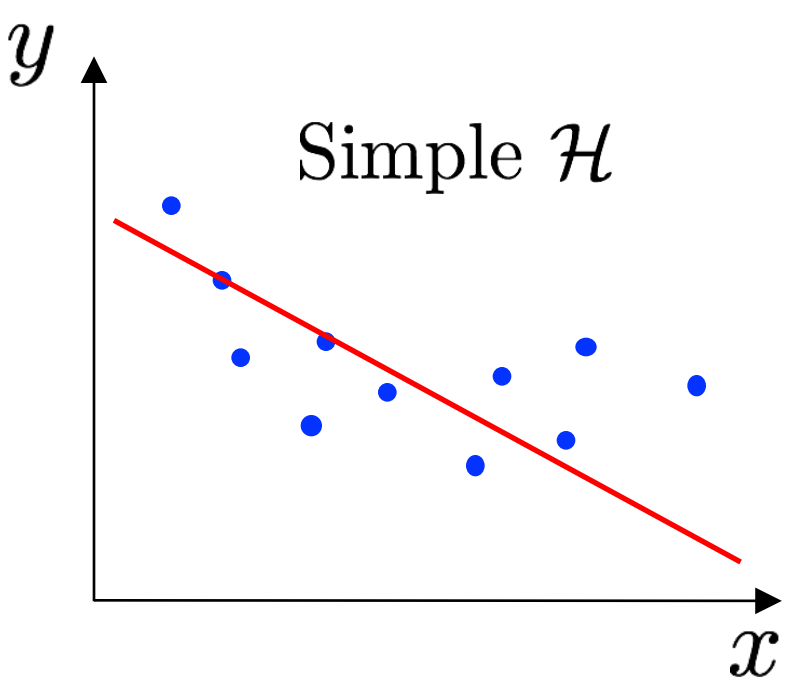
Training samples: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$

Empirical risk/Training error: $\widehat{L}(h) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell((\mathbf{x}_i, y_i), h)$  $\widehat{h} \in \arg\min_{h \in \mathcal{H}} \widehat{L}(h)$

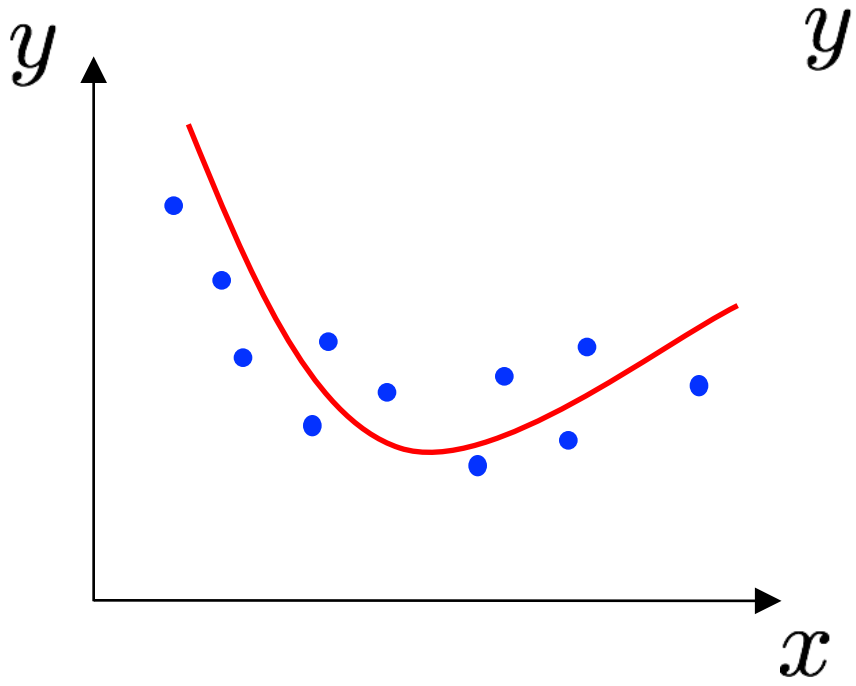
Empirical Risk Minimizer

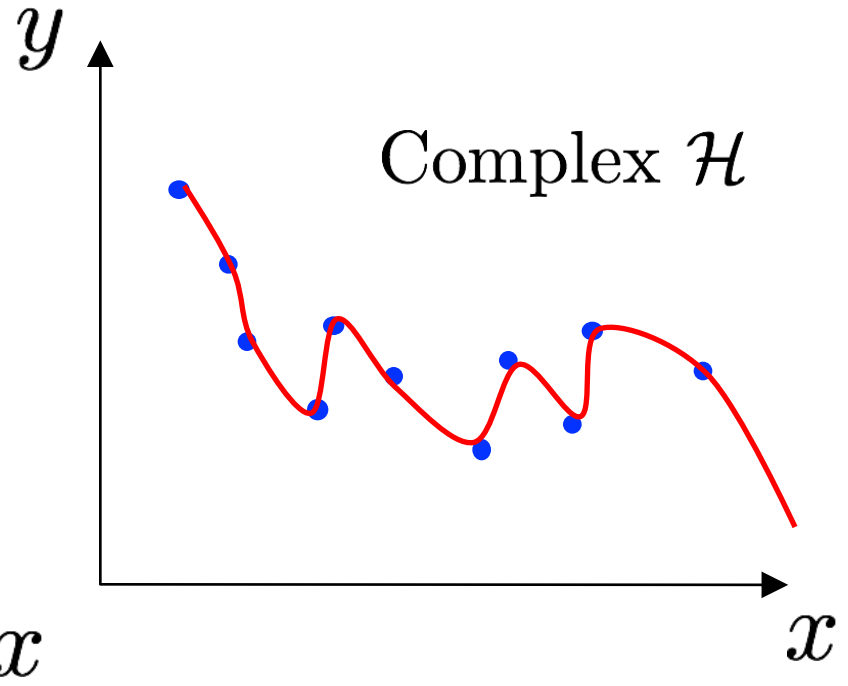
Expected risk of ERM: $L(\widehat{h})$

What Set of Hypotheses to Consider?



Underfitting

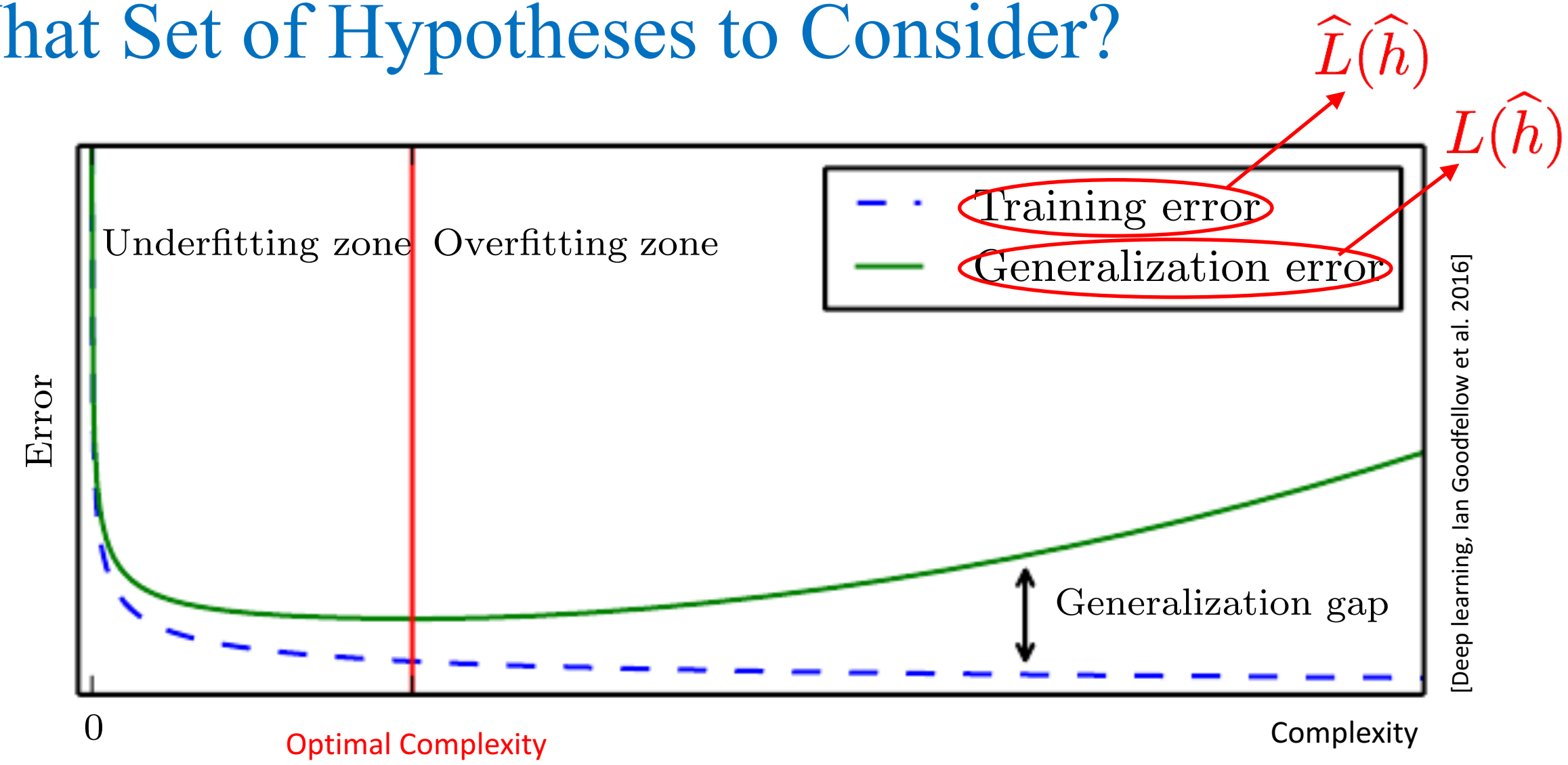




Overfitting

Trade-offs between “number of samples”, “Expected risk or ERM”, “Complexity of hypothesis class”

What Set of Hypotheses to Consider?



There are different ways of measuring complexity of a hypothesis class, but in general this trade-off exists

Simple Case

Assume:

$$L(h^*) = \mathbb{E}_{\mathbb{P}^*}\left[\ell((x,y),h^*)\right] = 0$$
$$|\mathcal{H}| < \infty$$
$$\text{zero-one loss: } \ell((x,y),h) = \mathbb{I}[y \neq h(x)]$$

Then, with probability at least $1-\delta$

$$L(\widehat{h}) - L(h^*) \leq \frac{\log|\mathcal{H}| + \log(1/\delta)}{n}$$

Proof?

Equivalent statement provides sample complexity

Remarks

Regularization typically reduces model complexity

How to estimate expected risk? Cross Validation!

Structure of the problems

- Summation in the objective
  - Stochastic optimization
  - Online optimization
  - Incremental methods

$$\min_{\mathbf{w}} \quad \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i - y_i)^2$$

$$\min_{\mathbf{w}} \quad \sum_{i=1}^{n} \left(\log \left(1 + \exp(\mathbf{w}^T \mathbf{x}_i) \right) - y_i \mathbf{w}^T \mathbf{x}_i \right)$$

- Large number of blocks
  - Block methods

$$\min_{\mathbf{w},v} \quad \frac{1}{n} \sum_{i=1}^{n} \max \left\{ 0, 1 - y_i (\mathbf{w}^T \mathbf{x}_i + v) \right\} + \lambda \|\mathbf{w}\|_2^2$$