Large Scale Optimization for Machine Learning

Meisam Razaviyayn

Lecture 12

razaviya@usc.edu

Agenda

- Review
- Regularization
- Cross validation
  - Parameter tuning
  - Termination criteria of optimization algorithms
- Structure of ERM

Recap: Empirical Risk Minimization

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

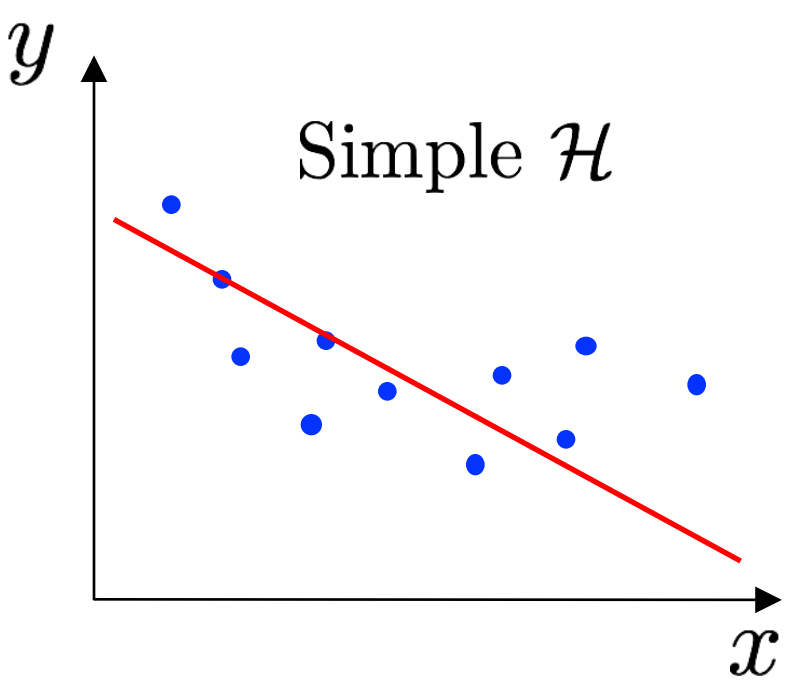
Training samples: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$

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

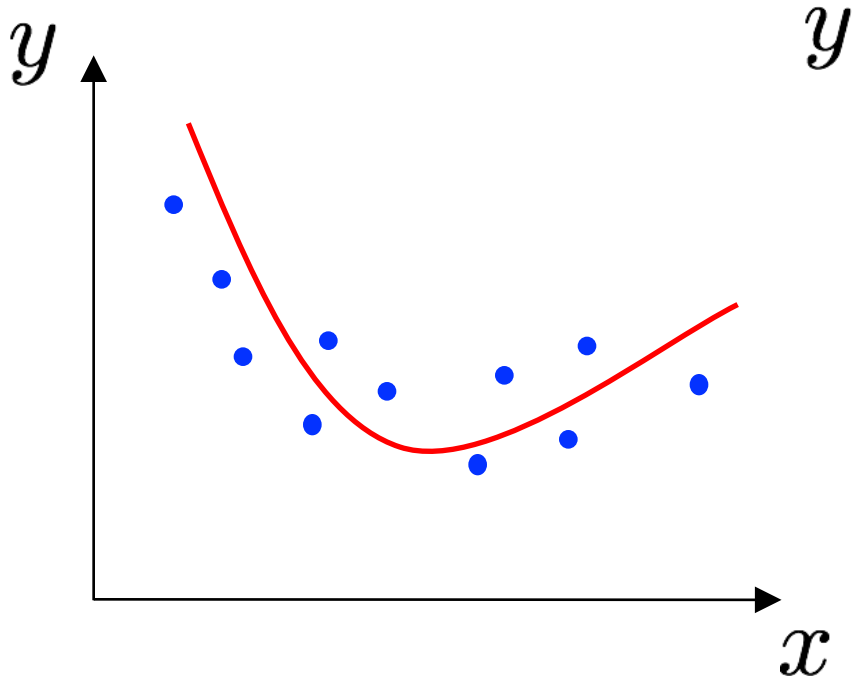
Empirical Risk Minimizer

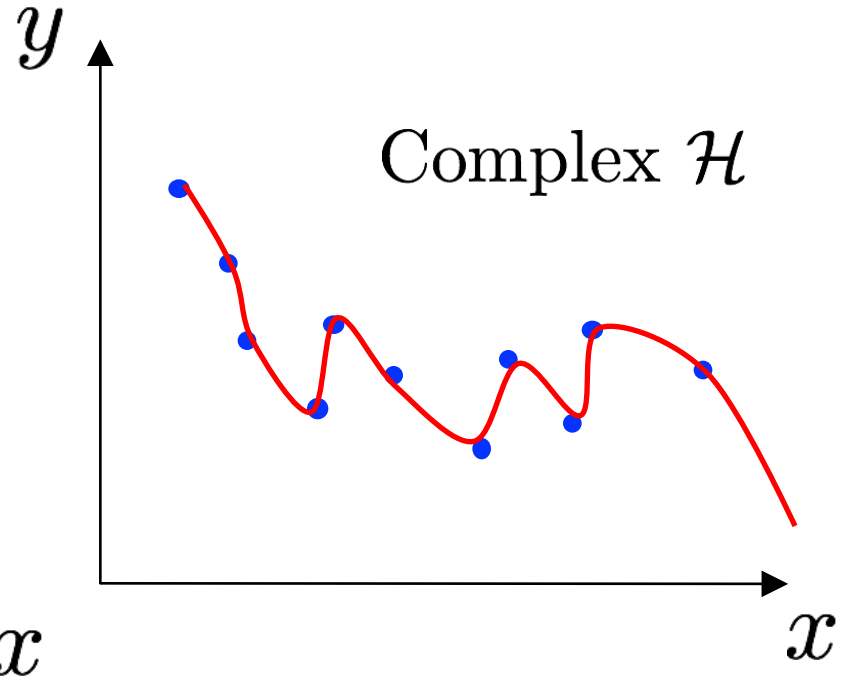
Expected risk of ERM: $L(\widehat{h})$

Recap: What Set of Hypotheses?



Underfitting



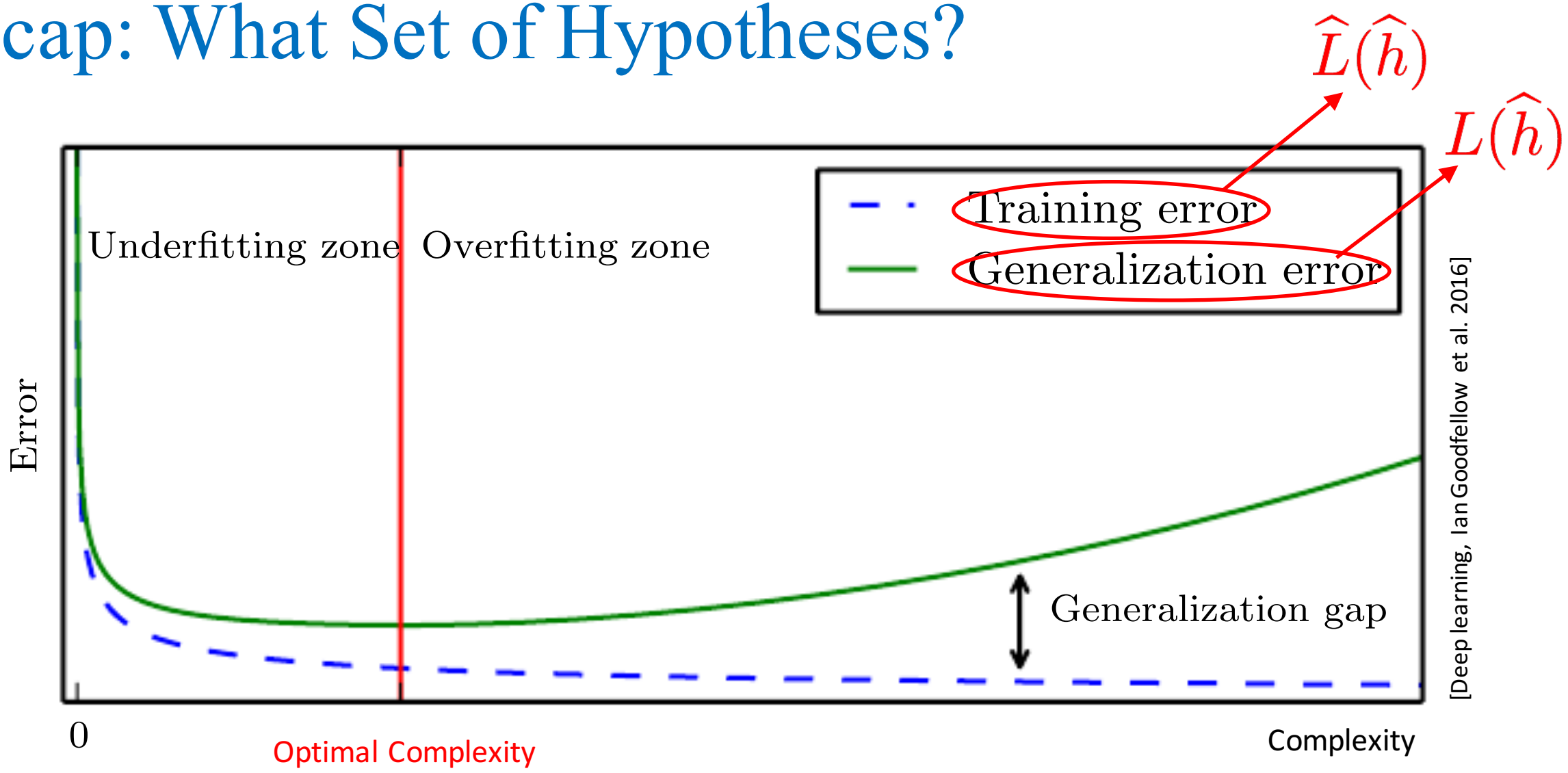


Overfitting

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

Occam's razor (William of Ockham)

Recap: What Set of Hypotheses?



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

Regularization

- Example: regression

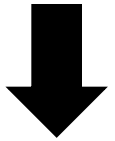
$$\mathcal{H} = \{h(\mathbf{x}) = \mathbf{w}^T\mathbf{x} \mid \mathbf{w} \in \mathbb{R}^d\}$$

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

$$\mathcal{H} = \{h(\mathbf{x}) = \mathbf{w}^T\mathbf{x} \mid \|\mathbf{w}\|_2^2 \leq \beta\}$$

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

$$\text{s.t.} \quad \|\mathbf{w}\|_2^2 \leq \beta$$



$$\min_{\mathbf{w}} \sum_{i=1}^{n} (\mathbf{w}^T\mathbf{x}_i - y_i)^2 + \lambda\|\mathbf{w}\|_2^2$$

Regularizer

Regularization

- **Goal**: reducing generalization error (expected risk) by reducing the complexity of $\mathcal{H}$

Empirical Risk Minimization

$$\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell((\mathbf{x}_i, y_i), h)$$

Regularized Empirical Risk Minimization

$$\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell((\mathbf{x}_i, y_i), h) + \lambda \mathcal{R}(h)$$

- Examples: Tikhonov regularization, total variation (TV) regularization, …
- What happens if we heavily regularize?
- How to
  - Find the correct regularizer?
  - Find the correct hypotheses class $\mathcal{H}$?

Cross Validation

- **Goal**: avoiding over/under fitting
- **Strategy**: leave some of your data for the evaluation of your fitted model $\hat{h}$

Popular Cross Validation Strategies:

- K-fold cross validation
  - Partition the samples to K partitions
  - Use K-1 partitions for training and one for validation
  - Repeat K times over all partitioning and take the average
- Leave-m-out
  - Choose m out of n samples and use them for validation and the rest for training
  - Repeat over all C(n,m) partitions and take the average (or randomly select)
  - Case m=1 is equivalent to n-fold cross validation → Leave-one-out

- **Slightly biased, but still very helpful!**

Cross validation can be used for:

- Choosing model fitting strategy
  - Example: SVM or logistic regression
- Type of regularization
  - Example: L_2 or L_1 norm
- Weight of the regularizer
- **Stopping criteria**
- Many other examples

Remarks

Regularization reduces model complexity

How to estimate expected risk and select models? Cross Validation!

Structure of ERMs

- Summation/expectations 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 optimization methods

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

Stochastic Optimization Framework

$$h^* = \arg\min_{h \in \mathcal{H}} \; \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}^*} \left[\ell((\mathbf{x}, y), h)\right] \qquad \widehat{h}_n = \arg\min_{h \in \mathcal{H}} \; \frac{1}{n}\sum_{i=1}^{n} \ell((\mathbf{x}_i, y_i), h)$$

Simplifying notations

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \; \mathbb{E}_{\xi} \left[\ell(\xi, \mathbf{w})\right] \qquad \widehat{\mathbf{w}}_n = \arg\min_{\mathbf{w}} \; \frac{1}{n}\sum_{i=1}^{n} \ell(\xi_i, \mathbf{w})$$

Different names: Empirical Risk Minimization, Sample Average Approximation

Assumption: Uniqueness of minimizer

What is the relation between the optimal $\mathbf{w}^*$ and estimated $\widehat{\mathbf{w}}_n$?

Sample Average Approximation (SAA)

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathbb{E}_\xi \left[\ell(\xi, \mathbf{w})\right]$$

$L(\mathbf{w})$

$$\widehat{\mathbf{w}}_n = \arg\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n} \ell(\xi_i, \mathbf{w})$$

$\widehat{L}_n(\mathbf{w})$

- For any fixed **w**, Law of large number implies

$$\widehat{L}_n(\mathbf{w}) \to L(\mathbf{w}) \text{ as } n \to \infty \text{ almost surely}$$

Figure

- Under some regularity conditions, by uniform convergence of LLN:

$$\widehat{\mathbf{w}}_n \to \mathbf{w}^* \text{ as } n \to \infty \text{ almost surely}$$

Sample Average Approximation (SAA)

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathbb{E}_{\xi}\left[\ell(\xi, \mathbf{w})\right] \qquad L(\mathbf{w})$$

$$\widehat{\mathbf{w}}_n = \arg\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n} \ell(\xi_i, \mathbf{w}) \qquad \widehat{L}_n(\mathbf{w})$$

$$\nu^* = \min_{\mathbf{w}} L(\mathbf{w}) \qquad \nu_n = \min_{\mathbf{w}} \widehat{L}_n(\mathbf{w})$$

- Under some regularity conditions, by LLN: $\nu_n \longrightarrow \nu^*$

Theorem : For all $n \geq 1$, $\mathbb{E}[\nu_n] \leq \mathbb{E}[\nu_{n+1}]$ **Proof?**

Training error is typically an under-estimator of the test error

SAA: Rate of Convergence

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \underbrace{\mathbb{E}_\xi\left[\ell(\xi, \mathbf{w})\right]}_{L(\mathbf{w})} \qquad \widehat{\mathbf{w}}_n = \arg\min_{\mathbf{w}} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(\xi_i, \mathbf{w})}_{\widehat{L}_n(\mathbf{w})}$$

Assume a strongly convex quadratic objective

$$\widehat{\mathbf{w}}_n = \arg\min_{\mathbf{w}} \; \widehat{L}_n(\mathbf{w}^*) + (\mathbf{w}-\mathbf{w}^*)^T \nabla \widehat{L}_n(\mathbf{w}^*) + \frac{1}{2}(\mathbf{w}-\mathbf{w}^*)^T \nabla^2 \widehat{L}_n(\mathbf{w}^*)(\mathbf{w}-\mathbf{w}^*)$$

$$\Rightarrow \sqrt{n}(\widehat{\mathbf{w}}_n - \mathbf{w}^*) = -\left(\nabla^2 \widehat{L}_n(\mathbf{w}^*)\right)^{-1}\left(\sqrt{n}\,\nabla \widehat{L}_n(\mathbf{w}^*)\right)$$

$\sqrt{n}\,\nabla \widehat{L}_n(\mathbf{w}^*)$ converges to $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ in probability

$\nabla^2 \widehat{L}_n(\mathbf{w}^*)$ converges to $\mathbf{H} \triangleq \nabla^2 L(\mathbf{w}^*)$ a.s.

$$\Rightarrow \sqrt{n}(\widehat{\mathbf{w}}_n - \mathbf{w}^*) \to \mathcal{N}(\mathbf{0}, \mathbf{H}^{-1}\boldsymbol{\Sigma}\mathbf{H}^{-1})$$ Slutsky's theorem

Want Hessian with large eigenvalues → Another justification for regularization

$$\boxed{\widehat{\mathbf{w}}_n - \mathbf{w}^* = O_p(1/\sqrt{n})}$$

Also true for general case
(under some regularity condition)