Foundations of Machine Learning

Introduction to ML

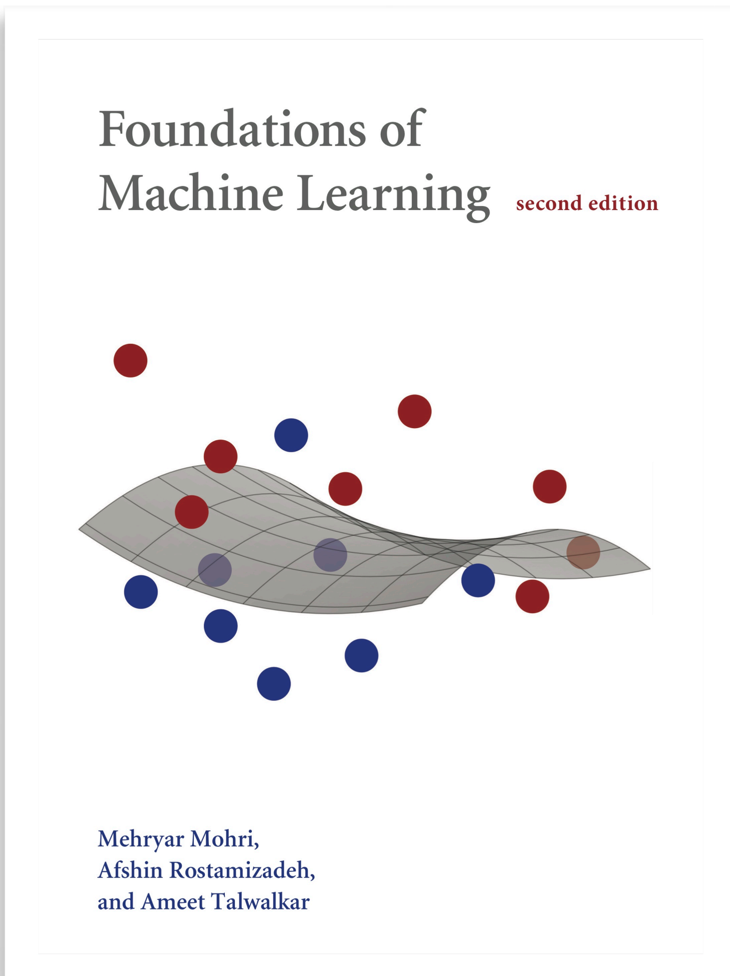
Mehryar Mohri
Courant Institute and Google Research
mohri@cims.nyu.edu

Logistics

- Prerequisites: basics in linear algebra, probability, and analysis of algorithms.
- Workload: about 3-4 homework assignments + project.
- Mailing list: join as soon as possible.

Course Material

- Textbook



- Slides: course web page.

  http://www.cs.nyu.edu/~mohri/ml23

This Lecture

- Basic definitions and concepts.
- Introduction to the problem of learning.
- Probability tools.

Machine Learning

- Definition: computational methods using experience to improve performance.
- Experience: → data-driven task, thus statistics, probability, and optimization.
- Computer science: learning algorithms, analysis of complexity, theoretical guarantees.
- Example: use document word counts to predict its topic.

Examples of Learning Tasks

- Text: document classification, spam detection.
- Language: NLP tasks (e.g., morphological analysis, POS tagging, context-free parsing, dependency parsing).
- Speech: recognition, synthesis, verification.
- Image: annotation, face recognition, OCR, handwriting recognition.
- Games (e.g., chess, backgammon, go).
- Unassisted control of vehicles (robots, car).
- Medical diagnosis, fraud detection, network intrusion.

Some Broad ML Tasks

- Classification: assign a category to each item (e.g., document classification).
- Regression: predict a real value for each item (prediction of stock values, economic variables).
- Ranking: order items according to some criterion (relevant web pages returned by a search engine).
- Clustering: partition data into 'homogenous' regions (analysis of very large data sets).
- Dimensionality reduction: find lower-dimensional manifold preserving some properties of the data.

General Objectives of ML

- Theoretical questions:
  - what can be learned, under what conditions?
  - are there learning guarantees?
  - analysis of learning algorithms.
- Algorithms:
  - more efficient and more accurate algorithms.
  - deal with large-scale problems.
  - handle a variety of different learning problems.

This Course

- Theoretical foundations:
  - learning guarantees.
  - analysis of algorithms.
- Algorithms:
  - main mathematically well-studied algorithms.
  - discussion of their extensions.
- Applications:
  - illustration of their use.

Topics

- Probability tools, concentration inequalities.
- PAC learning model, Rademacher complexity, VC-dimension, generalization bounds.
- Support vector machines (SVMs), margin bounds, kernel methods.
- Ensemble methods, boosting.
- Logistic regression and conditional maximum entropy models.
- On-line learning, weighted majority algorithm, Perceptron algorithm, mistake bounds.
- Regression, generalization, algorithms.
- Ranking, generalization, algorithms.
- Reinforcement learning, MDPs, bandit problems and algorithm.

Definitions and Terminology

- Example: item, instance of the data used.
- Features: attributes associated to an item, often represented as a vector (e.g., word counts).
- Labels: category (classification) or real value (regression) associated to an item.
- Data:
  - training data (typically labeled).
  - test data (labeled but labels not seen).
  - validation data (labeled, for tuning parameters).

General Learning Scenarios

- Settings:
  - batch: learner receives full (training) sample, which he uses to make predictions for unseen points.
  - on-line: learner receives one sample at a time and makes a prediction for that sample.
- Queries:
  - active: the learner can request the label of a point.
  - passive: the learner receives labeled points.

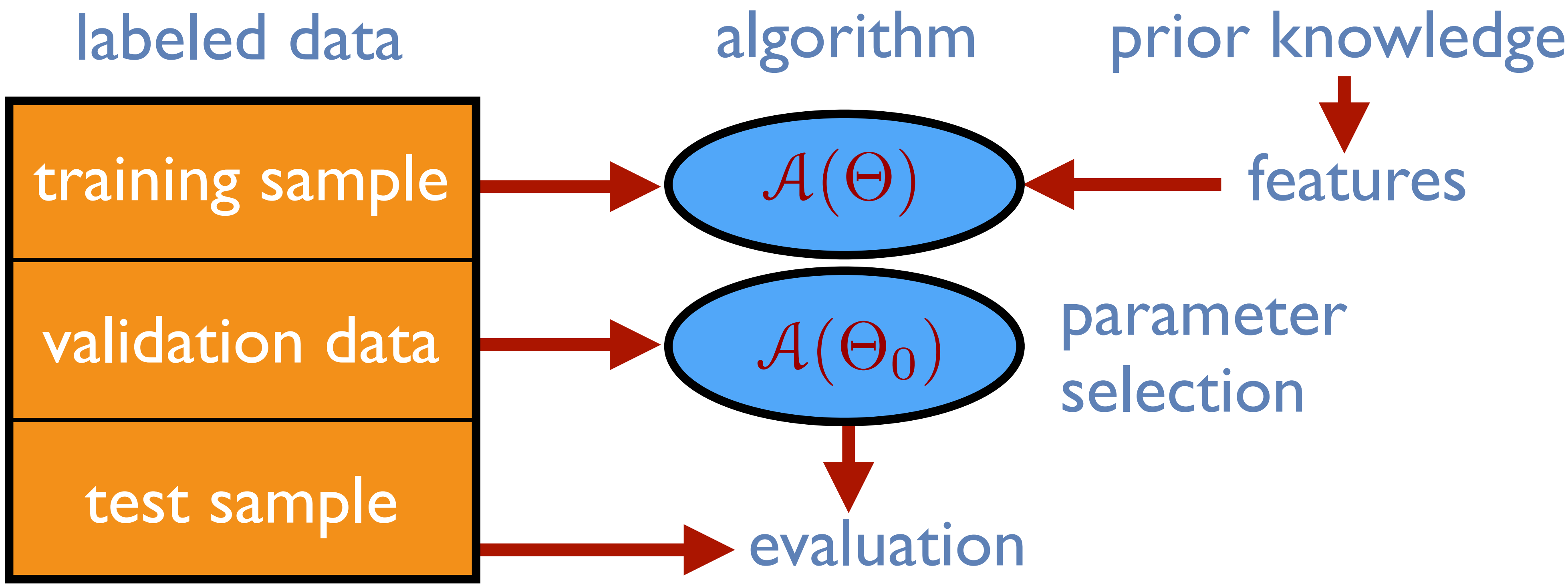
Standard Batch Scenarios

- Unsupervised learning: no labeled data.
- Supervised learning: uses labeled data for prediction on unseen points.
- Semi-supervised learning: uses labeled and unlabeled data for prediction on unseen points.
- Transduction: uses labeled and unlabeled data for prediction on seen points.

Example - SPAM Detection

- Problem: classify each e-mail message as SPAM or non-SPAM (binary classification problem).
- Potential data: large collection of SPAM and non-SPAM messages (labeled examples).

Learning Stages



This Lecture

- Basic definitions and concepts.
- Introduction to the problem of learning.
- Probability tools.

Definitions

- Spaces: input space X, output space Y.
- Loss function: $L\colon Y \times Y \to \mathbb{R}$.
  - $L(\widehat{y}, y)$: cost of predicting $\widehat{y}$ instead of y.
  - binary classification: 0-1 loss, $L(y, y') = 1_{y \neq y'}$.
  - regression: $Y \subseteq \mathbb{R}$, $l(y, y') = (y' - y)^2$.
- Hypothesis set: $H \subseteq Y^X$, subset of functions out of which the learner selects his hypothesis.
  - depends on features.
  - represents prior knowledge about task.

Supervised Learning Set-Up

- Training data: sample S of size m drawn i.i.d. from $X \times Y$ according to distribution D:

$$S = ((x_1, y_1), \ldots, (x_m, y_m)).$$

- Problem: find hypothesis $h \in H$ with small generalization error.
  - deterministic case: output label deterministic function of input, $y = f(x)$.
  - stochastic case: output probabilistic function of input.

Errors

- Generalization error: for $h \in H$, it is defined by

$$R(h) = \mathop{\mathrm{E}}_{(x,y)\sim D}[L(h(x), y)].$$

- Empirical error: for $h \in H$ and sample S, it is

$$\widehat{R}(h) = \frac{1}{m}\sum_{i=1}^{m} L(h(x_i), y_i).$$

- Bayes error:

$$R^\star = \inf_{\substack{h \\ h \text{ measurable}}} R(h).$$

  - in deterministic case, $R^\star = 0$.

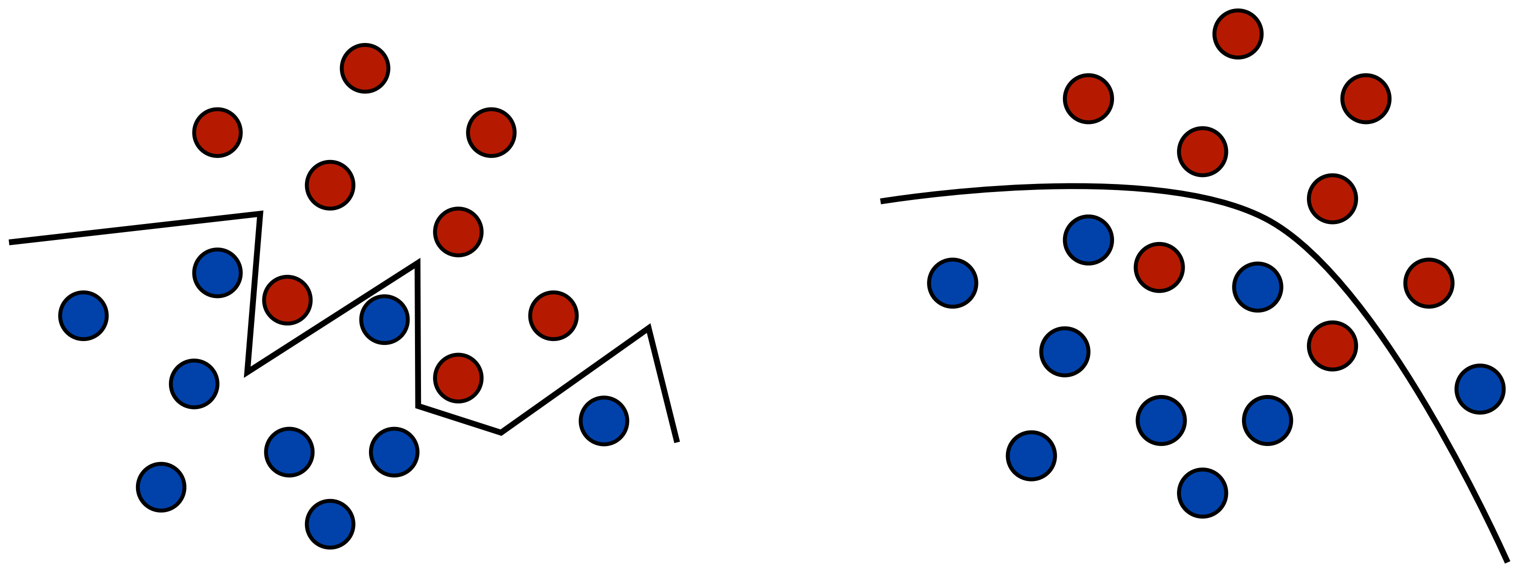
Noise

- Noise:
  - in binary classification, for any $x \in X$,

$$\text{noise}(x) = \min\{\Pr[1|x], \Pr[0|x]\}.$$

  - observe that $\mathrm{E}[\text{noise}(x)] = R^*$.

Learning ≠ Fitting



Notion of simplicity/complexity.

How do we define complexity?

Generalization

- Observations:
  - the best hypothesis on the sample may not be the best overall.
  - generalization is not memorization.
  - complex rules (very complex separation surfaces) can be poor predictors.
  - trade-off: complexity of hypothesis set vs sample size (underfitting/overfitting).

Model Selection

best in class

- General equality: for any $h \in H$,

$$R(h) - R^* = \underbrace{[R(h) - R(h^*)]}_{\text{estimation}} + \underbrace{[R(h^*) - R^*]}_{\text{approximation}}.$$

- Approximation: not a random variable, only depends on H.
- Estimation: only term we can hope to bound.
- How should we choose H?

Empirical Risk Minimization

- Select hypothesis set H.
- Find hypothesis $h \in H$ minimizing empirical error:

$$h = \operatorname*{argmin}_{h \in H} \widehat{R}(h).$$

  - but H may be too complex.
  - the sample size may not be large enough.

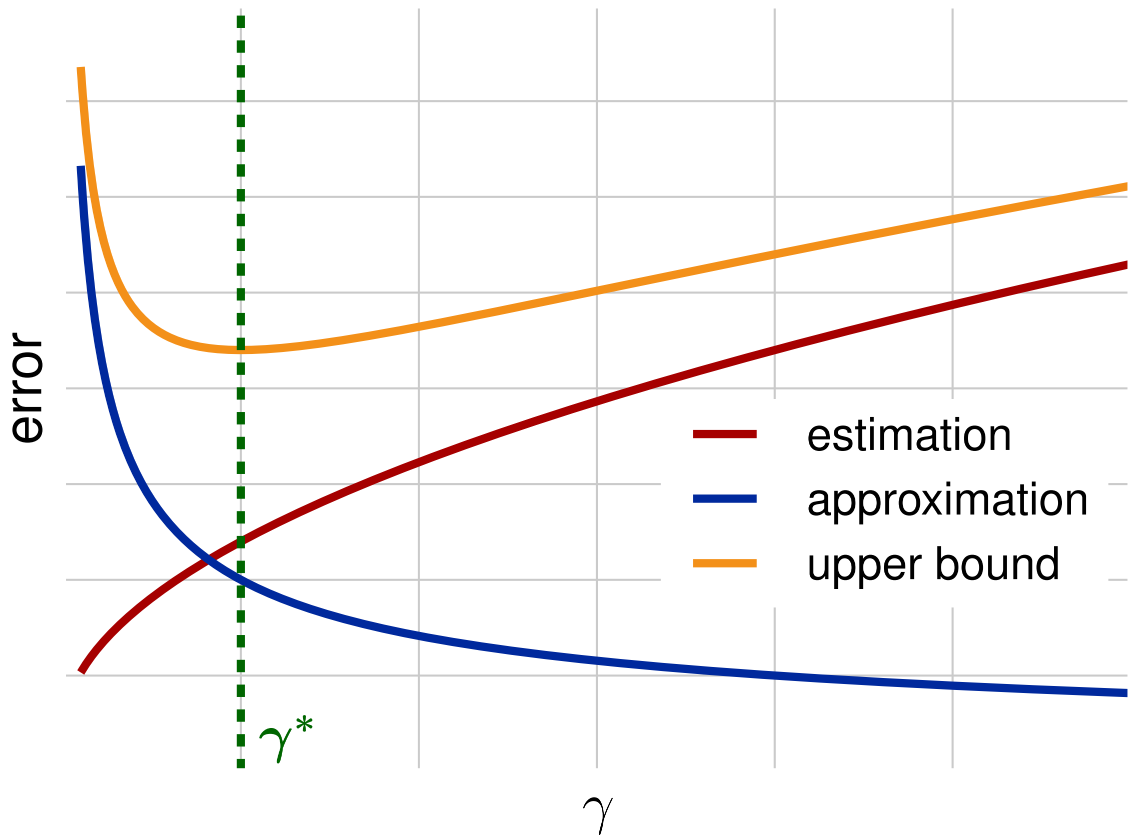
Generalization Bounds

- Definition: upper bound on $\Pr\left[\sup_{h\in H} |R(h) - \widehat{R}(h)| > \epsilon\right]$.
- Bound on estimation error for hypothesis h_0 given by ERM:

$$\begin{aligned} R(h_0) - R(h^*) &= R(h_0) - \widehat{R}(h_0) + \widehat{R}(h_0) - R(h^*) \\ &\leq R(h_0) - \widehat{R}(h_0) + \widehat{R}(h^*) - R(h^*) \\ &\leq 2 \sup_{h\in H} |R(h) - \widehat{R}(h)|. \end{aligned}$$

→ How should we choose H? (model selection problem)

Model Selection



$$\mathcal{H} = \bigcup_{\gamma \in \Gamma} \mathcal{H}_\gamma .$$

Structural Risk Minimization

(Vapnik, 1995)

- Principle: consider an infinite sequence of hypothesis sets ordered for inclusion,

$$H_1 \subset H_2 \subset \cdots \subset H_n \subset \cdots$$

$$h = \operatorname*{argmin}_{h \in H_n, n \in \mathbb{N}} \widehat{R}(h) + \text{penalty}(H_n, m).$$

  - strong theoretical guarantees.
  - typically computationally hard.

General Algorithm Families

- Empirical risk minimization (ERM):

$$h = \operatorname*{argmin}_{h \in H} \widehat{R}(h).$$

- Structural risk minimization (SRM): $H_n \subseteq H_{n+1}$,

$$h = \operatorname*{argmin}_{h \in H_n, n \in \mathbb{N}} \widehat{R}(h) + \textbf{penalty}(H_n, m).$$

- Regularization-based algorithms: $\lambda \geq 0$,

$$h = \operatorname*{argmin}_{h \in H} \widehat{R}(h) + \lambda \|h\|^2.$$

This Lecture

- Basic definitions and concepts.
- Introduction to the problem of learning.
- Probability tools.

Basic Properties

- Union bound: $\Pr[A \vee B] \leq \Pr[A] + \Pr[B]$.
- Inversion: if $\Pr[X \geq \epsilon] \leq f(\epsilon)$, then, for any $\delta > 0$, with probability at least $1-\delta$, $X \leq f^{-1}(\delta)$.
- Jensen's inequality: if f is convex, $f(\mathrm{E}[X]) \leq \mathrm{E}[f(X)]$.
- Expectation: if $X \geq 0$, $\mathrm{E}[X] = \int_0^{+\infty} \Pr[X > t]\, dt$.

Basic Inequalities

- Markov's inequality: if $X \geq 0$ and $\epsilon > 0$, then

$$\Pr[X \geq \epsilon] \leq \frac{\mathrm{E}[X]}{\epsilon}.$$

- Chebyshev's inequality: for any $\epsilon > 0$,

$$\Pr[|X - \mathrm{E}[X]| \geq \epsilon] \leq \frac{\sigma_X^2}{\epsilon^2}.$$

Hoeffding's Inequality

- Theorem: Let $X_1, \ldots, X_m$ be indep. rand. variables with the same expectation μ and $X_i \in [a, b]$, ($a < b$). Then, for any $\epsilon > 0$, the following inequalities hold:

$$\Pr\left[\mu - \frac{1}{m}\sum_{i=1}^{m} X_i > \epsilon\right] \leq \exp\left(-\frac{2m\epsilon^2}{(b-a)^2}\right)$$

$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m} X_i - \mu > \epsilon\right] \leq \exp\left(-\frac{2m\epsilon^2}{(b-a)^2}\right).$$

McDiarmid's Inequality

(McDiarmid, 1989)

- Theorem: let $X_1, \ldots, X_m$ be independent random variables taking values in U and $f: U^m \to \mathbb{R}$ a function verifying for all $i \in [1, m]$,

$$\sup_{x_1, \ldots, x_m, x'_i} |f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x'_i, \ldots, x_m)| \leq c_i.$$

Then, for all $\epsilon > 0$,

$$\Pr\Big[\big|f(X_1, \ldots, X_m) - \mathrm{E}[f(X_1, \ldots, X_m)]\big| > \epsilon\Big] \leq 2 \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^m c_i^2}\right).$$

Appendix

Markov's Inequality

- **Theorem**: let X be a non-negative random variable with $\mathrm{E}[X] < \infty$, then, for all $t > 0$,

$$\Pr[X \geq t\mathrm{E}[X]] \leq \frac{1}{t}.$$

- **Proof**:

$$\begin{aligned}
\Pr[X \geq t\,\mathrm{E}[X]] &= \sum_{x \geq t\mathrm{E}[X]} \Pr[X = x] \\
&\leq \sum_{x \geq t\,\mathrm{E}[X]} \Pr[X = x] \frac{x}{t\,\mathrm{E}[X]} \\
&\leq \sum_{x} \Pr[X = x] \frac{x}{t\,\mathrm{E}[X]} \\
&= \mathrm{E}\left[\frac{X}{t\,\mathrm{E}[X]}\right] = \frac{1}{t}.
\end{aligned}$$

Chebyshev's Inequality

- Theorem: let X be a random variable with $\mathrm{Var}[X] < \infty$, then, for all $t > 0$,

$$\Pr[|X - \mathrm{E}[X]| \geq t\sigma_X] \leq \frac{1}{t^2}.$$

- Proof: Observe that

$$\Pr[|X - \mathrm{E}[X]| \geq t\sigma_X] = \Pr[(X - \mathrm{E}[X])^2 \geq t^2\sigma_X^2].$$

The result follows Markov's inequality.

Weak Law of Large Numbers

- Theorem: let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables with the same mean μ and variance $\sigma^2<\infty$ and let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, then, for any $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr[|\overline{X}_n - \mu| \geq \epsilon] = 0.$$

- Proof: Since the variables are independent,

$$\mathrm{Var}[\overline{X}_n] = \sum_{i=1}^n \mathrm{Var}\left[\frac{X_i}{n}\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

- Thus, by Chebyshev's inequality,

$$\Pr[|\overline{X}_n - \mu| \geq \epsilon] \leq \frac{\sigma^2}{n\epsilon^2}.$$

Concentration Inequalities

- Some general tools for error analysis and bounds:
  - Hoeffding's inequality (additive).
  - Chernoff bounds (multiplicative).
  - McDiarmid's inequality (more general).

Hoeffding's Lemma

- Lemma: Let $X \in [a, b]$ be a random variable with $\mathrm{E}[X] = 0$ and $b \neq a$. Then for any $t > 0$,

$$\mathrm{E}[e^{tX}] \leq e^{\frac{t^2(b-a)^2}{8}}.$$

- Proof: by convexity of $x \mapsto e^{tx}$, for all $a \leq x \leq b$,

$$e^{tx} \leq \frac{b-x}{b-a} e^{ta} + \frac{x-a}{b-a} e^{tb}.$$

Thus,

$$\mathrm{E}[e^{tX}] \leq \mathrm{E}[\tfrac{b-X}{b-a} e^{ta} + \tfrac{X-a}{b-a} e^{tb}] = \tfrac{b}{b-a} e^{ta} + \tfrac{-a}{b-a} e^{tb} = e^{\phi(t)},$$

with,

$$\phi(t) = \log(\tfrac{b}{b-a} e^{ta} + \tfrac{-a}{b-a} e^{tb}) = ta + \log(\tfrac{b}{b-a} + \tfrac{-a}{b-a} e^{t(b-a)}).$$

- Taking the derivative gives:

$$\phi'(t) = a - \frac{ae^{t(b-a)}}{\frac{b}{b-a} - \frac{a}{b-a}e^{t(b-a)}} = a - \frac{a}{\frac{b}{b-a}e^{-t(b-a)} - \frac{a}{b-a}}.$$

- Note that: $\phi(0) = 0$ and $\phi'(0) = 0.$ Furthermore,

$$\begin{aligned}\Phi''(t) &= \frac{-abe^{-t(b-a)}}{[\frac{b}{b-a}e^{-t(b-a)} - \frac{a}{b-a}]^2} \\ &= \frac{\alpha(1-\alpha)e^{-t(b-a)}(b-a)^2}{[(1-\alpha)e^{-t(b-a)} + \alpha]^2} \\ &= \frac{\alpha}{[(1-\alpha)e^{-t(b-a)} + \alpha]} \frac{(1-\alpha)e^{-t(b-a)}}{[(1-\alpha)e^{-t(b-a)} + \alpha]}(b-a)^2 \\ &= u(1-u)(b-a)^2 \leq \frac{(b-a)^2}{4},\end{aligned}$$

with $\alpha = \dfrac{-a}{b-a}$. There exists $0 \leq \theta \leq t$ such that:

$$\phi(t) = \phi(0) + t\phi'(0) + \frac{t^2}{2}\phi''(\theta) \leq t^2 \frac{(b-a)^2}{8}.$$

Hoeffding's Theorem

- Theorem: Let $X_1, \ldots, X_m$ be independent random variables. Then for $X_i \in [a_i, b_i]$, the following inequalities hold for $S_m = \sum_{i=1}^m X_i$, for any $\epsilon > 0$,

$$\Pr[S_m - \mathrm{E}[S_m] \geq \epsilon] \leq e^{-2\epsilon^2 / \sum_{i=1}^m (b_i - a_i)^2}$$
$$\Pr[S_m - \mathrm{E}[S_m] \leq -\epsilon] \leq e^{-2\epsilon^2 / \sum_{i=1}^m (b_i - a_i)^2}.$$

- Proof: The proof is based on Chernoff's bounding technique: for any random variable X and $t > 0$, apply Markov's inequality and select t to minimize

$$\Pr[X \geq \epsilon] = \Pr[e^{tX} \geq e^{t\epsilon}] \leq \frac{\mathrm{E}[e^{tX}]}{e^{t\epsilon}}.$$

- Using this scheme and the independence of the random variables gives $\Pr[S_m - \mathrm{E}[S_m] \geq \epsilon]$

$$
\begin{aligned}
& \leq e^{-t\epsilon}\, \mathrm{E}[e^{t(S_m - \mathrm{E}[S_m])}] \\
& = e^{-t\epsilon} \Pi_{i=1}^m \, \mathrm{E}[e^{t(X_i - \mathrm{E}[X_i])}] \\
\text{(lemma applied to } X_i - \mathrm{E}[X_i]) \; & \leq e^{-t\epsilon} \Pi_{i=1}^m e^{t^2 (b_i - a_i)^2/8} \\
& = e^{-t\epsilon} e^{t^2 \sum_{i=1}^m (b_i - a_i)^2/8} \\
& \leq e^{-2\epsilon^2 / \sum_{i=1}^m (b_i - a_i)^2},
\end{aligned}
$$

  choosing $t = 4\epsilon / \sum_{i=1}^m (b_i - a_i)^2$.

- The second inequality is proved in a similar way.

Hoeffding's Inequality

- Corollary: for any $\epsilon > 0$, any distribution D and any hypothesis $h\colon X \to \{0,1\}$, the following inequalities hold:

$$\Pr[\widehat{R}(h) - R(h) \geq \epsilon] \leq e^{-2m\epsilon^2}$$
$$\Pr[\widehat{R}(h) - R(h) \leq -\epsilon] \leq e^{-2m\epsilon^2}.$$

- Proof: follows directly Hoeffding's theorem.

- Combining these one-sided inequalities yields

$$\Pr\left[\left|\widehat{R}(h) - R(h)\right| \geq \epsilon\right] \leq 2e^{-2m\epsilon^2}.$$

Chernoff's Inequality

- Theorem: for any $\epsilon > 0$, any distribution D and any hypothesis $h: X \to \{0, 1\}$, the following inequalities hold:

$$\Pr[\widehat{R}(h) \geq (1+\epsilon)R(h)] \leq e^{-m\, R(h)\, \epsilon^2/3}$$

$$\Pr[\widehat{R}(h) \leq (1-\epsilon)R(h)] \leq e^{-m\, R(h)\, \epsilon^2/2}.$$

- Proof: proof based on Chernoff's bounding technique.

McDiarmid's Inequality

(McDiarmid, 1989)

- Theorem: let $X_1, \ldots, X_m$ be independent random variables taking values in U and $f: U^m \to \mathbb{R}$ a function verifying for all $i \in [1, m]$,

$$\sup_{x_1,\ldots,x_m,x_i'} |f(x_1,\ldots,x_i,\ldots,x_m) - f(x_1,\ldots,x_i',\ldots,x_m)| \leq c_i.$$

Then, for all $\epsilon > 0$,

$$\Pr\Big[\big|f(X_1,\ldots,X_m) - \mathrm{E}[f(X_1,\ldots,X_m)]\big| > \epsilon\Big] \leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^m c_i^2}\right).$$

- Comments:
  - Proof: uses Hoeffding's lemma.
  - Hoeffding's inequality is a special case of McDiarmid's with

$$f(x_1, \ldots, x_m) = \frac{1}{m}\sum_{i=1}^{m} x_i \quad \text{and} \quad c_i = \frac{|b_i - a_i|}{m}.$$

Jensen's Inequality

- Theorem: let X be a random variable and f a measurable convex function. Then,

$$f(\mathrm{E}[X]) \leq \mathrm{E}[f(X)].$$

- Proof: definition of convexity, continuity of convex functions, and density of finite distributions.