Law of large numbers:

if we have $X_1, X_2, \cdots, X_n$ are i.i.d.,then

$\frac{1}{n}\sum_{i=1}^{n} X_i \Rightarrow E[X]$

This is asmptotic bound.

Concentration inequalities using method

- Markov's inequality
  $Pr[X \geq t] \leq \frac{E[X]}{t} \qquad \forall t > 0$
- Chebyshevs inequality： extension
  $Pr[|X - E[X] \geq t] \leq \frac{var(x)}{t^2} \qquad \forall t > 0$
  - suppose$X_1, X_2, X_3, \cdots, X_n$ are i.i.d. random variables with zero mean
    $Var(\bar{X}) = \frac{Var(X)}{n}$
    Applying Chebysheves inequality, for any $t \geq 0$, we have
    $Pr[|\bar{X} - E[X] \geq t] \leq \frac{var(x)}{nt^2}$
    large n or samll variance imply better concentration

Concentration inequalities for sub_Gaussian

- MGF(Moment generating function)
  $M_X(\lambda) = E[exp(\lambda X)]$
  - normal distribution $X \sim N(0, \sigma^2)$
    $M_X(\lambda) = exp(\frac{\lambda^2\sigma^2}{2})$
  - Rademacher random variable:
    $Pr[\epsilon = 1] = \frac{1}{2} \quad and \quad Pr[\epsilon = -1] = \frac{1}{2}$, in this case $\sigma^2 = 1$
    $M_\epsilon(\lambda) \leq exp(\frac{\lambda^2}{2})$
- Chernoff bounds
  $Pr[X - E[X] \geq t] = Pr[e^{\lambda(X - E[X])} \geq e^{\lambda t}] \leq \frac{E[e^{\lambda}(X - E[X])}{e^{\lambda t}}$(Applying the Markov's inequality)
  $\Rightarrow Pr[X - E[X] \geq t] \leq min_{\lambda \geq 0} E[e^{\lambda(X - E[X])}]e^{-\lambda t}$
- sub-Gaussian random variable
  $E[e^{\lambda(x-\mu)}] \leq e^{(\frac{\lambda^2\sigma^2}{2})}$
  - Gaussian distribution
  - Rademacher random variable
  - any bounded random variable

- sub-Gaussian concentration
  $Pr[X - E[X] \geq t] \leq 2e^{-\frac{t^2}{2\sigma^2}}$

Hoeffding's inequality

$Pr[\sum_{i=1}^{n}(X_i - \mu_i) \geq t] \leq e^{\frac{-t^2}{2\sum_{i=1}^{n}\sigma_i^2}}$
for bounded with mean μ_i and bounded on $[a_i, b_i]$
$Pr[\sum_{i=1}^{n}(X_i - \mu_i) \geq t] \leq e^{\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}$

Generalization

$Er_{out} = Er_{out} - Er_{in} + Er_{in}$
generalization:$Er_{out} - Er_{in}$(less complex model/hypothesis H)
training: Er_{in}(more complex model/hypothesis H)
Theory of generalization: bounding the generalization error

In-sample error vs out-of-sample error

- In expection of fixed $f \in H$
  $Er_{in}(f) = E_{s \sim D}[Er_{in}(f)]$
  - Er_{in} is unbiased estimator for Er_{out}
  - Law of large numbers:
    when $n \to \infty$, we have consistency

generalization for fixed model f: a lemma

- high probability bounds:
  $Pr[|Er_{in}(f) - Er_{out}(f)| \geq t] \leq 2e^{-2nt^2}$
  $Pr[|Er_{in}(f) - Er_{out}(f)| \leq t] \leq 1 - 2e^{-2nt^2}$
  (proved by Hoeffding's inequality for bounded random variables)
- Generalization bound-fixed f
  $Er_{out}(f) \leq Er_{in}(f) + \sqrt{\frac{log(\frac{2}{\sigma})}{2n}}$ (with probability at least $1 - \sigma \quad \forall \sigma > 0$)

generalization bound for finite model space

$\forall f \in H \quad Er_{out}(f) \leq Er_{in}(f) + \sqrt{\frac{log|H| + log\frac{2}{\sigma}}{2n}}$

- on the training side, we need: more complex model/hypothesis H
- on the generalization side, we need: less complex model/hypothesis H

Empirical Radmacher complexity

let $S = \{z_i = (x_i, y_i)\}_{i=1}^n$

$\hat{R}_s(H) := E_\epsilon[sup_{f\in H} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i)]$ (ϵ_i is a Rademacher random variable)

- Rademacher complexity
  The Rademacher complexity of H over the sample S woith respect to the distribution D is defined as:
  $R(H) := E_{s\sim i.i.d.D}[\hat{R}_s(H)]$
- Rademacher complexity: interpretation
  $R(H) = E_{S,\epsilon}[sup_{f\in H} \frac{\epsilon^T e_S}{n}]$($e_S$ is the vector of $f(x)$)
  $\epsilon^T e_S = ||\epsilon|| \cdot ||e_S|| \cdot cos(\alpha)$
  more complex H can generate more vectors e_S,thus have better chance to correlate teh random noise ϵ, on average.

Generalization bound using Rademacher complexity

$\forall f \in H \quad Er_{out}(f) \leq Er_{in}(f) + R(H) + \sqrt{\frac{log\frac{1}{\sigma}}{2n}}$

McDiarmid's inequality

let $S = \{z_1, \cdots, z_i, \cdots, z_n\}$

Suppose that $|h(z_1, \cdots, z_i, \cdots, z_n) - h(z_1, \cdots, z_i', \cdots, z_n)| \leq c_i$

we have $Pr[h(S) - E[h(S)] \geq t] \leq e^{\frac{2t^2}{\sum_{i=1}^n c_i^2}}$

- Look at the supremum of generalization error:
  $h(S) := sup_{f\in H}[Er_{out}(f) - Er_{in}(f; S)]$
  $h(S') - h(S) \leq \frac{1}{n}$
- Apply McDiarmid's inequality to $h(S)$, we have
  $Pr[h(S) - E[h(S)] \geq t] \leq e^{-2nt^2}$
  set $\sigma = e^{-2nt^2}$
  $Er_{out}(f) \leq Er_{in}(f) + E[h(S)] + \sqrt{\frac{log(\frac{1}{\sigma})}{2n}}$
  $E[h(S)] := 2R(L) = R(H)$

Growth function

$\forall n \in N \quad G_H(n) = max_{\{x_1, \cdots, x_n\}} |\{f(x_1), \cdots, f(x_n)\} : f \in H|$

- $G_H(n)$ counts the most dichotomies that can possibly generated on n points in X
- measure the richness of the hypothesis set H
- combinatorial concept, independent of distribution D
- Generalization bound using growth function

$$\forall f \in H \quad Er_{out}(f) \leq Er_{in}(f) + \sqrt{\frac{2logG_H(n)}{n}} + \sqrt{\frac{log\frac{1}{\sigma}}{2n}}$$

VC-Dimension

- The VC-dimension of a hypothesis set H is the size of teh largest dataset that can be shattered by H :
  $VCdim(H) := max\{n : G_H(n) = 2^n\}$
- let the hypothesis set H be a hyperplane in R^d, then
  $VCdim(H) = d + 1$
- generalization bound:
  $\forall f \in H \quad Er_{out}(f) \leq Er_{in}(f) + \sqrt{\frac{2dlog\frac{e \cdot n}{d}}{n}} + \sqrt{\frac{log\frac{1}{\sigma}}{2n}}$