# Contents

In this lecture we consider the classification problem, i.e. $\mathcal{Y} = \{-1, +1\}$.

Consider a family of weak classifiers

$$\mathcal{H} = \{h \colon \mathcal{X} \to \{-1, +1\}\}.$$

Let the empirical minimizer be

$$h_0 = \text{argmin} \, \frac{1}{n} \sum_{i=1}^{n} I(h(X_i) \neq Y_i)$$

and assume its expected error,

$$\frac{1}{2} > \varepsilon = Error(h_0), \ \varepsilon > 0$$

Examples:

- $\mathcal{X} = \mathbb{R}^d$, $\mathcal{H} = \{\text{sign}(wx + b) \colon w \in \mathbb{R}^d, b \in \mathbb{R}\}$
- Decision trees: restrict depth.
- Combination of simple classifiers:

  $$f = \sum_{t=1}^{T} \alpha_t h_t(x),$$

  where $h_t \in \mathcal{H}$, $\sum_{t=1}^{T} \alpha_t = 1$. For example,

$$h_1 = \begin{array}{|c|c|}\hline 1 & -1 \\ \hline 1 & -1 \\ \hline \end{array}, \qquad h_2 = \begin{array}{|c|c|}\hline 1 & 1 \\ \hline -1 & -1 \\ \hline \end{array}, \qquad h_3 = \begin{array}{|c|c|}\hline 1 & 1 \\ \hline 1 & 1 \\ \hline \end{array}$$

$$f = \tfrac{1}{7}(h_1 + 3h_2 + 3h_3) = \begin{array}{|c|c|}\hline 7 & 5 \\ \hline 1 & -1 \\ \hline \end{array}, \qquad \text{sign}(f) = \begin{array}{|c|c|}\hline 1 & 1 \\ \hline 1 & -1 \\ \hline \end{array}$$

**AdaBoost**

Assign weight to training examples $w_1(i) = 1/n$.

for $t = 1..T$

1) find "good" classifier $h_t \in \mathcal{H}$; Error $\varepsilon_t = \sum_{i=1}^{n} w_t(i) I(h(X_i) \neq Y_i)$

2) update weight for each $i$:

$$w_{t+1}(i) = \frac{w_t(i) e^{-\alpha_t Y_i h_t(X_i)}}{Z_t}$$

$$Z_t = \sum_{i=1}^{n} w_t(i) e^{-\alpha_t Y_i h_t(X_i)}$$

$$\alpha_t = \frac{1}{2} \ln \frac{1 - \varepsilon_t}{\varepsilon_t} > 0$$

3) t = t+1

end

Output the final classifier: $f = \text{sign}(\sum \alpha_t h_t(x))$.

**Theorem 2.1.** *Let $\gamma_t = 1/2 - \varepsilon_t$ (how much better $h_t$ is than tossing a coin). Then*

$$\frac{1}{n}\sum_{i=1}^{n} I(f(X_i) \neq Y_i) \leq \prod_{t=1}^{T} \sqrt{1-4\gamma_t^2}$$

*Proof.*

$$I(f(X_i) \neq Y_i) = I(Y_i f(X_i) = -1) = I(Y_i \sum_{t=1}^{T} \alpha_t h_t(X_i) \leq 0) \leq e^{-Y_i \sum_{t=1}^{T} \alpha_t h_t(X_i)}$$

Consider how weight of example $i$ changes:

$$\begin{aligned}
w_{T+1}(i) &= \frac{w_T(i) e^{-Y_i \alpha_T h_T(X_i)}}{Z_t} \\
&= \frac{e^{-Y_i \alpha_T h_T(X_i)}}{Z_t} \frac{w_{T-1}(i) e^{-Y_i \alpha_{T-1} h_{T-1}(X_i)}}{Z_{T-1}} \\
\ldots \\
&= \frac{e^{-Y_i \sum_{t=1}^{T} \alpha_t h_t(X_i)}}{\prod_{t=1}^{t} Z_t} \frac{1}{n}
\end{aligned}$$

Hence,

$$w_{T+1}(i) \prod Z_t = \frac{1}{n} e^{-Y_i \sum_{t=1}^{T} \alpha_t h_t(X_i)}$$

and therefore

$$\frac{1}{n}\sum_{i=1}^{n} I(f(X_i) \neq Y_i) \leq \frac{1}{n}\sum_{i=1}^{n} e^{-Y_i \sum_{t=1}^{T} \alpha_t h_t(X_i)} = \prod_{t=1}^{T} Z_t \sum_{i=1}^{n} w_{T+1}(i) = \prod_{t=1}^{T} Z_t$$

$$\begin{aligned}
Z_t &= \sum w_t(i) e^{-\alpha_t Y_i h_t(X_i)} \\
&= \sum_{i=1}^{n} w_t(i) e^{-\alpha_t} I(h_t(X_i) = Y_i) + \sum_{i=1}^{n} w_t(i) e^{+\alpha_t} I(h_t(X_i) \neq Y_i) \\
&= e^{+\alpha_t} \sum_{i=1}^{n} w_t(i) I(h_t(X_i) \neq Y_i) + e^{-\alpha_t} \sum_{i=1}^{n} w_t(i)(1 - I(h_t(X_i) \neq Y_i)) \\
&= e^{\alpha_t} \varepsilon_t + e^{-\alpha_t}(1-\varepsilon_t)
\end{aligned}$$

Minimize over $\alpha_t$ to get

$$\alpha_t = \frac{1}{2} \ln \frac{1-\varepsilon_t}{\varepsilon_t}$$

and

$$e^{\alpha_t} = \left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)^{1/2}.$$

Finally,

$$
\begin{aligned}
Z_t &= \left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)^{1/2}\varepsilon_t + \left(\frac{\varepsilon_t}{1-\varepsilon_t}\right)^{1/2}(1-\varepsilon_t) \\
&= 2(\varepsilon_t(1-\varepsilon_t))^{1/2} = 2\sqrt{(1/2-\gamma_t)(1/2+\gamma_t)} \\
&= \sqrt{1-4\gamma_t^2}
\end{aligned}
$$

□

As in the previous lecture, consider the classification setting. Let $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{+1, -1\}$, and

$$\mathcal{H} = \{\psi x + b,\ \psi \in \mathbb{R}^d,\ b \in \mathbb{R}\}$$

where $|\psi| = 1$.

We would like to maximize over the choice of hyperplanes the minimal distance from the data to the hyperplane:

$$\max_H \min_i d(x_i, H),$$

where

$$d(x_i, H) = y_i(\psi x_i + b).$$

Hence, the problem is formulated as maximizing the margin:

$$\max_{\psi,b} \underbrace{\min_i y_i(\psi x_i + b)}_{m \text{ (margin)}}.$$

Rewriting,

$$y_i(\psi' x_i + b') = \frac{y_i(\psi x_i + b)}{m} \geq 1,$$

$\psi' = \psi/m$, $b' = b/m$, $|\psi'| = |\psi|/m = 1/m$. Maximizing $m$ is therefore minimizing $|\psi'|$. Rename $\psi' \to \psi$, we have the following formulation:

$$\min |\psi| \quad \text{such that} \quad y_i(\psi x_i + b) \geq 1$$

Equivalently,

$$\min \frac{1}{2}\psi \cdot \psi \quad \text{such that} \quad y_i(\psi x_i + b) \geq 1$$

Introducing Lagrange multipliers:

$$\phi = \frac{1}{2}\psi \cdot \psi - \sum \alpha_i(y_i(\psi x_i + b) - 1),\ \ \alpha_i \geq 0$$

Take derivatives:

$$\frac{\partial \phi}{\partial \psi} = \psi - \sum \alpha_i y_i x_i = 0$$

$$\frac{\partial \phi}{\partial b} = -\sum \alpha_i y_i = 0$$

Hence,

$$\psi = \sum \alpha_i y_i x_i$$

and

$$\sum \alpha_i y_i = 0.$$

Substituting these into $\phi$,

$$\begin{aligned}\phi &= \frac{1}{2}\left(\sum \alpha_i y_i x_i\right)^2 - \sum_{i=1}^n \alpha_i \left( y_i \left( \sum_{j=1}^n \alpha_j y_j x_j x_i + b \right) - 1 \right) \\ &= \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i x_j - \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i x_j - b \sum \alpha_i y_i + \sum \alpha_i \\ &= \sum \alpha_i - \frac{1}{2} \sum \alpha_i \alpha_j y_i y_j x_i x_j\end{aligned}$$

The above expression has to be maximized this with respect to $\alpha_i$, $\alpha_i \geq 0$, which is a Quadratic Programming problem.

Hence, we have $\psi = \sum_{i=1}^n \alpha_i y_i x_i$.

Kuhn-Tucker condition:

$$\alpha_i \neq 0 \Leftrightarrow y_i(\psi x_i + b) - 1 = 0.$$

Throwing out non-support vectors $x_i$ does not affect hyperplane $\Rightarrow \alpha_i = 0$.

The mapping $\phi$ is a *feature* mapping:

$$x \in \mathbb{R}^d \longrightarrow \phi(x) = (\phi_1(x), \phi_2(x), ...) \in \mathcal{X}'$$

where $\mathcal{X}'$ is called *feature space.*

Support Vector Machines find optimal separating hyperplane in a very high-dimensional space. Let $K(x_i, x_j) = \sum_{k=1}^\infty \phi_k(x_i)\phi_k(x_j)$ be a scalar product in $\mathcal{X}'$. Notice that we don't need to know mapping $x \to \phi(x)$. We only need to know $K(x_i, x_j) = \sum_{k=1}^\infty \phi_k(x_i)\phi_k(x_j)$, a symmetric positive definite kernel.

Examples:

(1) Polynomial: $K(x_1, x_2) = (x_1 x_2 + 1)^\ell,\ \ell \geq 1$.

(2) Radial Basis: $K(x_1, x_2) = e^{-\gamma |x_1 - x_2|^2}$.

(3) Neural (two-layer): $K(x_1, x_2) = \frac{1}{1+e^{\alpha x_1 x_2 + \beta}}$ for some $\alpha, \beta$ (for some it's not positive definite).

Once $\alpha_i$ are known, the decision function becomes

$$\text{sign}\left(\sum \alpha_i y_i x_x \cdot x + b\right) = \text{sign}\left(\sum \alpha_i y_i K(x_i, x) + b\right)$$

Assume we have samples $z_1 = (x_1, y_1), \ldots, z_n = (x_n, y_n)$ as well as a new sample $z_{n+1}$. The classifier trained on the data $z_1, \ldots, z_n$ is $f_{z_1,\ldots,z_n}$.
The error of this classifier is

$$\text{Error}(z_1, \ldots, z_n) = \mathbb{E}_{z_{n+1}} I(f_{z_1,\ldots,z_n}(x_{n+1}) \neq y_{n+1}) = \mathbb{P}_{z_{n+1}}(f_{z_1,\ldots,z_n}(x_{n+1}) \neq y_{n+1})$$

and the *Average Generalization Error*

$$\text{A.G.E.} = \mathbb{E}\ \text{Error}(z_1, \ldots, z_n) = \mathbb{E}\mathbb{E}_{z_{n+1}} I(f_{z_1,\ldots,z_n}(x_{n+1}) \neq y_{n+1}).$$

Since $z_1, \ldots, z_n, z_{n+1}$ are i.i.d., in expectation training on $z_1, \ldots, z_i, \ldots, z_n$ and evaluating on $z_{n+1}$ is the same as training on $z_1, \ldots, z_{n+1}, \ldots, z_n$ and evaluating on $z_i$. Hence, for any $i$,

$$\text{A.G.E.} = \mathbb{E}\mathbb{E}_{z_i} I(f_{z_1,\ldots,z_{n+1},\ldots,z_n}(x_i) \neq y_i)$$

and

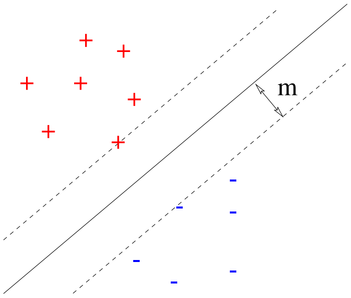
$$\text{A.G.E.} = \mathbb{E}\left[\underbrace{\frac{1}{n+1}\sum_{i=1}^{n+1} I(f_{z_1,\ldots,z_{n+1},\ldots,z_n}(x_i) \neq y_i)}_{\text{leave-one-out error}}\right].$$

Therefore, to obtain a bound on the generalization ability of an algorithm, it's enough to obtain a bound on its leave-one-out error. We now prove such a bound for SVMs. Recall that the solution of SVM is $\varphi = \sum_{i=1}^{n+1} \alpha_i^0 y_i x_i$.

**Theorem 4.1.**

$$L.O.O.E. \leq \frac{\min(\#\ \textit{support vect.}, D^2/m^2)}{n+1}$$

*where $D$ is the diameter of a ball containing all $x_i$, $i \leq n+1$ and $m$ is the margin of an optimal hyperplane.*



**Remarks:**

- dependence on sample size is $\frac{1}{n}$
- dependence on margin is $\frac{1}{m^2}$
- number of support vectors (sparse solution)

**Lemma 4.1.** *If $x_i$ is a support vector and it is misclassified by leaving it out, then $\alpha_i^0 \geq \frac{1}{D^2}$.*

Given Lemma 4.1, we prove Theorem 4.1 as follows.

*Proof.* Clearly,

$$\text{L.O.O.E.} \leq \frac{\#\text{ support vect.}}{n+1}.$$

Indeed, if $x_i$ is not a support vector, then removing it does not affect the solution. Using Lemma 4.1 above,

$$\sum_{i\in\text{supp.vect}} I(x_i \text{ is misclassified}) \leq \sum_{i\in\text{supp.vect}} \alpha_i^0 D^2 = D^2 \sum \alpha_i^0 = \frac{D^2}{m^2}.$$

In the last step we use the fact that $\sum \alpha_i^0 = \frac{1}{m^2}$. Indeed, since $|\varphi| = \frac{1}{m}$,

$$\begin{aligned}
\frac{1}{m^2} &= |\varphi|^2 = \varphi\cdot\varphi = \varphi\cdot\sum \alpha_i^0 y_i x_i \\
&= \sum \alpha_i^0 (y_i \varphi\cdot x_i) \\
&= \underbrace{\sum \alpha_i^0 (y_i(\varphi\cdot x_i + b) - 1)}_{0} + \sum \alpha_i^0 - b \underbrace{\sum \alpha_i^0 y_i}_{0} \\
&= \sum \alpha_i^0
\end{aligned}$$

□

We now prove Lemma 4.1. Let $u * v = K(u,v)$ be the dot product of $u$ and $v$, and $\|u\| = (K(u,u))^{1/2}$ be the corresponding $L_2$ norm. Given $x_1, \cdots, x_{n+1} \in \mathbb{R}^d$ and $y_1, \cdots, y_{n+1} \in \{-1,+1\}$, recall that the primal problem of training a support vector classifier is $\text{argmin}_\psi \frac{1}{2}\|\psi\|^2$ subject to $y_i(\psi * x_i + b) \geq 1$. Its dual problem is $\text{argmax}_\alpha \sum \alpha_i - \frac{1}{2}\left\|\sum \alpha_i y_i x_i\right\|^2$ subject to $\alpha_i \geq 0$ and $\sum \alpha_i y_i = 0$, and $\psi = \sum \alpha_i y_i x_i$. Since the Kuhn-Tucker condition can be satisfied, $\min_\psi \frac{1}{2}\psi * \psi = \max_\alpha \sum \alpha_i - \frac{1}{2}\left\|\sum \alpha_i y_i x_i\right\|^2 = \frac{1}{2m^2}$, where $m$ is the margin of an optimal hyperplane.

*Proof.* Define $w(\alpha) = \sum_i \alpha_i - \frac{1}{2}\left\|\sum \alpha_i y_i x_i\right\|^2$. Let $\alpha^0 = \text{argmax}_\alpha w(\alpha)$ subject to $\alpha_i \geq 0$ and $\sum \alpha_i y_i = 0$. Let $\alpha' = \text{argmax}_\alpha w(\alpha)$ subject to $\alpha_p = 0$, $\alpha_i \geq 0$ for $i \neq p$ and $\sum \alpha_i y_i = 0$. In other words, $\alpha^0$ corresponds to the support vector classifier trained from $\{(x_i, y_i) : i = 1, \cdots, n+1\}$ and $\alpha'$ corresponds to the support vector classifier trained from $\{(x_i, y_i) : i = 1, \cdots, p-1, p+1, \cdots, n+1\}$. Let $\gamma = \left(\begin{array}{ccccccc} \overset{1}{\downarrow} & & \overset{p-1}{\downarrow} & \overset{p}{\downarrow} & \overset{p+1}{\downarrow} & & \overset{n+1}{\downarrow} \\ 0, & \cdots, & 0, & 1, & 0, & \cdots, & 0 \end{array}\right)$. It follows that $w(\alpha^0 - \alpha_p^0\cdot\gamma) \leq w(\alpha') \leq w(\alpha^0)$. (For the dual problem, $\alpha'$ maximizes $w(\alpha)$ with a constraint that $\alpha_p = 0$, thus $w(\alpha')$ is no less than $w(\alpha^0 - \alpha_p^0\cdot\gamma)$, which is a special case that satisfies the constraints, including $\alpha_p = 0$. $\alpha^0$ maxmizes $w(\alpha)$ with a constraint $\alpha_p \geq 0$, which raises the constraint $\alpha_p = 0$, thus $w(\alpha') \leq w(\alpha^0)$. For the primal problem, the training problem corresponding to $\alpha'$ has less samples $(x_i, y_i)$, where $i \neq p$, to separate with maximum margin, thus its margin $m(\alpha')$ is no less than the margin $m(\alpha^0)$, and $w(\alpha') \leq w(\alpha^0)$. On the other hand, the hyperplane determined by $\alpha^0 - \alpha_p^0 \cdot \gamma$ might not separate $(x_i, y_i)$ for $i \neq p$ and corresponds to a equivalent or larger "margin" $1/\|\psi(\alpha^0 - \alpha_p^0 \cdot \gamma)\|$ than $m(\alpha')$).

Let us consider the inequality

$$\max_t w(\alpha' + t \cdot \gamma) - w(\alpha') \leq w(\alpha^0) - w(\alpha') \leq w(\alpha^0) - w(\alpha^0 - \alpha_p^0 \cdot \gamma).$$

For the left hand side, we have

$$\begin{aligned} w(\alpha' + t\gamma) &= \sum \alpha_i' + t - \frac{1}{2}\left\|\sum \alpha_i' y_i x_i + t \cdot y_p x_p\right\|^2 \\ &= \sum \alpha_i' + t - \frac{1}{2}\left\|\sum \alpha_i' y_i x_i\right\|^2 - t\left(\sum \alpha_i' y_i x_i\right) * (y_p x_p) - \frac{t^2}{2}\left\|y_p x_p\right\|^2 \\ &= w(\alpha') + t \cdot (1 - y_p \cdot \underbrace{(\sum \alpha_i' y_i x_i)}_{\psi'} * x_p) - \frac{t^2}{2}\left\|x_p\right\|^2 \end{aligned}$$

and $w(\alpha' + t\gamma) - w(\alpha') = t \cdot (1 - y_p \cdot \psi' * x_p) - \frac{t^2}{2}\left\|x_p\right\|^2$. Maximizing the expression over $t$, we find $t = (1 - y_p \cdot \psi' * x_p)/\|x_p\|^2$, and

$$\max_t w(\alpha' + t\gamma) - w(\alpha') = \frac{1}{2}\frac{(1 - y_p \cdot \psi' * x_p)^2}{\|x_p\|^2}.$$

For the right hand side,

$$\begin{aligned} w(\alpha^0 - \alpha_p^0 \cdot \gamma) &= \sum \alpha_i^0 - \alpha_p^0 - \frac{1}{2}\|\underbrace{\sum \alpha_i^0 y_i x_i}_{\psi_0} - \alpha_p^0 y_p x_p\|^2 \\ &= \sum \alpha_i^0 - \alpha_p^0 - \frac{1}{2}\|\psi_0\|^2 + \alpha_p^0 y_p \psi_0 * x_p - \frac{1}{2}\left(\alpha_p^0\right)^2\|x_p\|^2 \\ &= w(\alpha_0) - \alpha_p^0(1 - y_p \cdot \psi_0 * x_p) - \frac{1}{2}\left(\alpha_p^0\right)^2\|x_p\|^2 \\ &= w(\alpha_0) - \frac{1}{2}\left(\alpha_p^0\right)^2\|x_p\|^2. \end{aligned}$$

The last step above is due to the fact that $(x_p, y_p)$ is a support vector, and $y_p \cdot \psi_0 * x_p = 1$. Thus $w(\alpha^0) - w(\alpha^0 - \alpha_p^0 \cdot \gamma) = \frac{1}{2}\left(\alpha_p^0\right)^2\|x_p\|^2$ and $\frac{1}{2}\frac{\left(1 - y_p \cdot \psi' * x_p\right)^2}{\|x_p\|^2} \leq \frac{1}{2}\left(\alpha_p^0\right)^2\|x_p\|^2$. Thus

$$\begin{aligned} \alpha_p^0 &\geq \frac{|1 - y_p \cdot \psi' * x_p|}{\|x_p\|^2} \\ &\geq \frac{1}{D^2}. \end{aligned}$$

The last step above is due to the fact that the support vector classifier associated with $\psi'$ misclassifies $(x_p, y_p)$ according to assumption, and $y_p \cdot \psi' * x_p \leq 0$, and the fact that $\|x_p\| \leq D$. □

For a fixed $f \in \mathcal{F}$, if we observe $\frac{1}{n}\sum_{i=1}^n I\left(f(X_i) \neq Y_i\right)$ is small, can we say that $\mathbb{P}\left(f(X) \neq Y\right)$ is small? By the Law of Large Numbers,

$$\frac{1}{n}\sum_{i=1}^n I\left(f(X_i) \neq Y_i\right) \to \mathbb{E}I(f(X) \neq Y) = \mathbb{P}\left(f(X) \neq Y\right).$$

The Central Limit Theorem says

$$\frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n I\left(f(X_i) \neq Y_i\right) - \mathbb{E}I(f(X) \neq Y)\right)}{\sqrt{\text{Var}I}} \to \mathcal{N}(0,1).$$

Thus,

$$\frac{1}{n}\sum_{i=1}^n I\left(f(X_i) \neq Y_i\right) - \mathbb{E}I(f(X) \neq Y) \sim \frac{k}{\sqrt{n}}.$$

Let $Z_1, \cdots, Z_n \in \mathbb{R}$ be i.i.d. random variables. We're interested in bounds on $\frac{1}{n}\sum Z_i - \mathbb{E}Z$.

(1) Jensen's inequality: If $\phi$ is a convex function, then $\phi(\mathbb{E}Z) \leq \mathbb{E}\phi(X)$.

(2) Chebyshev's inequality: If $Z \geq 0$, then $\mathbb{P}\left(Z \geq t\right) \leq \frac{\mathbb{E}Z}{t}$.

Proof:

$$\begin{aligned}\mathbb{E}Z &= \mathbb{E}ZI(Z < t) + \mathbb{E}ZI(Z \geq t) \geq \mathbb{E}ZI(Z \geq t)\\ &\geq \mathbb{E}tI(Z \geq t) = t\mathbb{P}\left(Z \geq t\right).\end{aligned}$$

(3) Markov's inequality: Let $Z$ be a signed r.v. Then for any $\lambda > 0$

$$\mathbb{P}\left(Z \geq t\right) = \mathbb{P}\left(e^{\lambda Z} \geq e^{\lambda t}\right) \leq \frac{\mathbb{E}e^{\lambda Z}}{e^{\lambda t}}$$

and therefore

$$\mathbb{P}\left(Z \geq t\right) \leq \inf_{\lambda > 0} e^{-\lambda t}\mathbb{E}e^{\lambda Z}.$$

**Theorem 5.1.** *[Bennett] Assume $\mathbb{E}Z = 0$, $\mathbb{E}Z^2 = \sigma^2$, $|Z| < M = const$, $Z_1, \cdots, Z_n$ independent copies of $Z$, and $t \geq 0$. Then*

$$\mathbb{P}\left(\sum_{i=1}^n Z_i \geq t\right) \leq \exp\left(-\frac{n\sigma^2}{M^2}\phi\left(\frac{tM}{n\sigma^2}\right)\right),$$

*where $\phi(x) = (1+x)\log(1+x) - x$.*

*Proof.* Since $Z_i$ are i.i.d.,

$$\mathbb{P}\left(\sum_{i=1}^n Z_i \geq t\right) \leq e^{-\lambda t}\mathbb{E}e^{\lambda\sum_{i=1}^n Z_i} = e^{-\lambda t}\prod_{i=1}^n \mathbb{E}e^{\lambda Z_i} = e^{-\lambda t}\left(\mathbb{E}e^{\lambda Z}\right)^n.$$

Expanding,

$$\begin{aligned}
\mathbb{E}e^{\lambda Z} &= \mathbb{E}\sum_{k=0}^{\infty}\frac{(\lambda Z)^k}{k!} = \sum_{k=0}^{\infty}\lambda^k\frac{\mathbb{E}Z^k}{k!}\\
&= 1+\sum_{k=2}^{\infty}\frac{\lambda^k}{k!}\mathbb{E}Z^2Z^{k-2} \le 1+\sum_{k=2}^{\infty}\frac{\lambda^k}{k!}M^{k-2}\sigma^2\\
&= 1+\frac{\sigma^2}{M^2}\sum_{k=2}^{\infty}\frac{\lambda^k M^k}{k!} = 1+\frac{\sigma^2}{M^2}\left(e^{\lambda M}-1-\lambda M\right)\\
&\le \exp\left(\frac{\sigma^2}{M^2}\left(e^{\lambda M}-1-\lambda M\right)\right)
\end{aligned}$$

where the last inequality follows because $1+x \le e^x$.

Combining the results,

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{n} Z_i \ge t\right) &\le e^{-\lambda t}\exp\left(\frac{n\sigma^2}{M^2}\left(e^{\lambda M}-1-\lambda M\right)\right)\\
&= \exp\left(-\lambda t+\frac{n\sigma^2}{M^2}\left(e^{\lambda M}-1-\lambda M\right)\right)
\end{aligned}$$

Now, minimize the above bound with respect to $\lambda$. Taking derivative w.r.t. $\lambda$ and setting it to zero:

$$-t+\frac{n\sigma^2}{M^2}\left(Me^{\lambda M}-M\right)=0$$

$$e^{\lambda M}=\frac{tM}{n\sigma^2}+1$$

$$\lambda=\frac{1}{M}\log\left(1+\frac{tM}{n\sigma^2}\right).$$

The bound becomes

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{n} Z_i \ge t\right) &\le \exp\left(-\frac{t}{M}\log\left(1+\frac{tM}{n\sigma^2}\right)+\frac{n\sigma^2}{M^2}\left(\frac{tM}{n\sigma^2}+1-\log\left(1+\frac{tM}{n\sigma^2}\right)\right)\right)\\
&= \exp\left(\frac{n\sigma^2}{M^2}\left(\frac{tM}{n\sigma^2}-\log\left(1+\frac{tM}{n\sigma^2}\right)-\frac{tM}{n\sigma^2}\log\left(1+\frac{tM}{n\sigma^2}\right)\right)\right)\\
&= \exp\left(\frac{n\sigma^2}{M^2}\left(\frac{tM}{n\sigma^2}-\left(1+\frac{tM}{n\sigma^2}\right)\log\left(1+\frac{tM}{n\sigma^2}\right)\right)\right)\\
&= \exp\left(-\frac{n\sigma^2}{M^2}\phi\left(\frac{tM}{n\sigma^2}\right)\right)
\end{aligned}$$

□

Last time we proved Bennett's inequality: $\mathbb{E}X = 0$, $\mathbb{E}X^2 = \sigma^2$, $|X| < M = const$, $X_1, \cdots, X_n$ independent copies of $X$, and $t \geq 0$. Then

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) \leq \exp\left(-\frac{n\sigma^2}{M^2}\phi\left(\frac{tM}{n\sigma^2}\right)\right),$$

where $\phi(x) = (1+x)\log(1+x) - x$.

If $X$ is small, $\phi(x) = (1+x)(x - \frac{x^2}{2} + \cdots) - x = x + x^2 - \frac{x^2}{2} - x + \cdots = \frac{x^2}{2} + \cdots$.

If $X$ is large, $\phi(x) \sim x\log x$.

We can weaken the bound by decreasing $\phi(x)$. Take[1] $\phi(x) = \frac{x^2}{2+\frac{2}{3}x}$ to obtain **Bernstein's inequality**:

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) &\leq \exp\left(-\frac{n\sigma^2}{M^2}\left(\frac{\left(\frac{tM}{n\sigma^2}\right)^2}{2+\frac{2}{3}\frac{tM}{n\sigma^2}}\right)\right)\\ &= \exp\left(-\frac{t^2}{2n\sigma^2+\frac{2}{3}tM}\right)\\ &= e^{-u}\end{aligned}$$

where $u = \frac{t^2}{2n\sigma^2+\frac{2}{3}tM}$. Solve for $t$:

$$t^2 - \frac{2}{3}uMt - 2n\sigma^2 u = 0$$

$$t = \frac{1}{3}uM + \sqrt{\frac{u^2M^2}{9} + 2n\sigma^2 u}.$$

Substituting,

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq \sqrt{\frac{u^2M^2}{9} + 2n\sigma^2 u} + \frac{uM}{3}\right) \leq e^{-u}$$

or

$$\mathbb{P}\left(\sum_{i=1}^n X_i \leq \sqrt{\frac{u^2M^2}{9} + 2n\sigma^2 u} + \frac{uM}{3}\right) \geq 1 - e^{-u}$$

Using inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$,

$$\mathbb{P}\left(\sum_{i=1}^n X_i \leq \sqrt{2n\sigma^2 u} + \frac{2uM}{3}\right) \geq 1 - e^{-u}$$

For non-centered $X_i$, replace $X_i$ with $X_i - \mathbb{E}X$ or $\mathbb{E}X - X_i$. Then $|X_i - \mathbb{E}X| \leq 2M$ and so with high probability

$$\sum(X_i - \mathbb{E}X) \leq \sqrt{2n\sigma^2 u} + \frac{4uM}{3}.$$

Normalizing by $n$,

$$\frac{1}{n}\sum X_i - \mathbb{E}X \leq \sqrt{\frac{2\sigma^2 u}{n}} + \frac{4uM}{3n}$$

and

$$\mathbb{E}X - \frac{1}{n}\sum X_i \leq \sqrt{\frac{2\sigma^2 u}{n}} + \frac{4uM}{3n}.$$

[1] exercise: show that this is the best approximation

Whenever $\sqrt{\frac{2\sigma^2 u}{n}} \geq \frac{4uM}{3n}$, we have $u \leq \frac{n\sigma^2}{8M^2}$. So, $|\frac{1}{n}\sum X_i - \mathbb{E}X| \lesssim \sqrt{\frac{2\sigma^2 u}{n}}$ for $u \lesssim n\sigma^2$ (range of normal deviations). This is predicted by the Central Limit Theorem (condition for CLT is $n\sigma^2 \to \infty$). If $n\sigma^2$ does not go to infinity, we get Poisson behavior.

Recall from the last lecture that the we're interested in concentration inequalities because we want to know $\mathbb{P}\left(f(X) \neq Y\right)$ while we only observe $\frac{1}{n}\sum_{i=1}^n I\left(f(X_i) \neq Y_i\right)$. In Bernstein's inequality take $''X_i''$ to be $I(f(X_i) \neq Y_i)$. Then, since $2M = 1$, we get

$$\mathbb{E}I(f(X_i) \neq Y_i) - \frac{1}{n}\sum_{i=1}^n I\left(f(X_i) \neq Y_i\right) \leq \sqrt{\frac{2\mathbb{P}\left(f(X_i) \neq Y_i\right)\left(1 - \mathbb{P}\left(f(X_i) \neq Y_i\right)\right)u}{n}} + \frac{2u}{3n}$$

because $\mathbb{E}I(f(X_i) \neq Y_i) = \mathbb{P}\left(f(X_i) \neq Y_i\right) = \mathbb{E}I^2$ and therefore $\mathrm{Var}(I) = \sigma^2 = \mathbb{E}I^2 - (\mathbb{E}I)^2$. Thus,

$$\mathbb{P}\left(f(X_i) \neq Y_i\right) \leq \frac{1}{n}\sum_{i=1}^n I\left(f(X_i) \neq Y_i\right) + \sqrt{\frac{2\mathbb{P}\left(f(X_i) \neq Y_i\right)u}{n}} + \frac{2u}{3n}$$

with probability at least $1 - e^{-u}$. When the training error is zero,

$$\mathbb{P}\left(f(X_i) \neq Y_i\right) \leq \sqrt{\frac{2\mathbb{P}\left(f(X_i) \neq Y_i\right)u}{n}} + \frac{2u}{3n}.$$

If we forget about $2u/3n$ for a second, we obtain $\mathbb{P}\left(f(X_i) \neq Y_i\right)^2 \leq 2\mathbb{P}\left(f(X_i) \neq Y_i\right)u/n$ and hence

$$\mathbb{P}\left(f(X_i) \neq Y_i\right) \leq \frac{2u}{n}.$$

The above *zero-error rate* is better than $n^{-1/2}$ predicted by CLT.

Let $a_1, \ldots, a_n \in \mathbb{R}$ and let $\varepsilon_1, \ldots, \varepsilon_n$ be i.i.d. Rademacher random variables: $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 0.5$.

**Theorem 7.1.** *[Hoeffding] For $t \geq 0$,*

$$\mathbb{P}\left(\sum_{i=1}^n \varepsilon_i a_i \geq t\right) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n a_i^2}\right).$$

*Proof.* Similarly to the proof of Bennett's inequality (Lecture 5),

$$\mathbb{P}\left(\sum_{i=1}^n \varepsilon_i a_i \geq t\right) \leq e^{-\lambda t}\mathbb{E}\exp\left(\lambda\sum_{i=1}^n \varepsilon_i a_i\right) = e^{-\lambda t}\prod_{i=1}^n \mathbb{E}\exp(\lambda\varepsilon_i a_i).$$

Using inequality $\frac{e^x+e^{-x}}{2} \leq e^{x^2/2}$ (from Taylor expansion), we get

$$\mathbb{E}\exp(\lambda\varepsilon_i a_i) = \frac{1}{2}e^{\lambda a_i} + \frac{1}{2}e^{-\lambda a_i} \leq e^{\frac{\lambda^2 a_i^2}{2}}.$$

Hence, we need to minimize the bound with respect to $\lambda > 0$:

$$\mathbb{P}\left(\sum_{i=1}^n \varepsilon_i a_i \geq t\right) \leq e^{-\lambda t}e^{\frac{\lambda^2}{2}\sum_{i=1}^n a_i^2}.$$

Setting derivative to zero, we obtain the result. □

Now we change variable: $u = \frac{t^2}{2\sum_{i=1}^n a_i^2}$. Then $t = \sqrt{2u\sum_{i=1}^n a_i^2}$.

$$\mathbb{P}\left(\sum_{i=1}^n \varepsilon_i a_i \geq \sqrt{2u\sum_{i=1}^n a_i^2}\right) \leq e^{-u}$$

and

$$\mathbb{P}\left(\sum_{i=1}^n \varepsilon_i a_i \leq \sqrt{2u\sum_{i=1}^n a_i^2}\right) \geq 1 - e^{-u}.$$

Here $\sum_{i=1}^n a_i^2 = \mathrm{Var}(\sum_{i=1}^n \varepsilon_i a_i)$.

Rademacher sums will play important role in future. Consider again the problem of estimating $\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f$. We will see that by the Symmetrization technique,

$$\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f \sim \frac{1}{n}\sum_{i=1}^n f(X_i) - \frac{1}{n}\sum_{i=1}^n f(X_i').$$

In fact,

$$\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f\right| \leq \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \frac{1}{n}\sum_{i=1}^n f(X_i')\right| \leq 2\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f\right|.$$

The second inequality above follows by adding and subtracting $\mathbb{E}f$:

$$\begin{aligned}\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \frac{1}{n}\sum_{i=1}^n f(X_i')\right| &\leq \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f\right| + \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n f(X_i') - \mathbb{E}f\right| \\ &= 2\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f\right|\end{aligned}$$

while for the first inequality we use Jensen's inequality:

$$\begin{aligned}\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f\right| &= \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}f(X_i')\right| \\ &\leq \mathbb{E}_X\mathbb{E}_{X'}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}f(X_i')\right|.\end{aligned}$$

Note that $\frac{1}{n}\sum_{i=1}^n f(X_i) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}f(X_i')$ is equal in distribution to $\frac{1}{n}\sum_{i=1}^n \varepsilon_i(f(X_i) - f(X_i'))$.

We now prove Hoeffding-Chernoff Inequality:

**Theorem 7.2.** *Assume* $0 \leq X_i \leq 1$ *and* $\mu = \mathbb{E}X$. *Then*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu \geq t\right) \leq e^{-n\mathcal{D}(\mu+t,\mu)}$$

*where the KL-divergence* $\mathcal{D}(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$.

*Proof.* Note that $\phi(x) = e^{\lambda x}$ is convex and so $e^{\lambda x} = e^{\lambda(x\cdot 1 + (1-x)\cdot 0)} \leq xe^{\lambda} + (1-x)e^{\lambda\cdot 0} = 1 - x + xe^{\lambda}$. Hence,

$$\mathbb{E}e^{\lambda X} = 1 - \mathbb{E}X + \mathbb{E}Xe^{\lambda} = 1 - \mu + \mu e^{\lambda}.$$

Again, we minimize the following bound with respect to $\lambda > 0$:

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^n X_i \geq n(\mu+t)\right) &\leq e^{-\lambda n(\mu+t)}\mathbb{E}e^{\lambda\sum X_i} \\ &= e^{-\lambda n(\mu+t)}\left(\mathbb{E}e^{\lambda X}\right)^n \\ &\leq e^{-\lambda n(\mu+t)}\left(1-\mu+\mu e^{\lambda}\right)^n\end{aligned}$$

Take derivative w.r.t. $\lambda$:

$$-n(\mu+t)e^{-\lambda n(\mu+t)}(1-\mu+\mu e^{\lambda})^n + n(1-\mu+\mu e^{\lambda})^{n-1}\mu e^{\lambda}e^{-\lambda n(\mu+t)} = 0$$

$$-(\mu+t)(1-\mu+\mu e^{\lambda}) + \mu e^{\lambda} = 0$$

$$e^{\lambda} = \frac{(1-\mu)(\mu+t)}{\mu(1-\mu-t)}.$$

Substituting,

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^n X_i \geq n(\mu+t)\right) &\leq \left(\left(\frac{\mu(1-\mu-t)}{(1-\mu)(\mu+t)}\right)^{\mu+t}\left(1-\mu+\frac{(1-\mu)(\mu+t)}{1-\mu-t}\right)\right)^n \\ &= \left(\left(\frac{\mu}{\mu+t}\right)^{\mu+t}\left(\frac{1-\mu}{1-\mu-t}\right)^{1-\mu-t}\right)^n \\ &= \exp\left(-n\left((\mu+t)\log\frac{\mu+t}{\mu} + (1-\mu-t)\log\frac{1-\mu-t}{1-\mu}\right)\right),\end{aligned}$$

completing the proof. Moreover,

$$\mathbb{P}\left(\mu - \frac{1}{n}\sum_{i=1}^n X_i \geq t\right) = \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n Z_i - \mu_Z \geq t\right) \leq e^{-n\mathcal{D}(\mu_z+t,\mu_Z)} = e^{-n\mathcal{D}(1-\mu_X+t,1-\mu_X)}$$

where $Z_i = 1 - X_i$ (and thus $\mu_Z = 1 - \mu_X$). □

If $0 < \mu \leq 1/2$,

$$\mathcal{D}(1-\mu+t, 1-\mu) \geq \frac{t^2}{2\mu(1-\mu)}.$$

Hence, we get

$$\mathbb{P}\left(\mu - \frac{1}{n}\sum_{i=1}^n X_i \geq t\right) \leq e^{-\frac{nt^2}{2\mu(1-\mu)}} = e^{-u}.$$

Solving for $t$,

$$\mathbb{P}\left(\mu - \frac{1}{n}\sum_{i=1}^n X_i \geq \sqrt{\frac{2\mu(1-\mu)u}{n}}\right) \leq e^{-u}.$$

If $X_i \in \{0,1\}$ are i.i.d. Bernoulli trials, then $\mu = \mathbb{E}X = \mathbb{P}(X=1)$, $\mathrm{Var}(X) = \mu(1-\mu)$, and $\mathbb{P}\left(\mu - \frac{1}{n}\sum_{i=1}^n X_i \geq t\right) \leq e^{-\frac{nt^2}{2\mathrm{Var}(X)}}$.

The following inequality says that if we pick $n$ reals $a_1, \cdots, a_n \in \mathbb{R}$ and add them up each multiplied by a random sign $\pm 1$, then the expected value of the sum should not be far off from $\sqrt{\sum |a_i|^2}$.

**Theorem 7.3.** *[Khinchine inequality] Let $a_1, \cdots, a_n \in \mathbb{R}$, $\epsilon_i, \cdots, \epsilon_n$ be i.i.d. Rademacher random variables: $\mathbb{P}(\epsilon_i = 1) = \mathbb{P}(\epsilon_i = -1) = 0.5$, and $0 < p < \infty$. Then*

$$A_p \cdot \left(\sum_{i=1}^n |a_i|^2\right)^{1/2} \leq \left(\mathbb{E}\left|\sum_{i=1}^n a_i\epsilon_i\right|^p\right)^{1/p} \leq B_p \cdot \left(\sum_{i=1}^n |a_i|^2\right)^{1/2}$$

*for some constants $A_p$ and $B_p$ depending on $p$.*

*Proof.* Let $\sum |a_i|^2 = 1$ without lossing generality. Then

$$\begin{aligned}
\mathbb{E}\left|\sum a_i\epsilon_i\right|^p &= \int_0^\infty \mathbb{P}\left(\left|\sum a_i\epsilon_i\right|^p \geq s^p\right)\, ds^p \\
&= \int_0^\infty \mathbb{P}\left(\left|\sum a_i\epsilon_i\right| \geq s\right) \cdot ps^{p-1}\, ds^p \\
&= \int_0^\infty \mathbb{P}\left(\left|\sum a_i\epsilon_i\right| \geq s\right) \cdot ps^{p-1}\, ds^p \\
&\leq \int_0^\infty 2\exp(-\frac{s^2}{2}) \cdot ps^{p-1}\, ds^p \text{ , Hoeffding's inequality} \\
&= (B_p)^p \text{ , when } p \geq 2.
\end{aligned}$$

When $0 < p < 2$,

$$\begin{aligned}
\mathbb{E}\left|\sum a_i\epsilon_i\right|^p &\leq \mathbb{E}\left|\sum a_i\epsilon_i\right|^2 \\
&= \mathbb{E}\left|\sum a_i\epsilon_i\right|^{\frac{2}{3}p+(2-\frac{2}{3}p)} \\
&\leq \left(\mathbb{E}\left|\sum a_i\epsilon_i\right|^p\right)^{\frac{2}{3}}\left(\mathbb{E}\left|\sum a_i\epsilon_i\right|^{6-2p}\right)^{\frac{1}{3}} \text{ , Holder's inequality} \\
&\leq (B_{6-2p})^{2-\frac{2}{3}p}\cdot\left(\mathbb{E}\left|\sum a_i\epsilon_i\right|^p\right)^{\frac{2}{3}}.
\end{aligned}$$

Thus $\mathbb{E}\left|\sum a_i\epsilon_i\right|^p \leq (B_{6-2p})^{6-2p}$, completing the proof. □

Assume $f \in \mathcal{F} = \{f : \mathcal{X} \mapsto \mathbb{R}\}$ and $x_1, \ldots, x_n$ are i.i.d. Denote $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(x_i)$ and $\mathbb{P}f = \int f dP = \mathbb{E}f$. We are interested in bounding $\frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}f$.

Worst-case scenario is the value

$$\sup_{f\in\mathcal{F}} |\mathbb{P}_n f - \mathbb{P}f| .$$

The Glivenko-Cantelli property $GC(\mathcal{F}, P)$ says that

$$\mathbb{E}\sup_{f\in\mathcal{F}} |\mathbb{P}_n f - \mathbb{P}f| \to 0$$

as $n \to \infty$.

- Algorithm can output any $f \in \mathcal{F}$
- Objective is determined by $\mathbb{P}_n f$ (on the data)
- Goal is $\mathbb{P}f$
- Distribution $P$ is unknown

The most pessimistic requirement is

$$\sup_P \mathbb{E}\sup_{f\in\mathcal{F}} |\mathbb{P}_n f - \mathbb{P}f| \to 0$$

which we denote

$$\text{uniform}GC(\mathcal{F}).$$

**VC classes of sets**

Let $\mathcal{C} = \{C \subseteq X\}$, $f_C(x) = I(x \in C)$. The most pessimistic value is

$$\sup_P \mathbb{E}\sup_{C\in\mathcal{C}} |\mathbb{P}_n(C) - \mathbb{P}(C)| \to 0.$$

For any sample $\{x_1, \ldots, x_n\}$, we can look at the ways that $\mathcal{C}$ intersects with the sample:

$$\{C \cap \{x_1, \ldots, x_n\} : C \in \mathcal{C}\}.$$

Let

$$\triangle_n(\mathcal{C}, x_1, \ldots, x_n) = \text{card}\ \{C \cap \{x_1, \ldots, x_n\} : C \in \mathcal{C}\},$$

the number of different subsets picked out by $C \in \mathcal{C}$. Note that this number is at most $2^n$.

Denote

$$\triangle_n(\mathcal{C}) = \sup_{\{x_1,\ldots,x_n\}} \triangle_n(\mathcal{C}, x_1, \ldots, x_n) \leq 2^n.$$

We will see that for some classes, $\triangle_n(\mathcal{C}) = 2^n$ for $n \leq V$ and $\triangle_n(\mathcal{C}) < 2^n$ for $n > V$ for some constant $V$.

What if $\triangle_n(\mathcal{C}) = 2^n$ for all $n \geq 1$? That means we can always find $\{x_1, \ldots, x_n\}$ such that $C \in \mathcal{C}$ can pick out any subset of it: "$\mathcal{C}$ **shatters** $\{x_1, \ldots, x_n\}$". In some sense, we do not learn anything.

**Definition 8.1.** If $V < \infty$, then $\mathcal{C}$ is called a VC class. $V$ is called VC dimension of $\mathcal{C}$.

Sauer's lemma states the following:

**Lemma 8.2.**

$$\forall\{x_1,\ldots,x_n\},\qquad \triangle_n(\mathcal{C},x_1,\ldots,x_n)\le\left(\frac{en}{V}\right)^V \textit{ for } n\ge V.$$

Hence, $\mathcal{C}$ will pick out only very few subsets out of $2^n$ (because $\left(\frac{en}{V}\right)^V\sim n^V$).

**Lemma 8.3.** *The number $\triangle_n(\mathcal{C},x_1,\ldots,x_n)$ of subsets picked out by $\mathcal{C}$ is bounded by the number of subsets shattered by $\mathcal{C}$.*

*Proof.* Without loss of generality, we restrict $\mathcal{C}$ to $\mathcal{C}:=\{C\cap\{x_1,\ldots,x_n\}:C\in\mathcal{C}\}$, and we have $\text{card}(\mathcal{C})=\Delta_n(\mathcal{C},x_1,\cdots,x_n)$.

We will say that $\mathcal{C}$ is **hereditary** if and only if whenever $B\subseteq C\in\mathcal{C}$, $B\in\mathcal{C}$. If $\mathcal{C}$ is hereditary, then every $C\in\mathcal{C}$ is shattered by $\mathcal{C}$, and the lemma is obvious. Otherwise, we will transform $\mathcal{C}\to\mathcal{C}'$, hereditary, without changing the cardinality of $\mathcal{C}$ and without increasing the number of shattered subsets.

Define the operators $T_i$ for $i=1,\cdots,n$ as the following,

$$\begin{aligned} T_i(C) &= \begin{cases} C-\{x_i\} & \text{if } C-\{x_i\} \text{ is not in } \mathcal{C}\\ C & \text{otherwise}\end{cases}\\ T_i(\mathcal{C}) &= \{T_i(C):C\in\mathcal{C}\}.\end{aligned}$$

It follows that card $T_i(\mathcal{C})=$ card $\mathcal{C}$. Moreover, every $A\subseteq\{x_1,\cdots,x_n\}$ that is shattered by $T_i(\mathcal{C})$ is also shattered by $\mathcal{C}$. If $x_i\notin A$, then $\forall C\in\mathcal{C}, A\bigcap C=A\bigcap T_i(C)$, thus $\mathcal{C}$ and $T_i(\mathcal{C})$ both or neither shatter $A$. On the other hand, if $x_i\in A$ and $A$ is shattered by $T_i(\mathcal{C})$, then $\forall B\subseteq A,\exists C\in\mathcal{C}$, such that $B\bigcap\{x_i\}=A\bigcap T_i(C)$. This means that $x_i\in T_i(C)$, and that $C\backslash\{x_i\}\in\mathcal{C}$. Thus both $B\bigcup\{x_i\}$ and $B\backslash\{x_i\}$ are picked out by $\mathcal{C}$. Since either $B=B\bigcup\{x_i\}$ or $B=B\backslash\{x_i\}$, $B$ is picked out by $\mathcal{C}$. Thus $A$ is shattered by $\mathcal{C}$.

Apply the operator $T=T_1\circ\ldots\circ T_n$ until $T^{k+1}(\mathcal{C})=T^k(\mathcal{C})$. This will happen for at most $\sum_{C\in\mathcal{C}}\text{card}(C)$ times, since $\sum_{C\in\mathcal{C}}\text{card}(T_i(C))<\sum_{C\in\mathcal{C}}\text{card}(C)$ if $T_i(\mathcal{C})\neq\mathcal{C}$. The resulting collection $\mathcal{C}'$ is hereditary. This proves the lemma. □

Sauer's lemma is proved, since for arbitrary $\{x_1,\ldots,x_n\}$,

$$\begin{aligned}\triangle_n(\mathcal{C},x_1,\ldots,x_n) &\le \text{card (shattered subsets of }\{x_1,\ldots,x_n\})\\ &\le \text{card (subsets of size }\le V)\\ &=\sum_{i=0}^{V}\binom{n}{i}\\ &\le\left(\frac{en}{V}\right)^V.\end{aligned}$$

Recall the definition of VC-dimension. Consider some examples:

- $\mathcal{C} = \{(-\infty, a) \text{ and } (a, \infty) : a \in \mathbb{R}\}$. $VC(\mathcal{C}) = 2$.
- $\mathcal{C} = \{(a, b) \cup (c, d)\}$. $VC(\mathcal{C}) = 4$.
- $f_1, \ldots, f_d : \mathcal{X} \to \mathbb{R}$, $\mathcal{C} = \{\{x : \sum_{k=1}^d \alpha_k f_k(x) > 0\} : \alpha_1, \ldots, \alpha_d \in \mathbb{R}\}$

**Theorem 9.1.** *$VC(\mathcal{C})$ in the last example above is at most $d$.*

*Proof.* Observation: For any $\{x_1, \ldots, x_{d+1}\}$ if we cannot shatter $\{x_1, \ldots, x_{d+1}\} \longleftrightarrow \exists I \subseteq \{1 \ldots d+1\}$ s.t. we cannot pick out $\{x_i, i \in I\}$. If we can pick out $\{x_i, i \in I\}$, then for some $C \in \mathcal{C}$ there are $\alpha_1, \ldots, \alpha_d$ s.t. $\sum_{k=1}^d \alpha_k f_k(x) > 0$ for $i \in I$ and $\sum_{k=1}^d \alpha_k f_k(x) \leq 0$ for $i \notin I$.

Denote

$$\left(\sum_{k=1}^d \alpha_k f_k(x_1), \ldots, \sum_{k=1}^d \alpha_k f_k(x_{d+1})\right) = F(\alpha) \in \mathbb{R}^{d+1}.$$

By linearity,

$$F(\alpha) = \sum_{k=1}^d \alpha_k \left(f_k(x_1), \ldots, f(x_{d+1})\right) = \sum_{k=1}^d \alpha_k F_k \subseteq H \subset \mathbb{R}^{d+1}$$

and $H$ is a $d$-dim subspace. Hence, $\exists \phi \neq 0$, $\phi \cdot h = 0, \forall h \in H$ ($\phi$ orthogonal to $H$). Let $I = \{i : \phi_i > 0\}$, where $\phi = (\phi_1, \ldots, \phi_{d+1})$. If $I = \emptyset$ then take $-\phi$ instead of $\phi$ so that $\phi$ has positive coordinates.

Claim: We cannot pick out $\{x_i, i \in I\}$. Suppose we can: then $\exists \alpha_1, \ldots, \alpha_d$ s.t. $\sum_{k=1}^d \alpha_k f_k(x_i) > 0$ for $i \in I$ and $\sum_{k=1}^d \alpha_k f_k(x_i) \leq 0$ for $i \notin I$. But $\phi \cdot F(\alpha) = 0$ and so

$$\phi_1 \sum_{k=1}^d \alpha_k f_k(x_1) + \ldots + \phi_{d+1} \sum_{k=1}^d \alpha_k f_k(x_{d+1}) = 0.$$

Hence,

$$\underbrace{\sum_{i \in I} \phi_i \left(\sum_{k=1}^d \alpha_k f_k(x_i)\right)}_{>0} = \sum_{i \notin I} \underbrace{(-\phi_i)}_{\geq 0} \underbrace{\left(\sum_{k=1}^d \alpha_k f_k(x_i)\right)}_{\leq 0}.$$

Contradiction. □

- Half-spaces in $\mathbb{R}^d$: $\{\{\alpha_1 x_1 + \ldots + \alpha_d x_d + \alpha_{d+1} > 0\} : \alpha_1, \ldots, \alpha_{d+1} \in \mathbb{R}\}$.

By setting $f_1 = x_1, \ldots, f_d = x_d, f_{d+1} = 1$, we can use the previous result and therefore $VC(\mathcal{C}) \leq d+1$ for half-spaces.

Reminder: $\triangle_n(\mathcal{C}, x_1, \ldots, x_n) = \text{card}\{\{x_1, \ldots, x_n\} \cap C : C \in \mathcal{C}\}$.

**Lemma 9.1.** *If $\mathcal{C}$ and $\mathcal{D}$ are VC classes of sets,*

1. *$\mathcal{C} = \{C^c : C \in \mathcal{C}\}$ is VC*
2. *$\mathcal{C} \cap \mathcal{D} = \{C \cap D : C \in \mathcal{C}, D \in \mathcal{D}\}$ is VC*
3. *$\mathcal{C} \cup \mathcal{D} = \{C \cup D : C \in \mathcal{C}, D \in \mathcal{D}\}$ is VC*

(1) obvious - we can shatter $x_1, \ldots, x_n$ by $\mathcal{C}$ iff we can do the same by $\mathcal{C}^c$.

(a) By Sauer's Lemma,

$$\begin{aligned}\triangle_n(\mathcal{C} \cap \mathcal{D}, x_1, \ldots, x_n) &\leq \triangle_n(\mathcal{C}, x_1, \ldots, x_n)\triangle_n(\mathcal{C} \cap \mathcal{D}, x_1, \ldots, x_n) \\ &\leq \left(\frac{en}{V_{\mathcal{C}}}\right)^V_{\mathcal{C}} \left(\frac{en}{V_{\mathcal{D}}}\right)^V_{\mathcal{D}} \leq 2^n\end{aligned}$$

for large enough $n$.

(b) $(C \cup D) = (C^c \cap D^c)^c$, and the result follows from (1) and (2).

**Example 9.1.** Decision trees on $\mathbb{R}^d$ with linear decision rules: $\{C_1 \cap \ldots C_\ell\}$ is VC and $\bigcup_{\text{leaves}}\{C_1 \cap \ldots C_\ell\}$ is VC.

Neural networks with depth $\ell$ and binary leaves.

We are interested in bounding

$$\mathbb{P}\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i\in C)-\mathbb{P}\left(C\right)\right|\geq t\right)$$

In Lecture 7 we hinted at Symmetrization as a way to deal with the unknown $\mathbb{P}\left(C\right)$.

**Lemma 10.1.** *[Symmetrization] If* $t\geq\sqrt{\frac{2}{n}}$*, then*

$$\mathbb{P}\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i\in C)-\mathbb{P}\left(C\right)\right|\geq t\right)\leq 2\mathbb{P}\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i\in C)-\frac{1}{n}\sum_{i=1}^{n} I(X_i'\in C)\right|\geq t/2\right).$$

*Proof.* Suppose the event

$$\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i\in C)-\mathbb{P}\left(C\right)\right|\geq t$$

occurs. Let $X=(X_1,\ldots,X_n)\in\{\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i\in C)-\mathbb{P}\left(C\right)\right|\geq t\}$. Then

$$\exists C_X \quad \text{such that} \quad \left|\frac{1}{n}\sum_{i=1}^{n} I(X_i\in C_X)-\mathbb{P}\left(C_X\right)\right|\geq t.$$

For a fixed $C$,

$$\begin{aligned}\mathbb{P}_{X'}\left(\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i'\in C)-\mathbb{P}\left(C\right)\right|\geq t/2\right) &= \mathbb{P}\left(\left(\frac{1}{n}\sum_{i=1}^{n} I(X_i'\in C)-\mathbb{P}\left(C\right)\right)^2\geq t^2/4\right)\\ &\leq \text{ (by Chebyshev's Ineq) } \frac{4\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} I(X_i'\in C)-\mathbb{P}\left(C\right)\right)^2}{t^2}\\ &= \frac{4}{n^2t^2}\sum_{i,j}\mathbb{E}(I(X_i'\in C)-\mathbb{P}\left(C\right))(I(X_j'\in C)-\mathbb{P}\left(C\right))\\ = \frac{4}{n^2t^2}\sum_{i=1}^{n}\mathbb{E}(I(X_i'\in C)-\mathbb{P}\left(C\right))^2 &= \frac{4n\mathbb{P}\left(C\right)(1-\mathbb{P}\left(C\right)}{n^2t^2}\leq\frac{1}{nt^2}\leq\frac{1}{2}\end{aligned}$$

since we chose $t\geq\sqrt{\frac{2}{n}}$.

So,

$$\mathbb{P}_{X'}\left(\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i'\in C_X)-\mathbb{P}\left(C_X\right)\right|\leq t/2\middle|\exists C_X\right)\geq 1/2$$

if $t\geq\sqrt{2/n}$. Assume that the event

$$\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i'\in C_X)-\mathbb{P}\left(C_X\right)\right|\leq t/2$$

occurs. Recall that

$$\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i\in C_X)-\mathbb{P}\left(C_X\right)\right|\geq t.$$

Hence, it must be that

$$\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i\in C_X)-\frac{1}{n}\sum_{i=1}^{n} I(X_i'\in C_X)\right|\geq t/2.$$

We conclude

$$\begin{aligned}\frac{1}{2} &\leq \mathbb{P}_{X'}\left(\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i' \in C_X) - \mathbb{P}\left(C_X\right)\right| \leq t/2 \middle| \exists C_X\right) \\ &\leq \mathbb{P}_{X'}\left(\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i \in C_X) - \frac{1}{n}\sum_{i=1}^{n} I(X_i' \in C_X)\right| \geq t/2 \middle| \exists C_X\right) \\ &\quad \mathbb{P}_{X'}\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i \in C) - \frac{1}{n}\sum_{i=1}^{n} I(X_i' \in C)\right| \geq t/2 \middle| \exists C_X\right).\end{aligned}$$

Since indicators are $0,1$-valued,

$$\frac{1}{2} I\left(\underbrace{\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i \in C) - \mathbb{P}\left(C\right)\right| \geq t}_{\exists C_X}\right)$$

$$\leq \mathbb{P}_{X'}\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i \in C) - \frac{1}{n}\sum_{i=1}^{n} I(X_i' \in C)\right| \geq t/2 \middle| \exists C_X\right) \cdot I\left(\exists C_X\right)$$

$$\leq \mathbb{P}_{X,X'}\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i \in C) - \frac{1}{n}\sum_{i=1}^{n} I(X_i' \in C)\right| \geq t/2\right).$$

Now, take expectation with respect to $X_i$'s to obtain

$$\mathbb{P}_{X}\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i \in C) - \mathbb{P}\left(C\right)\right| \geq t\right)$$

$$\leq 2\cdot\mathbb{P}_{X,X'}\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i \in C) - \frac{1}{n}\sum_{i=1}^{n} I(X_i' \in C)\right| \geq t/2\right).$$

□

**Theorem 10.1.** *If $VC(\mathcal{C}) = V$, then*

$$\mathbb{P}\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i \in C) - \mathbb{P}\left(C\right)\right| \geq t\right) \leq 4\left(\frac{2en}{V}\right)^{V} e^{-\frac{nt^2}{8}}.$$

*Proof.*

$$2\mathbb{P}\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i \in C) - \frac{1}{n}\sum_{i=1}^{n} I(X_i' \in C)\right| \geq t/2\right)$$

$$= 2\mathbb{P}\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i\left(I(X_i \in C) - I(X_i' \in C)\right)\right| \geq t/2\right)$$

$$= 2\mathbb{E}_{X,X'}\mathbb{P}_{\varepsilon}\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i\left(I(X_i \in C) - I(X_i' \in C)\right)\right| \geq t/2\right).$$

The first equality is due to the fact that $X_i$ and $X_i'$ are i.i.d., and so switching their names (i.e. introducing random signs $\varepsilon_i$, $\mathbb{P}(\varepsilon_i = \pm 1) = 1/2$) does not have any effect. In the last line, it's important to see that the probability is taken with respect to $\varepsilon_i$'s, while $X_i$ and $X_i'$'s are fixed.

By Sauer's lemma,

$$\triangle_{2n}(\mathcal{C}, X_1, \ldots, X_n, X_1', \ldots, X_n') \le \left(\frac{2en}{V}\right)^V.$$

In other words, any class will be equivalent to one of $C_1, \ldots, C_N$ on the data, where $N \le \left(\frac{2en}{V}\right)^V$. Hence,

$$\begin{aligned}
&2\mathbb{E}_{X,X'}\mathbb{P}_\varepsilon\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(I(X_i\in C) - I(X_i'\in C)\right)\right| \ge t/2\right) \\
&= 2\mathbb{E}_{X,X'}\mathbb{P}_\varepsilon\left(\sup_{1\le k\le N}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(I(X_i\in C_k) - I(X_i'\in C_k)\right)\right| \ge t/2\right) \\
&= 2\mathbb{E}_{X,X'}\mathbb{P}_\varepsilon\left(\bigcup_{k=1}^N\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(I(X_i\in C_k) - I(X_i'\in C_k)\right)\right| \ge t/2\right) \\
&\overset{\text{union bound}}{\le} 2\mathbb{E}\sum_{k=1}^N \mathbb{P}_\varepsilon\left(\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(I(X_i\in C_k) - I(X_i'\in C_k)\right)\right| \ge t/2\right) \\
&\overset{\text{Hoeffding's inequality}}{\le} 2\mathbb{E}\sum_{k=1}^N 2\exp\left(-\frac{-n^2t^2}{8\sum_{i=1}^n\left(I(X_i\in C) - I(X_i'\in C)\right)^2}\right) \\
&\le 2\mathbb{E}\sum_{k=1}^N 2\exp\left(-\frac{-n^2t^2}{8n}\right) \le 2\left(\frac{2en}{V}\right)^V 2e^{-\frac{nt^2}{8}}.
\end{aligned}$$

□

Last time we proved the Pessimistic VC inequality:

$$\mathbb{P}\left(\sup_C \left|\frac{1}{n}\sum_{i=1}^n I(X_i \in C) - \mathbb{P}(C)\right| \geq t\right) \leq 4\left(\frac{2en}{V}\right)^V e^{-\frac{nt^2}{8}},$$

which can be rewritten with

$$t = \sqrt{\frac{8}{n}\left(\log 4 + V\log\frac{2en}{V} + u\right)}$$

as

$$\mathbb{P}\left(\sup_C \left|\frac{1}{n}\sum_{i=1}^n I(X_i \in C) - \mathbb{P}(C)\right| \leq \sqrt{\frac{8}{n}\left(\log 4 + V\log\frac{2en}{V} + u\right)}\right) \geq 1 - e^{-u}.$$

Hence, the rate is $\sqrt{\frac{V\log n}{n}}$. In this lecture we will prove Optimistic VC inequality, which will improve on this rate when $\mathbb{P}(C)$ is small.

As before, we have pairs $(X_i, Y_i)$, $Y_i = \pm 1$. These examples are labeled according to some unknown $C_0$ such that $Y = 1$ if $X = C_0$ and $Y = 0$ if $X \notin C_0$.

Let $\mathcal{C} = \{C : C \subseteq \mathcal{X}\}$, a set of classifiers. $C$ makes a mistake if

$$X \in C \setminus C_0 \cup C_0 \setminus C = C \triangle C_0.$$

Similarly to last lecture, we can derive bounds on

$$\sup_C \left|\frac{1}{n}\sum_{i=1}^n I(X_i \in C\triangle C_0) - \mathbb{P}(C\triangle C_0)\right|,$$

where $\mathbb{P}(C\triangle C_0)$ is the generalization error.

Let $\mathcal{C}' = \{C\triangle C_0 : C \in \mathcal{C}\}$. One can prove that $VC(\mathcal{C}') \leq VC(\mathcal{C})$ and $\triangle_n(C', X_1, \ldots, X_n) \leq \triangle_n(C, X_1, \ldots, X_n)$.

By Hoeffding-Chernoff, if $\mathbb{P}(C) \leq \frac{1}{2}$,

$$\mathbb{P}\left(\mathbb{P}(C) - \frac{1}{n}\sum_{i=1}^n I(X_i \in C) \leq \sqrt{\frac{2\mathbb{P}(C)\, t}{n}}\right) \geq 1 - e^{-t}.$$

**Theorem 11.1.** *[Optimistic VC inequality]*

$$\mathbb{P}\left(\sup_C \frac{\mathbb{P}(C) - \frac{1}{n}\sum_{i=1}^n I(X_i \in C)}{\sqrt{\mathbb{P}(C)}} \geq t\right) \leq 4\left(\frac{2en}{V}\right)^V e^{-\frac{nt^2}{4}}.$$

*Proof.* Let $C$ be fixed. Then

$$\mathbb{P}_{(X_i')}\left(\frac{1}{n}\sum_{i=1}^n I(X_i' \in C) \geq \mathbb{P}(C)\right) \geq \frac{1}{4}$$

whenever $\mathbb{P}(C) \geq \frac{1}{n}$. Indeed, $\mathbb{P}(C) \geq \frac{1}{n}$ since $\sum_{i=1}^n I(X_i' \in C) \geq n\mathbb{P}(C) \geq 1$. Otherwise $\mathbb{P}\left(\sum_{i=1}^n I(X_i' \in C) = 0\right) = \prod_{i=1}^n \mathbb{P}(X_i' \notin C) = (1 - \mathbb{P}(C))^n$ can be as close to 0 as we want.

Similarly to the proof of the previous lecture, let

$$(X_i) \in \left\{\sup_C \frac{\mathbb{P}(C) - \frac{1}{n}\sum_{i=1}^n I(X_i \in C)}{\sqrt{\mathbb{P}(C)}} \geq t\right\}.$$

Hence, there exists $C_X$ such that

$$\frac{\mathbb{P}\left(C\right) - \frac{1}{n}\sum_{i=1}^n I(X_i \in C)}{\sqrt{\mathbb{P}\left(C\right)}} \geq t.$$

**Exercise 1.** *Show that if*

$$\frac{\mathbb{P}\left(C_X\right) - \frac{1}{n}\sum_{i=1}^n I(X_i \in C_X)}{\sqrt{\mathbb{P}\left(C_X\right)}} \geq t$$

*and*

$$\frac{1}{n}\sum_{i=1}^n I(X_i' \in C_X) \geq \mathbb{P}\left(C_X\right),$$

*then*

$$\frac{\frac{1}{n}\sum_{i=1}^n I(X_i' \in C_X) - \frac{1}{n}\sum_{i=1}^n I(X_i \in C_X)}{\sqrt{\frac{1}{n}\sum_{i=1}^n I(X_i \in C_X) + \frac{1}{n}\sum_{i=1}^n I(X_i' \in C_X)}} \geq \frac{t}{\sqrt{2}}.$$

*Hint: use the fact that $\phi(s) = \frac{s-a}{\sqrt{s}} = \sqrt{s} - \frac{s}{\sqrt{s}}$ is increasing in $s$.*

From the above exercise it follows that

$$\begin{aligned}\frac{1}{4} &\leq \mathbb{P}_{(X')}\left(\frac{1}{n}\sum_{i=1}^n I(X_i' \in C_X) \geq \mathbb{P}\left(C_X\right) \middle| \exists C_X\right)\\ &\leq \mathbb{P}_{(X')}\left(\frac{\frac{1}{n}\sum_{i=1}^n I(X_i' \in C_X) - \frac{1}{n}\sum_{i=1}^n I(X_i \in C_X)}{\sqrt{\frac{1}{n}\sum_{i=1}^n I(X_i \in C_X) + \frac{1}{n}\sum_{i=1}^n I(X_i' \in C_X)}} \geq \frac{t}{\sqrt{2}} \middle| \exists C_X\right)\end{aligned}$$

Since indicator is $0, 1$-valued,

$$\begin{aligned}&\frac{1}{4} I\left(\underbrace{\sup_C \frac{\mathbb{P}\left(C\right) - \frac{1}{n}\sum_{i=1}^n I(X_i \in C)}{\sqrt{\mathbb{P}\left(C\right)}} \geq t}_{\exists C_X}\right)\\ &\leq \mathbb{P}_{(X')}\left(\frac{\frac{1}{n}\sum_{i=1}^n I(X_i' \in C_X) - \frac{1}{n}\sum_{i=1}^n I(X_i \in C_X)}{\sqrt{\frac{1}{n}\sum_{i=1}^n I(X_i \in C_X) + \frac{1}{n}\sum_{i=1}^n I(X_i' \in C_X)}} \geq \frac{t}{\sqrt{2}} \middle| \exists C_X\right) \cdot I\left(\exists C_X\right)\\ &\leq \mathbb{P}_{(X')}\left(\sup_C \frac{\frac{1}{n}\sum_{i=1}^n I(X_i' \in C) - \frac{1}{n}\sum_{i=1}^n I(X_i \in C)}{\sqrt{\frac{1}{n}\sum_{i=1}^n I(X_i \in C) + \frac{1}{n}\sum_{i=1}^n I(X_i' \in C)}} \geq \frac{t}{\sqrt{2}}\right).\end{aligned}$$

Hence,

$$\begin{aligned}&\frac{1}{4}\mathbb{P}\left(\sup_C \frac{\mathbb{P}\left(C_X\right) - \frac{1}{n}\sum_{i=1}^n I(X_i \in C_X)}{\sqrt{\mathbb{P}\left(C_X\right)}} \geq t\right)\\ &\leq \mathbb{P}\left(\sup_C \frac{\frac{1}{n}\sum_{i=1}^n I(X_i' \in C) - \frac{1}{n}\sum_{i=1}^n I(X_i \in C)}{\sqrt{\frac{1}{n}\sum_{i=1}^n I(X_i \in C) + \frac{1}{n}\sum_{i=1}^n I(X_i' \in C)}} \geq \frac{t}{\sqrt{2}}\right)\\ &= \mathbb{E}\mathbb{P}_\varepsilon\left(\sup_C \frac{\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(I(X_i' \in C) - I(X_i \in C)\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^n I(X_i \in C) + \frac{1}{n}\sum_{i=1}^n I(X_i' \in C)}} \geq \frac{t}{\sqrt{2}}\right).\end{aligned}$$

There exist $C_1, \ldots, C_N$, with $N \le \triangle_{2n}(\mathcal{C}, X_1, \ldots, X_n, X'_1, \ldots, X'_n)$. Therefore,

$$
\begin{aligned}
&\mathbb{E}\mathbb{P}_\varepsilon\left(\sup_C \frac{\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(I(X'_i \in C) - I(X_i \in C)\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^n I(X_i \in C) + \frac{1}{n}\sum_{i=1}^n I(X'_i \in C)}} \ge \frac{t}{\sqrt{2}}\right)\\
=&\mathbb{E}\mathbb{P}_\varepsilon\left(\bigcup_{k\le N}\left\{\frac{\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(I(X'_i \in C_k) - I(X_i \in C_k)\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^n I(X_i \in C_k) + \frac{1}{n}\sum_{i=1}^n I(X'_i \in C_k)}} \ge \frac{t}{\sqrt{2}}\right\}\right)\\
\le&\mathbb{E}\sum_{k=1}^N \mathbb{P}_\varepsilon\left(\frac{\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(I(X'_i \in C_k) - I(X_i \in C_k)\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^n I(X_i \in C_k) + \frac{1}{n}\sum_{i=1}^n I(X'_i \in C_k)}} \ge \frac{t}{\sqrt{2}}\right)\\
=&\mathbb{E}\sum_{k=1}^N \mathbb{P}_\varepsilon\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(I(X'_i \in C_k) - I(X_i \in C_k)\right) \ge \frac{t}{\sqrt{2}}\sqrt{\frac{1}{n}\sum_{i=1}^n I(X_i \in C_k) + \frac{1}{n}\sum_{i=1}^n I(X'_i \in C_k)}\right)
\end{aligned}
$$

The last expression can be upper-bounded by Hoeffding's inequality as follows:

$$
\begin{aligned}
&\mathbb{E}\sum_{k=1}^N \mathbb{P}_\varepsilon\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(I(X'_i \in C_k) - I(X_i \in C_k)\right) \ge \frac{t}{\sqrt{2}}\sqrt{\frac{1}{n}\sum_{i=1}^n \left(I(X_i \in C_k) + I(X'_i \in C_k)\right)}\right)\\
&\le \mathbb{E}\sum_{k=1}^N \exp\left(-\frac{t^2\frac{1}{n}\sum_{i=1}^n\left(I(X_i \in C_k) + I(X'_i \in C_k)\right)}{2\frac{1}{n^2}2\sum\left(I(X'_i \in C_k) - I(X_i \in C_k)\right)^2}\right)
\end{aligned}
$$

since upper sum in the exponent is bigger than the lower sum (compare term-by-term)

$$
\le \mathbb{E}\sum_{k=1}^N e^{-\frac{nt^2}{4}} \le \left(\frac{2en}{V}\right)^V e^{-\frac{nt^2}{4}}.
$$

□

**VC-subgraph classes of functions**

Let $\mathcal{F} = \{f : \mathcal{X} \mapsto \mathbb{R}\}$ and

$$C_f = \{(x,t) \in \mathcal{X} \times \mathbb{R} : 0 \le t \le f(x) \text{ or } f(x) \le t \le 0\}.$$

Define class of sets $\mathcal{C} = \{C_f : f \in \mathcal{F}\}$.

**Definition 12.1.** *If $\mathcal{C}$ is a VC class of sets, then $\mathcal{F}$ is VC-subgraph class of functions and, by definition, $VC(\mathcal{F}) = VC(\mathcal{C})$.*

Note that equivalent definition of $C_f$ is

$$C'_f = \{(x,t) \in \mathcal{X} \times \mathbb{R} : |f(x)| \ge |t|\}.$$

**Example 12.1.** $\mathcal{C} = \{C \subseteq \mathcal{X}\}$, $\mathcal{F}(\mathcal{C}) = \{I(X \in C) : C \in \mathcal{C}\}$. Then $\mathcal{F}(\mathcal{C})$ is VC-subgraph class if and only if $\mathcal{C}$ is a VC class of sets.

Assume $d$ functions are fixed: $\{f_1, \ldots, f_d\} : \mathcal{X} \mapsto \mathbb{R}$. Let

$$\mathcal{F} = \left\{ \sum_{i=1}^{d} \alpha_i f_i(x) : \alpha_1, \ldots, \alpha_d \in \mathbb{R} \right\}.$$

Then $VC(\mathcal{F}) \le d+1$. To prove this, it's easier to use the second definition.

**Packing and covering numbers**

Let $f, g \in \mathcal{F}$ and assume we have a distance function $d(f,g)$.

**Example 12.2.** If $X_1, \ldots, X_n$ are data points, then

$$d_1(f,g) = \frac{1}{n}\sum_{i=1}^{n} |f(X_i) - g(X_i)|$$

and

$$d_2(f,g) = \left( \frac{1}{n}\sum_{i=1}^{n} \left(f(X_i) - g(X_i)\right)^2 \right)^{1/2}.$$

**Definition 12.2.** *Given $\varepsilon > 0$ and $f_1, \ldots, f_N \in \mathcal{F}$, we say that $f_1, \ldots, f_N$ are $\varepsilon$-separated if $d(f_i, f_j) > \varepsilon$ for any $i \ne j$.*

**Definition 12.3.** *The $\varepsilon$-packing number, $\mathcal{D}(\mathcal{F}, \varepsilon, d)$, is the maximal cardinality of an $\varepsilon$-separated set.*

Note that $\mathcal{D}(\mathcal{F}, \varepsilon, d)$ is decreasing in $\varepsilon$.

**Definition 12.4.** *Given $\varepsilon > 0$ and $f_1, \ldots, f_N \in \mathcal{F}$, we say that the set $f_1, \ldots, f_N$ is an $\varepsilon$-cover of $\mathcal{F}$ if for any $f \in \mathcal{F}$, there exists $1 \le i \le N$ such that $d(f, f_i) \le \varepsilon$.*

**Definition 12.5.** *The $\varepsilon$-covering number, $\mathcal{N}(\mathcal{F}, \varepsilon, d)$, is the minimal cardinality of an $\varepsilon$-cover of $\mathcal{F}$.*

**Lemma 12.1.**

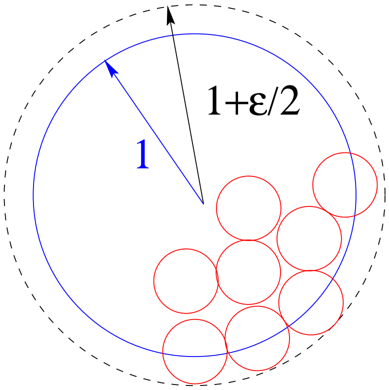
$$\mathcal{D}(\mathcal{F}, 2\varepsilon, d) \le \mathcal{N}(\mathcal{F}, \varepsilon, d) \le \mathcal{D}(\mathcal{F}, \varepsilon, d).$$

*Proof.* To prove the first inequality, assume that $\mathcal{D}(\mathcal{F}, 2\varepsilon, d) > \mathcal{N}(\mathcal{F}, \varepsilon, d)$. Let the packing corresponding to the packing number $\mathcal{D}(\mathcal{F}, 2\varepsilon, d) = D$ be $f_1, \ldots, f_D$. Let the covering corresponding to the covering number $\mathcal{N}(\mathcal{F}, \varepsilon, d) = N$ be $f'_1, \ldots, f'_N$. Since $D > N$, there exist $f_i$ and $f_j$ such that for some $f'_k$

$$d(f_i, f'_k) \le \varepsilon \text{ and } d(f_j, f'_k) \le \varepsilon.$$

Therefore, by triangle inequality, $d(f_i, f_j) \le 2\varepsilon$, which is a contradiction.

To prove the second inequality, assume $f_1, \ldots, f_D$ is an optimal packing. For any $f \in \mathcal{F}$, $f_1, \ldots, f_D, f$ would also be $\varepsilon$-packing if $d(f, f_i) > \varepsilon$ for all $i$. Since $f_1, \ldots, f_D$ is optimal, this cannot be true, and, therefore, for any $f \in \mathcal{F}$ there exists $f_i$ such that $d(f, f_i) \le \varepsilon$. Hence $f_1, \ldots, f_D$ is also a cover. Hence, $\mathcal{N}(\mathcal{F}, \varepsilon, d) \le \mathcal{D}(\mathcal{F}, \varepsilon, d)$. □



**Example 12.3.** Consider the $L_1$-ball $\{x \in \mathbb{R}^d, |x| \le 1\} = B_1(0)$ and $d(x, y) = |x - y|_1$. Then

$$\mathcal{D}(B_1(0), \varepsilon, d) \le \left(\frac{2+\varepsilon}{\varepsilon}\right)^d \le \left(\frac{3}{\varepsilon}\right)^d,$$

where $\varepsilon \le 1$. Indeed, let $f_1, \ldots, f_D$ be optimal $\varepsilon$-packing. Then the volume of the ball with $\varepsilon/2$-fattening (so that the center of small balls fall within the boundary) is

$$\text{Vol}\left(1 + \frac{\varepsilon}{2}\right) = C_d \left(1 + \frac{\varepsilon}{2}\right)^d.$$

Moreover, the volume of each of the small balls

$$\text{Vol}\left(\frac{\varepsilon}{2}\right) = C_d \left(\frac{\varepsilon}{2}\right)^d$$

and the volume of all the small balls is

$$DC_d \left(\frac{\varepsilon}{2}\right)^d.$$

Therefore,

$$D \leq \left( \frac{2+\varepsilon}{\varepsilon} \right)^d .$$

**Definition 12.6.** $\log \mathcal{N}(\mathcal{F}, \varepsilon, d)$ *is called metric entropy.*

For example, $\log \mathcal{N}(B_1(0), \varepsilon, d) \leq d \log \frac{3}{\varepsilon}$.

**Theorem 13.1.** *Assume $\mathcal{F}$ is a VC-subgraph class and $VC(\mathcal{F}) = V$. Suppose $-1 \leq f(x) \leq 1$ for all $f \in \mathcal{F}$ and $x \in \mathcal{X}$. Let $x_1, \ldots, x_n \in \mathcal{X}$ and define $d(f,g) = \frac{1}{n}\sum_{i=1}^n |f(x_i) - g(x_i)|$. Then*

$$\mathcal{D}(\mathcal{F}, \varepsilon, d) \leq \left(\frac{8e}{\varepsilon}\log\frac{7}{\varepsilon}\right)^V.$$

*(which is $\leq \left(\frac{K}{\varepsilon}\right)^{V+\delta}$ for some $\delta$.)*

*Proof.* Let $m = \mathcal{D}(\mathcal{F}, \varepsilon, d)$ and $f_1, \ldots, f_m$ be $\varepsilon$-separated, i.e.

$$\frac{1}{n}\sum_{i=1}^n |f_r(x_i) - f_\ell(x_i)| > \varepsilon.$$

Let $(z_1, t_1), \ldots, (z_k, t_k)$ be constructed in the following way: $z_i$ is chosen uniformly from $x_1, \ldots, x_n$ and $t_i$ is uniform on $[-1, 1]$.

Consider $f_r$ and $f_\ell$ from the $\varepsilon$-packing. Let $C_{f_r}$ and $C_{f_\ell}$ be subgraphs of $f_r$ and $f_\ell$. Then

$$\begin{aligned}
&\mathbb{P}\left(C_{f_r} \text{ and } C_{f_\ell} \text{ pick out different subsets of } (z_1,t_1),\ldots,(z_k,t_k)\right)\\
&= \mathbb{P}\left(\text{At least one point } (z_i,t_i) \text{ is picked by } C_{f_r} \text{ or } C_{f_\ell} \text{ but not picked by the other}\right)\\
&= 1 - \mathbb{P}\left(\text{All points } (z_i,t_i) \text{ are picked either by both or by none}\right)\\
&= 1 - \mathbb{P}\left((z_i,t_i) \text{ is picked either by both or by none}\right)^k
\end{aligned}$$

□

Since $z_i$ is drawn uniformly from $x_1, \ldots, x_n$,

$$\begin{aligned}
&\mathbb{P}\left((z_1,t_1) \text{ is picked by both } C_{f_r}, C_{f_\ell} \text{ or by neither}\right)\\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{P}\left((x_i,t_1) \text{ is picked by both } C_{f_r}, C_{f_\ell} \text{ or by neither}\right)\\
&= \frac{1}{n}\sum_{i=1}^n \left(1 - \frac{1}{2}|f_r(x_i) - f_\ell(x_i)|\right)\\
&= 1 - \frac{1}{2}\frac{1}{n}\sum_{i=1}^n |f_r(x_i) - f_\ell(x_i)|\\
&= 1 - \frac{1}{2}d(f_r, f_\ell) \leq 1 - \varepsilon/2 \leq e^{-\varepsilon/2}
\end{aligned}$$

Substituting,

$$\begin{aligned}
&\mathbb{P}\left(C_{f_r} \text{ and } C_{f_\ell} \text{ pick out different subsets of } (z_1, t_1), \ldots, (z_k, t_k)\right) \\
&= 1 - \mathbb{P}\left((z_1, t_1) \text{ is picked by both } C_{f_r}, C_{f_\ell} \text{ or by neither}\right)^k \\
&\geq 1 - \left(e^{-\varepsilon/2}\right)^k \\
&= 1 - e^{-k\varepsilon/2}
\end{aligned}$$

There are $\binom{m}{2}$ ways to choose $f_r$ and $f_\ell$, so

$$\mathbb{P}\left(\text{All pairs } C_{f_r} \text{ and } C_{f_\ell} \text{ pick out different subsets of } (z_1, t_1), \ldots, (z_k, t_k)\right) \geq 1 - \binom{m}{2} e^{-k\varepsilon/2}.$$

What $k$ should we choose so that $1 - \binom{m}{2} e^{-k\varepsilon/2} > 0$? Choose

$$k > \frac{2}{\varepsilon} \log \binom{m}{2}.$$

Then there exist $(z_1, t_1), \ldots, (z_k, t_k)$ such that all $C_{f_\ell}$ pick out different subsets. But $\{C_f : f \in \mathcal{F}\}$ is VC, so by Sauer's lemma, we can pick out at most $\left(\frac{ek}{V}\right)^V$ out of these $k$ points. Hence, $m \leq \left(\frac{ek}{V}\right)^V$ as long as $k > \frac{2}{\varepsilon} \log \binom{m}{2}$. The latter holds for $k = \frac{2}{\varepsilon} \log m^2$. Therefore,

$$m \leq \left(\frac{e}{V} \frac{2}{\varepsilon} \log m^2\right)^V = \left(\frac{4e}{V\varepsilon} \log m\right)^V,$$

where $m = \mathcal{D}(\mathcal{F}, \varepsilon, d)$. Hence, we get

$$m^{1/V} \leq \frac{4e}{\varepsilon} \log m^{1/V}$$

and defining $m^{1/V} = s$,

$$s \leq \frac{4e}{\varepsilon} \log s.$$

Note that $\frac{s}{\log s}$ is increasing for $s \geq e$ and so for large enough $s$, the inequality will be violated. We now check that the inequality is violated for $s' = \frac{8e}{\varepsilon} \log \frac{7}{\varepsilon}$. Indeed, one can show that

$$\frac{4e}{\varepsilon} \log \left(\frac{7}{\varepsilon}\right)^2 > \frac{4e}{\varepsilon} \log \left(\frac{8e}{\varepsilon} \log \frac{7}{\varepsilon}\right)$$

since

$$\frac{49}{8e\varepsilon} > \log \frac{7}{\epsilon}.$$

Hence, $m^{1/V} = s \leq s'$ and, thus,

$$\mathcal{D}(\mathcal{F}, \varepsilon, d) \leq \left(\frac{8e}{\varepsilon} \log \frac{7}{\varepsilon}\right)^V.$$

For $f \in F \subseteq [-1,1]^n$, define $R(f) = \frac{1}{n}\sum_{i=1}^n \varepsilon_i f_i$. Let $d(f,g) := \left(\frac{1}{n}\sum_{i=1}^n (f_i - g_i)^2\right)^{1/2}$.

**Theorem 14.1.**

$$\mathbb{P}\left(\forall f \in F, R(f) \leq \frac{2^{9/2}}{\sqrt{n}} \int_0^{d(0,f)} \log^{1/2} \mathcal{D}(F,\varepsilon,d)d\varepsilon + 2^{7/2}d(0,f)\sqrt{\frac{u}{n}}\right) \geq 1 - e^{-u}$$

*for any $u > 0$.*

*Proof.* Without loss of generality, assume $0 \in F$.

*Kolmogorov's chaining technique*: define a sequence of subsets

$$\{0\} = F_0 \subseteq F_1 \ldots \subseteq F_j \subseteq \ldots \subseteq F$$

where $F_j$ is defined such that

(1) $\forall f,g \in F_j, d(f,g) > 2^{-j}$
(2) $\forall f \in F$, we can find $g \in F_j$ such that $d(f,g) \leq 2^{-j}$

How to construct $F_{j+1}$ if we have $F_j$:

- $F_{j+1} := F_j$
- Find $f \in F$, $d(f,g) > 2^{-(j+1)}$ for all $g \in F_{j+1}$
- Repeat until you cannot find such $f$

Define projection $\pi_j : F \mapsto F_j$ as follows: for $f \in F$ find $g \in F_j$ with $d(f,g) \leq 2^{-j}$ and set $\pi_j(f) = g$.

For any $f \in F$,

$$\begin{aligned} f &= \pi_0(f) + (\pi_1(f) - \pi_0(f)) + (\pi_2(f) - \pi_1(f)) \ldots \\ &= \sum_{j=1}^{\infty}(\pi_j(f) - \pi_{j-1}(f)) \end{aligned}$$

Moreover,

$$\begin{aligned} d(\pi_{j-1}(f), \pi_j(f)) &\leq d(\pi_{j-1}(f), f) + d(f, \pi_j(f)) \\ &\leq 2^{-(j-1)} + 2^{-j} = 3 \cdot 2^{-j} \leq 2^{-j+2} \end{aligned}$$

Define the links

$$L_{j-1,j} = \{f - g : f \in F_j, g \in F_{j-1}, d(f,g) \leq 2^{-j+2}\}.$$

Since $R$ is linear, $R(f) = \sum_{j=1}^{\infty} R(\pi_j(f) - \pi_{j-1}(f))$. We first show how to control $R$ on the links. Assume $\ell \in L_{j-1,j}$. Then by Hoeffding's inequality

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \ell_i \geq t\right) &\leq \exp\left(-\frac{t^2}{2\sum \frac{1}{n^2}\ell_i^2}\right)\\ &= \exp\left(-\frac{nt^2}{2\frac{1}{n}\sum_{i=1}^{n}\ell_i^2}\right)\\ &\leq \exp\left(-\frac{nt^2}{2\cdot 2^{-2j+4}}\right)\end{aligned}$$

Note that

$$\mathrm{card} L_{j-1,j} \leq \mathrm{card} F_{j-1} \cdot \mathrm{card} F_j \leq (\mathrm{card} F_j)^2.$$

$$\begin{aligned}\mathbb{P}\left(\forall \ell \in L_{j-1,j}, R(\ell) = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \ell_i \leq t\right) &\geq 1 - (\mathrm{card} F_j)^2 e^{-\frac{nt^2}{2\cdot 2^{-2j+5}}}\\ &= 1 - \frac{1}{(\mathrm{card} F_j)^2} e^{-u}\end{aligned}$$

after changing the variable such that

$$t = \sqrt{\frac{2^{-2j+5}}{n}\left(4\log(\mathrm{card} F_j) + u\right)} \leq \sqrt{\frac{2^{-2j+5}}{n}4\log(\mathrm{card} F_j)} + \sqrt{\frac{2^{-2j+5}}{n}u}.$$

Hence,

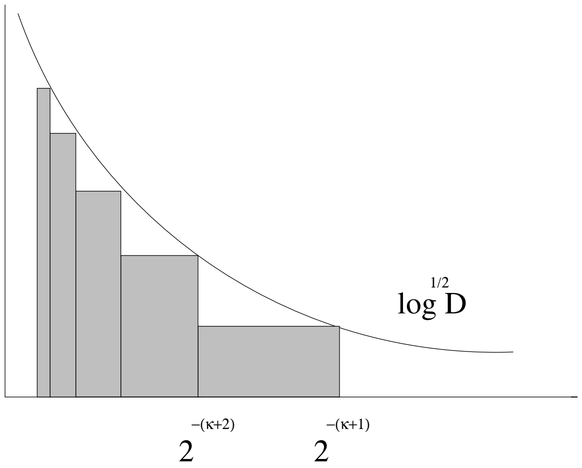
$$\mathbb{P}\left(\forall \ell \in L_{j-1,j}, R(\ell) \leq \frac{2^{7/2}2^{-j}}{\sqrt{n}}\log^{1/2}(\mathrm{card} F_j) + 2^{5/2}2^{-j}\sqrt{\frac{u}{n}}\right) \geq 1 - \frac{1}{(\mathrm{card} F_j)^2}e^{-u}.$$

If $F_{j-1} = F_j$ then by definition $\pi_{j-1}(f) = \pi_f$ and $L_{j-1,j} = \{0\}$.

By union bound for all steps,

$$\begin{aligned}&\mathbb{P}\left(\forall j \geq 1, \forall \ell \in L_{j-1,j}, R(\ell) \leq \frac{2^{7/2}2^{-j}}{\sqrt{n}}\log^{1/2}(\mathrm{card} F_j) + 2^{5/2}2^{-j}\sqrt{\frac{u}{n}}\right)\\ &\geq 1 - \sum_{j=1}^{\infty}\frac{1}{(\mathrm{card} F_j)^2}e^{-u}\\ &\geq 1 - \left(\frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2}\right)e^{-u}\\ &= 1 - (\pi^2/6 - 1)e^{-u} \geq 1 - e^{-u}\end{aligned}$$

Recall that $R(f) = \sum_{j=1}^{\infty} R(\pi_j(f) - \pi_{j-1}(f))$. If $f$ is close to 0, $-2^{k+1} < d(0, f) \leq 2^{-k}$. Find such a $k$. Then $\pi_0(f) = \ldots = \pi_k(f) = 0$ and so



$$
\begin{aligned}
R(f) &= \sum_{j=k+1}^{\infty} R(\pi_j(f) - \pi_{j-1}(f)) \\
&\le \sum_{j=k+1}^{\infty} \left( \frac{2^{7/2}}{\sqrt{n}} 2^{-j} \log^{1/2}(\mathrm{card} F_j) + 2^{5/2} 2^{-j} \sqrt{\frac{u}{n}} \right) \\
&\le \sum_{j=k+1}^{\infty} \left( \frac{2^{7/2}}{\sqrt{n}} 2^{-j} \log^{1/2} \mathcal{D}(F, 2^{-j}, d) \right) + 2^{5/2} 2^{-k} \sqrt{\frac{u}{n}}
\end{aligned}
$$

Note that $2^{-k} < 2d(f,0)$, so

$$2^{5/2} 2^{-k} < 2^{7/2} d(f,0).$$

Furthermore,

$$
\begin{aligned}
\frac{2^{9/2}}{\sqrt{n}} \sum_{j=k+1}^{\infty} \left( 2^{-(j+1)} \log^{1/2} \mathcal{D}(F, 2^{-j}, d) \right) &\le \frac{2^{9/2}}{\sqrt{n}} \int_0^{2^{-(k+1)}} \log^{1/2} \mathcal{D}(F, \varepsilon, d) d\varepsilon \\
&\le \frac{2^{9/2}}{\sqrt{n}} \underbrace{\int_0^{d(0,f)} \log^{1/2} \mathcal{D}(F, \varepsilon, d) d\varepsilon}_{\text{Dudley's entropy integral}}
\end{aligned}
$$

since $2^{-(k+1)} < d(0,f)$.

$\square$

**Lemma 15.1.** *Let $\xi, \nu$ - random variables. Assume that*

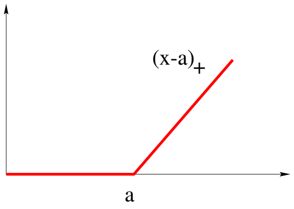
$$\mathbb{P}(\nu \geq t) \leq \Gamma e^{-\gamma t}$$

*where $\Gamma \geq 1$, $t \geq 0$, and $\gamma > 0$. Furthermore, for all $a > 0$ assume that*

$$\mathbb{E}\phi(\xi) \leq \mathbb{E}\phi(\nu)$$

*where $\phi(x) = (x-a)_+$. Then*

$$\mathbb{P}(\xi \geq t) \leq \Gamma \cdot e \cdot e^{-\gamma t}.$$



*Proof.* Since $\phi(x) = (x-a)_+$, we have $\phi(\xi) \geq \phi(t)$ whenever $\xi \geq t$.

$$\begin{aligned}\mathbb{P}(\xi \geq t) &\leq \mathbb{P}(\phi(\xi) \geq \phi(t)) \\ &\leq \frac{\mathbb{E}\phi(\xi)}{\phi(t)} \leq \frac{\mathbb{E}\phi(\nu)}{\phi(t)} = \frac{\mathbb{E}(\nu-a)_+}{(t-a)_+}\end{aligned}$$

Furthermore,

$$\begin{aligned}\mathbb{E}(\nu-a)_+ &= \mathbb{E}\int_0^{(\nu-a)_+} 1dx \\ &= \mathbb{E}\int_0^{\infty} I(x \leq (\nu-a)_+)dx \\ &= \int_0^{\infty} \mathbb{E}I(x \leq (\nu-a)_+)dx \\ &= \int_0^{\infty} \mathbb{P}((\nu-a)_+ \geq x)\,dx \\ &= \int_0^{\infty} \mathbb{P}(\nu \geq a+x)\,dx \\ &\leq \int_0^{\infty} \Gamma e^{-\gamma a-\gamma x}dx = \frac{\Gamma e^{-\gamma a}}{\gamma}.\end{aligned}$$

Hence,

$$\mathbb{P}(\xi \geq t) \leq \frac{\Gamma e^{-\gamma a}}{\gamma(t-a)_+} = \frac{\Gamma \cdot e \cdot e^{-\gamma t}}{1} = \Gamma \cdot e \cdot e^{-\gamma t}$$

where we chose optimal $a = t - \frac{1}{\gamma}$ to minimize $\frac{\Gamma e^{-\gamma a}}{\gamma}$. □

**Lemma 15.2.** *Let $x = (x_1, \ldots, x_n)$, $x' = (x'_1, \ldots, x'_n)$. If for functions $\varphi_1(x, x')$, $\varphi_2(x, x')$, $\varphi_3(x, x')$*

$$\mathbb{P}\left(\varphi_1(x, x') \geq \varphi_2(x, x') + \sqrt{\varphi_3(x, x') \cdot t}\right) \leq \Gamma e^{-\gamma t}$$

*then*

$$\mathbb{P}\left(\mathbb{E}_{x'}\varphi_1(x, x') \geq \mathbb{E}_{x'}\varphi_2(x, x') + \sqrt{\mathbb{E}_{x'}\varphi_3(x, x') \cdot t}\right) \leq \Gamma \cdot e \cdot e^{-\gamma t}.$$

*(i.e. if the inequality holds, then it holds with averaging over one of the copies)*

*Proof.* First, note that $\sqrt{ab} = \inf_{\delta>0}(\delta a + \frac{b}{4\delta})$ with $\delta_* = \sqrt{\frac{b}{4a}}$ achieving the infima. Hence,

$$\begin{aligned}\{\varphi_1 \geq \varphi_2 + \sqrt{\varphi_3 t}\} &= \{\exists \delta > 0, \varphi_1 \geq \varphi_2 + \delta\varphi_3 + \frac{t}{4\delta}\} \\ &= \{\exists \delta > 0, (\varphi_1 - \varphi_2 - \delta\varphi_3)4\delta \geq t\} \\ &= \{\underbrace{\sup_{\delta>0}(\varphi_1 - \varphi_2 - \delta\varphi_3)4\delta}_{\nu} \geq t\}\end{aligned}$$

and similarly

$$\{\mathbb{E}_{x'}\varphi_1 \geq \mathbb{E}_{x'}\varphi_2 + \sqrt{\mathbb{E}_{x'}\varphi_3 t}\} = \{\underbrace{\sup_{\delta>0}(\mathbb{E}_{x'}\varphi_1 - \mathbb{E}_{x'}\varphi_2 - \delta\mathbb{E}_{x'}\varphi_3)4\delta}_{\xi} \geq t\}.$$

By assumption, $\mathbb{P}(\nu \geq t) \leq \Gamma e^{-\gamma t}$. We want to prove $\mathbb{P}(\xi \geq t) \leq \Gamma \cdot e \cdot e^{-\gamma t}$. By the previous lemma, we only need to check whether $\mathbb{E}\phi(\xi) \leq \mathbb{E}\phi(\nu)$.

$$\begin{aligned}\xi &= \sup_{\delta>0} \mathbb{E}_{x'}(\varphi_1 - \varphi_2 - \delta\varphi_3)4\delta \\ &\leq \mathbb{E}_{x'} \sup_{\delta>0}(\varphi_1 - \varphi_2 - \delta\varphi_3)4\delta \\ &= \mathbb{E}_{x'}\nu\end{aligned}$$

Thus,

$$\phi(\xi) \leq \phi(\mathbb{E}_{x'}\nu) \leq \mathbb{E}_{x'}\phi(\nu)$$

by Jensen's inequality ($\phi$ is convex). Hence,

$$\mathbb{E}\phi(\xi) \leq \mathbb{E}\mathbb{E}_{x'}\phi(\nu) = \mathbb{E}\phi(\nu).$$

□

We will now use Lemma 15.2. Let $\mathcal{F} = \{f : \mathcal{X} \mapsto [c, c+1]\}$. Let $x_1, \ldots, x_n, x'_1, \ldots, x'_n$ be i.i.d. random variables. Define

$$F = \{(f(x_1) - f(x'_1), \ldots, f(x_n) - f(x'_n)) : f \in \mathcal{F}\} \subseteq [-1, 1]^n.$$

Define

$$d(f,g) = \left(\frac{1}{n}\sum_{i=1}^n \left((f(x_i) - f(x_i')) - (g(x_i) - g(x_i'))\right)^2\right)^{1/2}.$$

In Lecture 14, we proved

$$\mathbb{P}_\varepsilon\left(\forall f \in \mathcal{F}, \frac{1}{n}\sum_{i=1}^n \varepsilon_i(f(x_i) - f(x_i')) \leq \frac{2^{9/2}}{\sqrt{n}}\int_0^{d(0,f)} \log^{1/2}\mathcal{D}(\mathcal{F},\varepsilon,d)d\varepsilon \right.$$
$$\left. +2^{7/2}d(0,f)\sqrt{\frac{t}{n}}\right) \geq 1 - e^{-t}.$$

Complement of the above is

$$\mathbb{P}_\varepsilon\left(\exists f \in \mathcal{F}, \frac{1}{n}\sum_{i=1}^n \varepsilon_i(f(x_i) - f(x_i')) \geq \frac{2^{9/2}}{\sqrt{n}}\int_0^{d(0,f)} \log^{1/2}\mathcal{D}(\mathcal{F},\varepsilon,d)d\varepsilon + 2^{7/2}d(0,f)\sqrt{\frac{t}{n}}\right) \leq e^{-t}.$$

Taking expectation with respect to $x, x'$, we get

$$\mathbb{P}\left(\exists f \in \mathcal{F},\ \frac{1}{n}\sum_{i=1}^n \varepsilon_i(f(x_i) - f(x_i')) \geq \frac{2^{9/2}}{\sqrt{n}}\int_0^{d(0,f)} \log^{1/2}\mathcal{D}(\mathcal{F},\varepsilon,d)d\varepsilon + 2^{7/2}d(0,f)\sqrt{\frac{t}{n}}\right) \leq e^{-t}.$$

Hence (see below)

$$\mathbb{P}\left(\exists f \in \mathcal{F},\ \frac{1}{n}\sum_{i=1}^n (f(x_i) - f(x_i')) \geq \frac{2^{9/2}}{\sqrt{n}}\int_0^{d(0,f)} \log^{1/2}\mathcal{D}(\mathcal{F},\varepsilon,d)d\varepsilon + 2^{7/2}d(0,f)\sqrt{\frac{t}{n}}\right) \leq e^{-t}.$$

To see why the above step holds, notice that $d(f,g)$ is invariant under permutations $x_i \leftrightarrow x_i'$. We can remove $\varepsilon_i$ since $x$ and $x'$ are i.i.d and we can switch $x_i$ and $x_i'$. To the right of "$\geq$" sign, only distance $d(f,g)$ depends on $x, x'$, but it's invariant to the permutations.

By Lemma 15.2 (minus technical detail "$\exists f$"),

$$\mathbb{P}\left(\exists f \in \mathcal{F},\ \mathbb{E}_{x'}\frac{1}{n}\sum_{i=1}^n (f(x_i) - f(x_i')) \geq \mathbb{E}_{x'}\frac{2^{9/2}}{\sqrt{n}}\int_0^{d(0,f)} \log^{1/2}\mathcal{D}(\mathcal{F},\varepsilon,d)d\varepsilon \right.$$
$$\left. +2^{7/2}\sqrt{\frac{\mathbb{E}_{x'}d(0,f)^2 t}{n}}\right) \leq e \cdot e^{-t},$$

where

$$\mathbb{E}_{x'}\frac{1}{n}\sum_{i=1}^n (f(x_i) - f(x_i')) = \frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}f$$

and

$$\mathbb{E}_{x'}d(0,f)^2 = \mathbb{E}_{x'}\frac{1}{n}\sum_{i=1}^n (f(x_i) - f(x_i'))^2.$$

The Dudley integral above will be bounded by something non-random in the later lectures.

In Lecture 15, we proved the following *Generalized VC inequality*

$$\mathbb{P}\left(\forall f \in \mathcal{F},\ \mathbb{E}f - \frac{1}{n}\sum_{i=1}^n f(x_i) \le \frac{2^{9/2}}{\sqrt{n}}\mathbb{E}_{x'}\int_0^{d(0,f)} \log^{1/2}\mathcal{D}(\mathcal{F},\varepsilon,d)d\varepsilon + 2^{7/2}\sqrt{\frac{\mathbb{E}_{x'}d(0,f)^2 t}{n}}\right) \ge 1 - e^{-t}$$

$$d(f,g) = \left(\frac{1}{n}\sum_{i=1}^n \left(f(x_i) - f(x_i') - g(x_i) + g(x_i')\right)^2\right)^{1/2}$$

**Definition 16.1.** *We say that $\mathcal{F}$ satisfies uniform entropy condition if*

$$\forall n,\ \forall(x_1,\ldots,x_n),\ \mathcal{D}(\mathcal{F},\varepsilon,d_x) \le \mathcal{D}(\mathcal{F},\varepsilon)$$

*where* $d_x(f,g) = \left(\frac{1}{n}\sum_{i=1}^n (f(x_i) - g(x_i))^2\right)^{1/2}$

**Lemma 16.1.** *If $\mathcal{F}$ satisfies uniform entropy condition, then*

$$\mathbb{E}_{x'}\int_0^{d(0,f)} \log^{1/2}\mathcal{D}(\mathcal{F},\varepsilon,d)d\varepsilon \le \int_0^{\sqrt{\mathbb{E}_{x'}d(0,f)^2}} \log^{1/2}\mathcal{D}(\mathcal{F},\varepsilon/2)d\varepsilon$$

*Proof.* Using inequality $(a+b)^2 \le 2(a^2+b^2)$,

$$\begin{aligned}
d(f,g) &= \left(\frac{1}{n}\sum_{i=1}^n \left(f(x_i) - g(x_i) + g(x_i') - f(x_i')\right)^2\right)^{1/2} \\
&\le \left(\frac{2}{n}\sum_{i=1}^n \left((f(x_i) - g(x_i))^2 + (g(x_i') - f(x_i'))^2\right)\right)^{1/2} \\
&= 2\left(\frac{1}{2n}\sum_{i=1}^n \left((f(x_i) - g(x_i))^2 + (g(x_i') - f(x_i'))^2\right)\right)^{1/2} \\
&= 2d_{x,x'}(f,g)
\end{aligned}$$

Since $d(f,g) \le 2d_{x,x'}(f,g)$, we also have

$$\mathcal{D}(\mathcal{F},\varepsilon,d) \le \mathcal{D}(\mathcal{F},\varepsilon/2,d_{x,x'}).$$

Indeed, let $f_1,...,f_N$ be optimal $\varepsilon$-packing w.r.t. distance $d$. Then

$$\varepsilon \le d(f_i,f_j) \le 2d_{x,x'}(f_i,f_j)$$

and, hence,

$$\varepsilon/2 \le d_{x,x'}(f_i,f_j).$$

So, $f_1,...,f_N$ is $\varepsilon/2$-packing w.r.t. $d_{x,x'}$. Therefore, can pack at least $N$ and so $\mathcal{D}(\mathcal{F},\varepsilon,d) \le \mathcal{D}(\mathcal{F},\varepsilon/2,d_{x,x'})$.

$$\begin{aligned}\mathbb{E}_{x'}\int_0^{d(0,f)}\log^{1/2}\mathcal{D}(\mathcal{F},\varepsilon,d)d\varepsilon &\le \mathbb{E}_{x'}\int_0^{d(0,f)}\log^{1/2}\mathcal{D}(\mathcal{F},\varepsilon/2,d_{x,x'})d\varepsilon\\ &\le \int_0^{d(0,f)}\log^{1/2}\mathcal{D}(\mathcal{F},\varepsilon/2)d\varepsilon\end{aligned}$$

Let $\phi(x) = \int_0^x \log^{1/2}\mathcal{D}(\mathcal{F},\varepsilon)d\varepsilon$. It is concave because $\phi'(x) = \log^{1/2}\mathcal{D}(\mathcal{F},\varepsilon/2)$ is decreasing when $x$ is increasing (can pack less with larger balls). Hence, by Jensen's inequality,

$$\mathbb{E}_{x'}\phi(d(0,f)) \le \phi(\mathbb{E}_{x'}d(0,f)) = \phi(\mathbb{E}_{x'}\sqrt{d(0,f)^2}) \le \phi(\sqrt{\mathbb{E}_{x'}d(0,f)^2}).$$

□

**Lemma 16.2.** *If $\mathcal{F} = \{f\colon \mathcal{X}\to[0,1]\}$, then*

$$\mathbb{E}_{x'}d(0,f)^2 \le 2\max\left(\mathbb{E}f, \frac{1}{n}\sum_{i=1}^n f(x_i)\right)$$

*Proof.*

$$\begin{aligned}\mathbb{E}_{x'}d(0,f)^2 &= \mathbb{E}_{x'}\frac{1}{n}\sum_{i=1}^n (f(x_i)-f(x_i'))^2\\ &= \frac{1}{n}\sum_{i=1}^n\left(f^2(x_i)-2f(x_i)\mathbb{E}f+\mathbb{E}f^2\right)\\ &\le \frac{1}{n}\sum_{i=1}^n(f^2(x_i)+\mathbb{E}f^2) \le \frac{1}{n}\sum_{i=1}^n f(x_i)+\mathbb{E}f\\ &\le 2\max\left(\mathbb{E}f,\frac{1}{n}\sum_{i=1}^n f(x_i)\right)\end{aligned}$$

□

**Theorem 16.1.** *If $\mathcal{F}$ satisfies Uniform Entropy Condition and $\mathcal{F} = \{f\colon\mathcal{X}\to[0,1]\}$. Then*

$$\mathbb{P}\left(\forall f\in\mathcal{F},\ \mathbb{E}f-\frac{1}{n}\sum_{i=1}^n f(x_i) \le \frac{2^{9/2}}{\sqrt{n}}\int_0^{\sqrt{2\mathbb{E}f}}\log^{1/2}\mathcal{D}(\mathcal{F},\varepsilon/2)d\varepsilon + 2^{7/2}\sqrt{\frac{2\mathbb{E}f\cdot t}{n}}\right) \ge 1-e^{-t}.$$

*Proof.* If $\mathbb{E}f \ge \frac{1}{n}\sum_{i=1}^n f(x_i)$, then

$$2\max\left(\mathbb{E}f,\frac{1}{n}\sum_{i=1}^n f(x_i)\right) = 2\mathbb{E}f.$$

If $\mathbb{E}f \le \frac{1}{n}\sum_{i=1}^n f(x_i)$,

$$\mathbb{E}f - \frac{1}{n}\sum_{i=1}^n f(x_i) \le 0$$

and the bound trivially holds. □

Another result:

$$\mathbb{P}\left(\forall f \in \mathcal{F},\ \frac{1}{n}\sum_{i=1}^{n} f(x_i) - \mathbb{E}f \leq \frac{2^{9/2}}{\sqrt{n}}\int_0^{\sqrt{2\frac{1}{n}\sum_{i=1}^n f(x_i)}} \log^{1/2}\mathcal{D}(\mathcal{F},\varepsilon/2)d\varepsilon + 2^{7/2}\sqrt{\frac{2(\frac{1}{n}\sum_{i=1}^n f(x_i))t}{n}}\right)$$

$$\geq 1 - e^{-t}.$$

**Example 16.1.** [VC-type entropy condition]

$$\log \mathcal{D}(\mathcal{F},\varepsilon) \leq \alpha \log \frac{2}{\varepsilon}.$$

For VC-subgraph classes, entropy condition is satisfied. Indeed, in Lecture 13, we proved that $\mathcal{D}(\mathcal{F},\varepsilon,d) \leq \left(\frac{8e}{\varepsilon}\log\frac{7}{\varepsilon}\right)^V$ for a VC-subgraph class $\mathcal{F}$ with $VC(\mathcal{F}) = V$, where $d(f,g) = d_1(f,g) = \frac{1}{n}\sum_{i=1}^n |f(x_i) - g(x_i)|$. Note that if $f, g : \mathcal{X} \mapsto [0,1]$, then

$$d_2(f,g) = \left(\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - g(x_i))^2\right)^{1/2} \leq \left(\frac{1}{n}\sum_{i=1}^{n}|f(x_i) - g(x_i)|\right)^{1/2}.$$

Hence, $\varepsilon < d_2(f,g) \leq \sqrt{d_1(f,g)}$ implies

$$\mathcal{D}(\mathcal{F},\varepsilon,d_2) \leq \mathcal{D}(\mathcal{F},\varepsilon^2,d_1) \leq \left(\frac{8e}{\varepsilon^2}\log\frac{7}{\varepsilon^2}\right)^V = \mathcal{D}(\mathcal{F},\varepsilon).$$

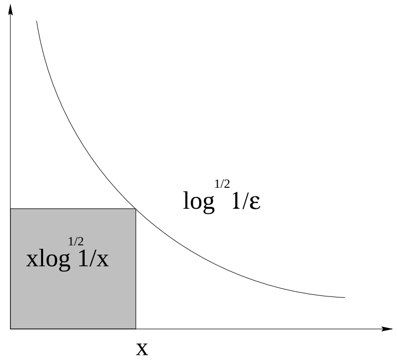
The entropy is

$$\log \mathcal{D}(\mathcal{F},\varepsilon) \leq \log\left(\frac{8e}{\varepsilon^2}\log\frac{7}{\varepsilon^2}\right)^V = V\log\left(\frac{8e}{\varepsilon^2}\log\frac{7}{\varepsilon^2}\right) \leq K \cdot V \log\frac{2}{\varepsilon},$$

where $K$ is an absolute constant.

We now give an upper bound on the Dudley integral for VC-type entropy condition.

$$\int_0^x \sqrt{\log\frac{1}{\varepsilon}}d\varepsilon \leq \begin{cases} 2x\log^{1/2}\frac{1}{x} & , \quad x \leq \frac{1}{e} \\ 2x & , \quad x \geq \frac{1}{e} \end{cases}.$$



*Proof.* First, check the inequality for $x \leq 1/e$. Taking derivatives,

$$\sqrt{\log\frac{1}{x}} \leq 2\sqrt{\log\frac{1}{x}} + \frac{x}{\sqrt{\log\frac{1}{x}}}\left(-\frac{1}{x}\right)$$

$$\log\frac{1}{x} \le 2\log\frac{1}{x} - 1$$
$$1 \le \log\frac{1}{x}$$
$$x \le 1/e$$

Now, check for $x \ge 1/e$.

$$\begin{aligned}\int_0^x \sqrt{\log\frac{1}{\varepsilon}}d\varepsilon &= \int_0^{\frac{1}{e}} \sqrt{\log\frac{1}{\varepsilon}}d\varepsilon + \int_{\frac{1}{e}}^x \sqrt{\log\frac{1}{\varepsilon}}d\varepsilon \\ &\le \frac{2}{e} + \int_{\frac{1}{e}}^x 1dx \\ &= \frac{2}{e} + x - \frac{1}{e} = x + \frac{1}{e} \le 2x\end{aligned}$$

□

Using the above result, we get

$$\mathbb{P}\left(\forall f \in \mathcal{F},\ \mathbb{E}f - \frac{1}{n}\sum_{i=1}^n f(x_i) \le K\sqrt{\frac{\alpha}{n}\mathbb{E}f\log\frac{1}{\mathbb{E}f}} + K\sqrt{\frac{t\mathbb{E}f}{n}}\right) \ge 1 - e^{-t}.$$

Without loss of generality, we can assume $\mathbb{E}f \ge \frac{1}{n}$, and, therefore, $\log\frac{1}{\mathbb{E}f} \le \log n$. Hence,

$$\mathbb{P}\left(\forall f \in \mathcal{F},\ \frac{\mathbb{E}f - \frac{1}{n}\sum_{i=1}^n f(x_i)}{\sqrt{\mathbb{E}f}} \le K\sqrt{\frac{\alpha\log n}{n}} + K\sqrt{\frac{t}{n}}\right) \ge 1 - e^{-t}.$$

Consider the classification setting, i.e. $\mathcal{Y} = \{-1, +1\}$. Denote the set of weak classifiers

$$\mathcal{H} = \{h : \mathcal{X} \mapsto [-1, +1]\}$$

and assume $\mathcal{H}$ is a VC-subgraph. Hence, $\mathcal{D}(\mathcal{H}, \varepsilon, d_x) \leq K \cdot V \log 2/\varepsilon$. A voting algorithm outputs

$$f = \sum_{i=1}^{T} \lambda_i h_i, \text{ where } h_i \in \mathcal{H}, \sum_{i=1}^{T} \lambda_i \leq 1, \ \lambda_i > 0.$$

Let

$$\mathcal{F} = \text{conv } \mathcal{H} = \left\{ \sum_{i=1}^{T} \lambda_i h_i, \ h_i \in \mathcal{H}, \ \sum_{i=1}^{T} \lambda_i \leq 1, \ \lambda_i \geq 0, \ T \geq 1 \right\}.$$

Then $\text{sign}(f(x))$ is the prediction of the label $y$. Let

$$\mathcal{F}_d = \text{conv}_d \ \mathcal{H} = \left\{ \sum_{i=1}^{d} \lambda_i h_i, \ h_i \in \mathcal{H}, \ \sum_{i=1}^{T} \lambda_i \leq 1, \ \lambda_i \geq 0 \right\}.$$

**Theorem 17.1.** *For any $x = (x_1, \ldots, x_n)$, if*

$$\log \mathcal{D}(\mathcal{H}, \varepsilon, d_x) \leq KV \log 2/\varepsilon$$

*then*

$$\log \mathcal{D}(conv_d \ \mathcal{H}, \varepsilon, d_x) \leq KVd \log 2/\varepsilon.$$

*Proof.* Let $h^1, \ldots, h^D$ be $\varepsilon$-packing of $\mathcal{H}$ with respect to $d_x$, $D = \mathcal{D}(\mathcal{H}, \varepsilon, d_x)$.

Note that $d_x$ is a norm.

$$d_x(f, g) = \left( \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - g(x_i))^2 \right)^{1/2} = \|f - g\|_x.$$

If $f = \sum_{i=1}^{d} \lambda_i h_i$, for all $h_i$ we can find $h^{k_i}$ such that $d(h_i, h^{k_i}) \leq \varepsilon$. Let $f' = \sum_{i=1}^{d} \lambda_i h^{k_i}$. Then

$$d(f, f') = \|f - f'\|_x = \left\| \sum_{i=1}^{d} \lambda_i (h_i - h^{k_i}) \right\|_x \leq \sum_{i=1}^{d} \lambda_i \|h_i - h^{k_i}\|_x \leq \varepsilon.$$

Define

$$\mathcal{F}_{D,d} = \left\{ \sum_{i=1}^{d} \lambda_i h_i, \ h_i \in \{h^1, \ldots, h^D\}, \ \sum_{i=1}^{d} \lambda_i \leq 1, \ \lambda_i \geq 0 \right\}.$$

Hence, we can approximate any $f \in \mathcal{F}_d$ by $f' \in \mathcal{F}_{D,d}$ within $\varepsilon$.

Now, let $f = \sum_{i=1}^{d} \lambda_i h_i \in \mathcal{F}_{D,d}$ and consider the following construction. We will choose $Y_1(x), \ldots, Y_k(x)$ from $h_1, \ldots, h_d$ according to $\lambda_1, \ldots, \lambda_d$:

$$\mathbb{P}\left(Y_j(x) = h_i(x)\right) = \lambda_i \ \text{ and } \ \mathbb{P}\left(Y_j(x) = 0\right) = 1 - \sum_{i=1}^{d} \lambda_i.$$

Note that with this construction

$$\mathbb{E} Y_j(x) = \sum_{i=1}^{d} \lambda_i h_i(x) = f(x).$$

Furthermore,

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{k}\sum_{j=1}^{k} Y_j - f\right\|_x^2 &= \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{k}\sum_{j=1}^{k} Y_j(x_i) - f(x_i)\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\frac{1}{k}\sum_{j=1}^{k}(Y_j(x_i) - \mathbb{E}Y_j(x_i))\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k^2}\sum_{j=1}^{k}\mathbb{E}(Y_j(x_i) - \mathbb{E}Y_j(x_i))^2 \\
&\le \frac{4}{k}
\end{aligned}$$

because $|Y_j(x_i) - \mathbb{E}Y_j(x_i)| \le 2$. Choose $k = 4/\varepsilon^2$. Then

$$\mathbb{E}\left\|\frac{1}{k}\sum_{j=1}^{k} Y_j - f\right\|_x^2 = \mathbb{E}d_x\left(\frac{1}{k}\sum_{j=1}^{k} Y_j, f\right)^2 \le \varepsilon^2.$$

So, there exists a deterministic combination $\frac{1}{k}\sum_{j=1}^{k} Y_j$ such that $d_x(\frac{1}{k}\sum_{j=1}^{k} Y_j, f) \le \varepsilon$.
Define

$$\mathcal{F}'_{D,d} = \left\{\frac{1}{k}\sum_{j=1}^{k} Y_j : \ k = 4/\varepsilon^2, \ Y_j \in \{h_1, \ldots, h_d\} \subseteq \{h^1, \ldots, h^D\}\right\}$$

Hence, we can approximate any $f = \sum_{i=1}^{d}\lambda_i h_i \in \mathcal{F}_{D,d}$, $h_i \in \{h^1, \ldots, h^D\}$, by $f' \in \mathcal{F}'_{D,d}$ within $\varepsilon$.
Let us now bound the cardinality of $\mathcal{F}'_{D,d}$. To calculate the number of ways to choose $k$ functions out of $h_1, \ldots, h_d$, assume each of $h_i$ is chosen $k_d$ times such that $k = k_1 + \ldots + k_d$. We can formulate the problem as finding the number of strings of the form

$$\underbrace{00\ldots0}_{k_1}1\underbrace{00\ldots0}_{k_2}1\ldots1\underbrace{00\ldots0}_{k_d}.$$

In this string, there are $d-1$ "1"s and $k$ "0"s, and total length is $k+d-1$. The number of such strings is $\binom{k+d-1}{k}$. Hence,

$$\begin{aligned}
\text{card } \mathcal{F}'_{D,d} &\le \binom{D}{d} \times \binom{k+d}{k} \\
&\le \frac{D^{D-d}D^d}{d^d(D-d)^{D-d}} \frac{(k+d)^{k+d}}{k^k d^d} \\
&= \left(\frac{D(k+d)}{d^2}\right)^d \left(\frac{D}{D-d}\right)^{D-d} \left(\frac{k+d}{k}\right)^k \\
&= \left(\frac{D(k+d)}{d^2}\right)^d \left(1+\frac{d}{D-d}\right)^{D-d} \left(1+\frac{d}{k}\right)^k \\
&\text{using inequality } 1+x \le e^x \\
&\le \left(\frac{D(k+d)e^2}{d^2}\right)^d
\end{aligned}$$

where $k = 4/\varepsilon^2$ and $D = \mathcal{D}(\mathcal{F}, \varepsilon, d_x)$.

Therefore, we can approximate any $f \in \mathcal{F}_d$ by $f'' \in \mathcal{F}_{D,d}$ within $\varepsilon$ and $f'' \in \mathcal{F}_{D,d}$ by $f' \in \mathcal{F}'_{D,d}$ within $\varepsilon$. Hence, we can approximate any $f \in \mathcal{F}_d$ by $f' \in \mathcal{F}'_{D,d}$ within $2\varepsilon$. Moreover,

$$\begin{aligned}
\log \mathcal{N}(\mathcal{F}_d = \text{conv}_d\, \mathcal{H}, 2\varepsilon, d_x) &\le d \log \frac{e^2 D(k+d)}{d^2} \\
&= d\left(2 + \log D + \log \frac{k+d}{d^2}\right) \\
&\le d\left(2 + KV \log \frac{2}{\varepsilon} + \log\left(1+\frac{4}{\varepsilon^2}\right)\right) \\
&\le KVd \log \frac{2}{\varepsilon}
\end{aligned}$$

since $\frac{k+d}{d^2} \le 1+k$ and $d \ge 1$, $V \ge 1$. □

In this lecture, we show that although the VC-hull classes might be considerably larger than the VC-classes, they are small enough to have finite uniform entropy integral.

**Theorem 18.1.** *Let $(\mathcal{X},\mathcal{A},\mu)$ be a measurable space, $\mathcal{F}\subset\{f|f:\mathcal{X}\to\mathbb{R}\}$ be a class of measurable functions with measurable square integrable envelope $F$ (i.e., $\forall x\in\mathcal{X},\forall f\in\mathcal{F},|f(x)|<F(x)$, and $\|F\|_2=(\int F^2 d\mu)^{1/2}<\infty$), and the $\epsilon$-net of $\mathcal{F}$ satisfies $N(\mathcal{F},\epsilon\|F\|_2,\|\cdot\|)\le C\left(\frac{1}{\epsilon}\right)^V$ for $0<\epsilon<1$. Then there exists a constant $K$ that depends only on $C$ and $V$ such that $\log N(conv\mathcal{F},\epsilon\|F\|_2,\|\cdot\|)\le K\left(\frac{1}{\epsilon}\right)^{\frac{2\cdot V}{V+2}}$.*

*Proof.* Let $N(\mathcal{F},\epsilon\|F\|_2,\|\cdot\|_2)\le C\left(\frac{1}{\epsilon}\right)^V\overset{\triangle}{=}n$. Then $\epsilon=C^{1/v}n^{-1/V}$, and $\epsilon\|F\|_2=C^{1/V}\|F\|_2\cdot n^{-1/V}$. Let $L=C^{1/V}\|F\|_2$. Then $N(\mathcal{F},Ln^{-1/V},\|\cdot\|_2)\le n$ (i.e., the $L\cdot n^{-1/V}$-net of $\mathcal{F}$ contains at most $n$ elements). Construct $\mathcal{F}_1\subset\mathcal{F}_2\subset\cdots\subset\mathcal{F}_n\subset\cdots$ such that each $\mathcal{F}_n$ is a $L\cdot n^{-1/V}$-net, and contains at most $n$ elements. Let $W=\frac{1}{2}+\frac{1}{V}$. We proceed to show that there exists constants $C_k$ and $D_k$ that depend only on $C$ and $V$ and are upper bounded ($\sup_k C_k\vee D_k<\infty$), such that

$$\log N(\text{conv}\mathcal{F}_{n\cdot k^q},C_kL\cdot n^{-W},\|\cdot\|_2)\quad\le\quad D_k\cdot n \tag{18.1}$$

for $n,k\ge 1$, and $q\ge 3+V$. This implies the theorem, since if we let $k\to\infty$, we have $\log N(\text{conv}\mathcal{F},C_\infty L\cdot n^{-W},\|\cdot\|_2)\le D_\infty\cdot n$. Let $\epsilon=C_\infty C^{1/V}n^{-W}$, and $K=D_\infty C_\infty^{\frac{2\cdot V}{V+2}}C^{\frac{2}{V+2}}$, we get $C_\infty L\cdot n^{-W}=C_\infty C^{1/V}\|F\|_2 n^{-W}=\epsilon\|F\|_2$, $n=\left(\frac{C_\infty C^{1/V}}{\epsilon}\right)^{1/W}$ and $\log N(\text{conv}\mathcal{F},\epsilon\|F\|_2,\|\cdot\|_2)\le K\cdot\left(\frac{1}{\epsilon}\right)^{\frac{2\cdot V}{V+2}}$. Inequality 18.1 will proved in two steps: (1)

$$\log N(\text{conv}\mathcal{F}_n,C_1L\cdot n^{-W},\|\cdot\|_2)\quad\le\quad D_1\cdot n \tag{18.2}$$

by induction on $n$, using Kolmogorov's chaining technique, and (2) for fixed $n$,

$$\log N(\text{conv}\mathcal{F}_{n\cdot k^q},C_kL\cdot n^{-W},\|\cdot\|_2)\quad\le\quad D_k\cdot n \tag{18.3}$$

by induction on $k$, using the results of (1) and Kolmogorov's chaining technique.

For any fixed $n_0$ and any $n\le n_0$, we can choose large enough $C_1$ such that $C_1Ln_0^{-W}\ge\|F\|_2$. Thus $N(\text{conv}\mathcal{F}_n,C_1L\cdot n^{-W},\|\cdot\|_2)=1$ and 18.2 holds trivially. For general $n$, fix $m=n/d$ for large enough $d>1$. For any $f\in\mathcal{F}_n$, there exists a projection $\pi_m f\in\mathcal{F}_m$ such that $\|f-\pi_m f\|\le C^{\frac{1}{V}}m^{-\frac{1}{V}}\|F\|=Lm^{-\frac{1}{V}}$ by definition of $\mathcal{F}_m$. Since $\sum_{f\in\mathcal{F}_n}\lambda_f\cdot f=\sum_{f\in\mathcal{F}_m}\mu_f\cdot f+\sum_{f\in\mathcal{F}_n}\lambda_f\cdot(f-\pi_m f)$, we have $\text{conv}\mathcal{F}_n\subset\text{conv}\mathcal{F}_m+\text{conv}\mathcal{G}_n$, and the number of elements $|\mathcal{G}_n|\le|\mathcal{F}_n|\le n$, where $\mathcal{G}_n=\{f-\pi_m f:f\in\mathcal{F}_n\}$. We will find $\frac{1}{2}C_1Ln^{-\frac{1}{W}}$-nets for both $\mathcal{F}_m$ and $\mathcal{G}_n$, and bound the number of elements for them to finish to induction step. We need the following lemma to bound the number of elements for the $\frac{1}{2}C_1Ln^{-\frac{1}{W}}$-net of $\mathcal{G}_n$.

**Lemma 18.2.** *Let $(\mathcal{X},\mathcal{A},\mu)$ be a measurable space and $\mathcal{F}$ be an arbitrary set of $n$ measurable functions $f:\mathcal{X}\to\mathbb{R}$ of finite $L_2(\mu)$- diameter $diam\mathcal{F}$ ($\forall f,g\in\mathcal{F},\int(f-g)^2d\mu<\infty$). Then $\forall\epsilon>0$, $N(conv\mathcal{F},\epsilon diam\mathcal{F},\|\cdot\|_2)\le\left(e+en\epsilon^2\right)^{2/\epsilon^2}$.*

*Proof.* Let $\mathcal{F} = \{f_1, \cdots, f_n\}$. $\forall \sum_{i=1}^n \lambda_i f_i$, let $Y_1, \cdots, Y_k$ be i.i.d. random variables such that $P(Y_i = f_j) = \lambda_j$ for all $j = 1, \cdots, n$. It follows that $\mathbb{E}Y_i = \sum \lambda_j f_j$ for all $i = 1, \cdots, k$, and

$$\mathbb{E}\left(\frac{1}{k}\sum_{i=1}^k Y_i - \sum_{j=1}^n \lambda_j f_j\right) \le \frac{1}{k}\mathbb{E}\left(Y_1 - \sum_{j=1}^n \lambda_j f_j\right) \le \frac{1}{k}\left(\text{diam}\mathcal{F}\right)^2.$$

Thus at least one realization of $\frac{1}{k}\sum_{i=1}^k Y_i$ has a distance at most $k^{-1/2}\text{diam}\mathcal{F}$ to $\sum \lambda_i f_i$. Since all realizations of $\frac{1}{k}\sum_{i=1}^k Y_i$ has the form $\frac{1}{k}\sum_{i=1}^k f_{j_k}$, there are at most $\binom{n+k-1}{k}$ of such forms. Thus

$$\begin{aligned} N(k^{-1/2}\text{diam}\mathcal{F}, \text{conv}\mathcal{F}, \|\cdot\|_2) &\le \binom{n+k-1}{k} \\ \le \frac{(k+n)^{k+n}}{k^k n^n} &= \left(\frac{k+n}{k}\right)^k \left(\frac{k+n}{n}\right)^n \\ \le e^k \left(\frac{k+n}{k}\right)^k &= (e + en\epsilon^2)^{2/\epsilon^2} \end{aligned}$$

□

By triangle inequality and definition of $\mathcal{G}_n$, $\text{diam}\mathcal{G}_n = \sup_{g_1, g_2 \in \mathcal{G}_n} \|g_1 - g_2\|_2 \le 2 \cdot Lm^{-1/V}$. Let $\epsilon \cdot \text{diam}\mathcal{G}_n = \epsilon \cdot 2Lm^{-1/V} = \frac{1}{2}C_1 L n^{-W}$. It follows that $\epsilon = \frac{1}{4}C_1 m^{1/V} \cdot n^{-W}$, and

$$\begin{aligned} N(\text{conv}\mathcal{G}_n, \epsilon\text{diam}\mathcal{G}_n, \|\cdot\|_2) &\le \left(e + en \cdot \frac{1}{16}C_1^2 m^{2/V} \cdot n^{-2W}\right)^{32 \cdot C_1^{-2} m^{2/V} n^{2 \cdot W}} \\ &= \left(e + \frac{e}{16}C_1^2 d^{-2/V}\right)^{32 \cdot C_1^{-2} d^{2/V} n} \end{aligned}$$

By definition of $\mathcal{F}_m$ and and induction assumption, $\log N(\text{conv}\mathcal{F}_m, C_1 L \cdot m^{-W}, \|\cdot\|_2) \le D_1 \cdot m$. In other words, the $C_1 L \cdot m^{-W}$-net of $\text{conv}\mathcal{F}_m$ contains at most $e^{D_1 m}$ elements. This defines a partition of $\text{conv}\mathcal{F}_m$ into at most $e^{D_1 m}$ elements. Each element is isometric to a subset of a ball of radius $C_1 L m^{-W}$. Thus each set can be partitioned into $\left(\frac{3C_1 L m^{-W}}{\frac{1}{2}C_1 L n^{-W}}\right)^m = \left(6d^W\right)^{n/d}$ sets of diameter at most $\frac{1}{2}C_1 L n^{-W}$ according to the following lemma.

**Lemma 18.3.** *The packing number of a ball of radius $R$ in $\mathbb{R}^d$ satisfies $D(B(0,r), \epsilon, \|\cdot\|) \le \left(\frac{3R}{\epsilon}\right)^d$ for the usual norm, where $0 < \epsilon \le R$.*

As a result, the $C_1 L n^{-W}$-net of $\text{conv}\mathcal{F}_n$ has at most $e^{D_1 n/d}\left(6d^W\right)^{n/d}\left(e + eC_1^2 d^{-2/V}\right)^{8d^{2/V}C_1^{-2}n}$ elements. This can be upper-bounded by $e^n$ by choosing $C_1$ and $d$ depending only on $V$, and $D_1 = 1$.

For $k > 1$, construct $\mathcal{G}_{n,k}$ such that $\text{conv}\mathcal{F}_{nk^q} \subset \text{conv}\mathcal{F}_{n(k-1)^q} + \text{conv}\mathcal{G}_{n,k}$ in a similar way as before. $\mathcal{G}_{n,k}$ contains at most $nk^q$ elements, and each has a norm smaller than $L\left(n\left(k-1\right)^q\right)^{-1/V}$. To bound the cardinality of a $Lk^{-2}n^{-W}$-net, we set $\epsilon \cdot 2L\left(n\left(k-1\right)^q\right)^{-1/V} = Lk^{-2}n^{-W}$, get $\epsilon = \frac{1}{2}n^{-1/2}\left(k-1\right)^{q/V}k^{-2}$,

and

$$\begin{aligned} N(\text{conv}\mathcal{G}_{n,k}, \epsilon \text{diam}\mathcal{G}_{n,k}, \|\cdot\|_2) &\leq \left(e + enk^q\epsilon^2\right)^{2/\epsilon^2} \Rightarrow \\ N(\text{conv}\mathcal{G}_{n,k}, \epsilon \text{diam}\mathcal{G}_{n,k}, \|\cdot\|_2) &\leq \left(e + \frac{e}{4}k^{-4+q+2q/V}\right)^{8\cdot n\cdot k^4(k-1)^{-2q/V}} \end{aligned}$$

. As a result, we get

$$\begin{aligned} C_k &= C_{k-1} + \frac{1}{k^2} \\ D_k &= D_{k-1} + 8k^4(k-1)^{-2q/V}\log(e + \frac{e}{4}k^{-4+q+2q/V}). \end{aligned}$$

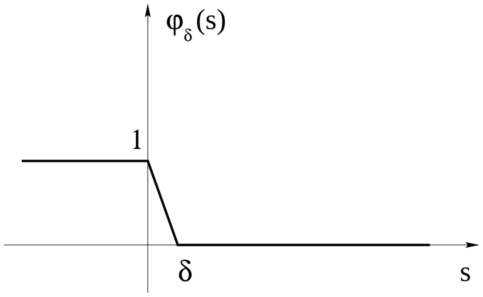
For $2q/V - 4 \geq 2$, the resulting sequences $C_k$ and $D_k$ are bounded. □

In a classification setup, we are given $\{(x_i, y_i) : x_i \in \mathcal{X}, y_i \in \{-1,+1\}\}_{i=1,\cdots,n}$, and are required to construct a classifier $y = \mathrm{sign}(f(x))$ with minimum testing error. For any $x$, the term $y \cdot f(x)$ is called **margin** can be considered as the confidence of the prediction made by $\mathrm{sign}(f(x))$. Classifiers like SVM and AdaBoost are all **maximal margin classifiers**. Maximizing margin means, penalizing small margin, controling the complexity of all possible outputs of the algorithm, or controling the generalization error.

We can define $\phi_\delta(s)$ as in the following plot, and control the error $\mathbb{P}(y \cdot f(x) \le 0)$ in terms of $\mathbb{E}\phi_\delta(y \cdot f(x))$:

$$\begin{aligned}
\mathbb{P}(y\cdot f(x)\le 0) &= \mathbb{E}_{x,y} I(y\cdot f(x)\le 0)\\
&\le \mathbb{E}_{x,y}\phi_\delta(y\cdot f(x))\\
&= \mathbb{E}\phi_\delta(y\cdot f(x))\\
&= \underbrace{\mathbb{E}_n\phi_\delta(y\cdot f(x))}_{\text{observed error}} + \underbrace{(\mathbb{E}(y\cdot f(x)) - \mathbb{E}_n\phi_\delta(y\cdot f(x)))}_{\text{generalization capability}},
\end{aligned}$$

where $\mathbb{E}_n\phi_\delta(y\cdot f(x)) \overset{\triangle}{=} \frac{1}{n}\sum_{i=1}^n \phi_\delta(y\cdot f(x))$.



Let us define $\phi_\delta(y\mathcal{F}) \overset{\triangle}{=} \{\phi_\delta(y\cdot f(x)) : f \in \mathcal{F}\}$. The function $\phi_\delta$ satisfies Lipschetz condition $|\phi_\delta(a) - \phi_\delta(b)| \le \frac{1}{\delta}|a-b|$. Thus given any $\{z_i = (x_i, y_i)\}_{i=1,\cdots,n}$,

$$\begin{aligned}
d_z(\phi_\delta(y\cdot f(x)), \phi_\delta(y\cdot g(x))) &= \left(\frac{1}{n}\sum_{i=1}^n (\phi_\delta(y_i f(x_i)) - \phi_\delta(y_i\cdot g(x_i)))^2\right)^{1/2} && \text{,definition of } d_z\\
&\le \frac{1}{\delta}\left(\frac{1}{n}\sum_{i=1}^n (y_i f(x_i) - y_i\cdot g(x_i))^2\right)^{1/2} && \text{,Lipschetz condition}\\
&= \frac{1}{\delta} d_x(f(x), g(x)) \text{ ,definition of } d_x,
\end{aligned}$$

and the packing numbers for $\phi_\delta(y\mathcal{F})$ and $\mathcal{F}$ satisfies inequality $D(\phi_\delta(y\mathcal{F}), \epsilon, d_z) \le D(\mathcal{F}, \epsilon\cdot\delta, d_x)$.

Recall that for a VC-subgraph class $\mathcal{H}$, the packing number satisfies $D(\mathcal{H}, \epsilon, d_x) \le C(\frac{1}{\epsilon})^V$, where $C$ is a constant, and $V$ is a constant. For its corresponding VC-hull class, there exists $K(C,V)$, such that $\log D(\mathcal{F} = \mathrm{conv}(\mathcal{H}), \epsilon, d_x) \le K(\frac{1}{\epsilon})^{\frac{2V}{V+2}}$. Thus $\log D(\phi_\delta(y\mathcal{F}), \epsilon, d_z) \le \log D(\mathcal{F}, \epsilon\cdot\delta, d_x) \le K(\frac{1}{\epsilon\cdot\delta})^{\frac{2V}{V+2}}$.

On the other hand, for a VC-subgraph class $\mathcal{H}$, $\log D(\mathcal{H}, \epsilon, d_x) \le KV\log\frac{2}{\epsilon}$, where $V$ is the VC dimension of $\mathcal{H}$. We proved that $\log D(\mathcal{F}_d = \mathrm{conv}_d\mathcal{H}, \epsilon, d_x) \le K\cdot V\cdot d\log\frac{2}{\epsilon}$. Thus $\log D(\phi_\delta(y\mathcal{F}_d), \epsilon, d_x) \le K\cdot V\cdot d\log\frac{2}{\epsilon\delta}$.

*Remark* 19.1. For a VC-subgraph class $\mathcal{H}$, let $V$ is the VC dimension of $\mathcal{H}$. The packing number satisfies $D(\mathcal{H},\epsilon,d_x) \le \left(\frac{k}{\epsilon}\log\frac{k}{\epsilon}\right)^V$. D Haussler (1995) also proved the following two inequalities related to the packing number: $D(\mathcal{H},\epsilon,\|\cdot\|_1) \le \left(\frac{k}{\epsilon}\right)^V$, and $D(\mathcal{H},\epsilon,d_x) \le K\left(\frac{1}{\epsilon}\right)^V$.

Since conv($\mathcal{H}$) satisfies the **uniform entroy condition** (Lecture 16) and $f \in [-1,1]^{\mathcal{X}}$, with a probability of at least $1-e^{-u}$,

$$\begin{aligned}
\mathbb{E}\phi_\delta(y\cdot f(x)) - \mathbb{E}_n\phi_\delta(y\cdot f(x)) &\le \frac{K}{\sqrt{n}}\int_0^{\sqrt{\mathbb{E}\phi_\delta}}\sqrt{\left(\frac{1}{\epsilon\cdot\delta}\right)^{\frac{2V}{V+2}}}d\epsilon + K\sqrt{\frac{\mathbb{E}\phi_\delta\cdot u}{n}} \\
&= Kn^{-\frac{1}{2}}\delta^{-\frac{V}{V+2}}\left(\mathbb{E}\phi_\delta\right)^{\frac{1}{V+2}} + K\sqrt{\frac{\mathbb{E}\phi_\delta\cdot u}{n}} \qquad (19.1)
\end{aligned}$$

for all $f \in \mathcal{F} = \text{conv}\mathcal{H}$. The term $\mathbb{E}\phi_\delta$ to estimate appears in both sides of the above inequality. We give a bound $\mathbb{E}\phi_\delta \le x^*(\mathbb{E}_n\phi_\delta, n, \delta)$ as the following. Since

$$\begin{aligned}
&\mathbb{E}\phi_\delta \le \mathbb{E}_n\phi_\delta \\
&\mathbb{E}\phi_\delta \le Kn^{-\frac{1}{2}}\delta^{-\frac{V}{V+2}}\left(\mathbb{E}\phi_\delta\right)^{\frac{1}{V+2}} \quad\Rightarrow\quad \mathbb{E}\phi_\delta \le Kn^{-\frac{1}{2}\frac{V+2}{V+1}}\delta^{-\frac{V}{V+1}} \\
&\mathbb{E}\phi_\delta \le K\sqrt{\frac{\mathbb{E}\phi_\delta\cdot u}{n}} \quad\Rightarrow\quad \mathbb{E}\phi_\delta \le K\frac{u}{n},
\end{aligned}$$

It follows that with a probability of at least $1-e^{-u}$,

$$\mathbb{E}\phi_\delta \le K\cdot\left(\mathbb{E}_n\phi_\delta + n^{-\frac{1}{2}\frac{V+2}{V+1}}\delta^{-\frac{V}{V+1}} + \frac{u}{n}\right) \qquad (19.2)$$

for some constant $K$. We proceed to bound $\mathbb{E}\phi_\delta$ for $\delta \in \{\delta_k = \exp(-k) : k \in \mathbb{N}\}$. Let $\exp(-u_k) = \left(\frac{1}{k+1}\right)^2 e^{-u}$ , it follows that $u_k = u + 2\cdot\log(k+1) = u + 2\cdot\log(\log\frac{1}{\delta_k}+1)$. Thus with a probability of at least $1-\sum_{k\in\mathbb{N}}\exp(-u_k) = 1-\sum_{k\in\mathbb{N}}\left(\frac{1}{k+1}\right)^2 e^{-u} = 1-\frac{\pi^2}{6}\cdot e^{-u} < 1-2\cdot e^{-u}$,

$$\begin{aligned}
\mathbb{E}\phi_{\delta_k}(y\cdot f(x)) &\le K\cdot\left(\mathbb{E}_n\phi_{\delta_k}(y\cdot f(x)) + n^{-\frac{1}{2}\frac{V+2}{V+1}}\delta_k^{-\frac{V}{V+1}} + \frac{u_k}{n}\right) \\
&= K\cdot\left(\mathbb{E}_n\phi_{\delta_k}(y\cdot f(x)) + n^{-\frac{1}{2}\frac{V+2}{V+1}}\delta_k^{-\frac{V}{V+1}} + \frac{u+2\cdot\log(\log\frac{1}{\delta_k}+1)}{n}\right) \qquad (19.3)
\end{aligned}$$

for all $f \in \mathcal{F}$ and all $\delta_k \in \{\delta_k : k \in \mathbb{N}\}$. Since $\mathbb{P}(y\cdot f(x) \le 0) = \mathbb{E}_{x,y}I(y\cdot f(x) < 0) \le \mathbb{E}_{x,y}\phi_\delta(y\cdot f(x))$, and $\mathbb{E}_n\phi_\delta(y\cdot f(x)) = \frac{1}{n}\sum_{i=1}^n\phi_\delta(y_i\cdot f(x_i)) \le \frac{1}{n}\sum_{i=1}^n I(y_i\cdot f(x_i) \le \delta) = \mathbb{P}_n(y_i\cdot f(x_i) \le \delta)$, with probability at least $1-2\cdot e^{-u}$,

$$\mathbb{P}(y\cdot f(x)) \le 0) \le K\cdot\inf_\delta\left(\mathbb{P}_n(y\cdot f(x) \le \delta) + n^{-\frac{V+2}{2(V+1)}}\delta^{-\frac{V}{V+1}} + \frac{u}{n} + \frac{2\log(\log\frac{1}{\delta}+1)}{n}\right).$$

As in the previous lecture, let $\mathcal{H} = \{h : \mathcal{X} \mapsto [-1, 1]\}$ be a VC-subgraph class and $f \in \mathcal{F} = \text{conv } \mathcal{H}$. The classifier is $\text{sign}(f(x))$. The set

$$\{y \neq \text{sign}(f(x))\} = \{yf(x) \leq 0\}$$

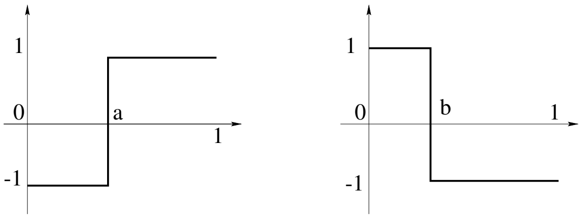
is the set of misclassified examples and $\mathbb{P}\left(yf(x) \leq 0\right)$ is the misclassification error.

Assume the examples are labeled according to $C_0 = \{x \in \mathcal{X} : y = 1\}$. Let $C = \{\text{sign}(f(x)) > 0\}$. Then $C_0 \triangle C$ are misclassified examples.

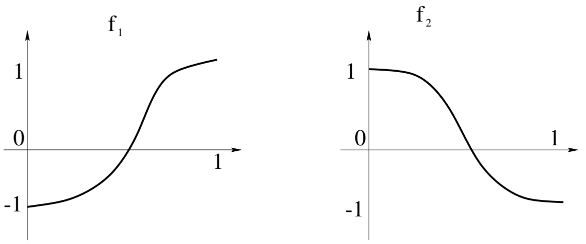
$$\mathbb{P}\left(C\triangle C_0\right) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \in C\triangle C_0) + \underbrace{\mathbb{P}\left(C\triangle C_0\right) - \frac{1}{n}\sum_{i=1}^{n} I(x_i \in C\triangle C_0)}_{\text{small. estimate uniformly over sets } c} .$$

For voting classifiers, the collection of sets $\mathcal{C}$ can be ”very large”.

**Example 20.1.** Let $\mathcal{H}$ be the class of simple step-up and step-down functions on the $[0, 1]$ interval, parametrized by $a$ and $b$.



Then $VC(\mathcal{H}) = 2$. Let $\mathcal{F} = \text{conv } \mathcal{H}$. First, rescale the functions: $f = \sum_{i=1}^{T} \lambda_i h_i = 2\sum_{i=1}^{T} \lambda_i \left(\frac{h_i+1}{2}\right) - 1 = 2f' - 1$ where $f' = \sum_{i=1}^{T} \lambda_i h_i'$, $h_i' = \frac{h_i+1}{2}$. We can generate any non-decreasing function $f'$ such that $f'(0) = 0$ and $f'(1) = 1$. Similarly, we can generate any non-increasing $f'$ such that $f'(0) = 1$ and $f'(1) = 0$. Rescaling back to $f$, we can get any non-increasing and non-decreasing functions of the form

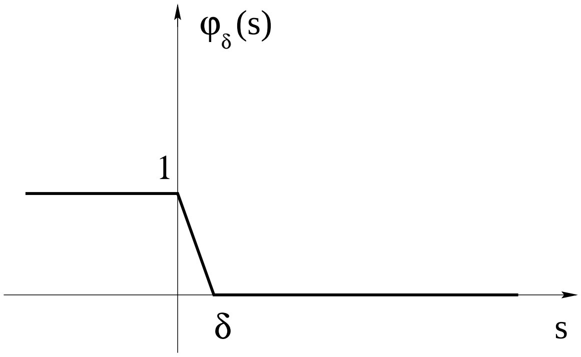


Any function with sum of jumps less than 1 can be written as $f = \frac{1}{2}(f_1 + f_2)$. Hence, we can generate basically all sets by $\{f(x) > 0\}$, i.e. conv $\mathcal{H}$ is bad.

Recall that $\mathbb{P}\left(yf(x) \leq 0\right) = \mathbb{E}I(yf(x) \leq 0)$. Define function $\varphi_\delta(s)$ as follows:

Then,

$$I(s \leq 0) \leq \varphi_\delta\left(s\right) \leq I(s \leq \delta).$$



Hence,

$$\begin{aligned}\mathbb{P}\left(yf(x)\leq 0\right) &\leq \mathbb{E}\varphi_\delta\left(yf(x)\right)\\ &= \frac{1}{n}\sum_{i=1}^n \varphi_\delta\left(y_i f(x_i)\right) + \left(\mathbb{E}\varphi_\delta\left(yf(x)\right) - \frac{1}{n}\sum_{i=1}^n \varphi_\delta\left(y_i f(x_i)\right)\right)\\ &\leq \frac{1}{n}\sum_{i=1}^n I(y_i f(x_i)\leq\delta) + \left(\mathbb{E}\varphi_\delta\left(yf(x)\right) - \frac{1}{n}\sum_{i=1}^n \varphi_\delta\left(y_i f(x_i)\right)\right)\end{aligned}$$

By going from $\frac{1}{n}\sum_{i=1}^n I(y_i f(x_i)\leq 0)$ to $\frac{1}{n}\sum_{i=1}^n I(y_i f(x_i)\leq\delta)$, we are penalizing small confidence predictions. The margin $yf(x)$ is a measure of the confidence of the prediction.

For the sake of simplicity, denote $\mathbb{E}\varphi_\delta = \mathbb{E}\varphi_\delta\left(yf(x)\right)$ and $\bar{\varphi_\delta} = \frac{1}{n}\sum_{i=1}^n \varphi_\delta\left(y_i f(x_i)\right)$.

**Lemma 20.2.** *Let $\mathcal{F}_d = conv_d\ \mathcal{H} = \{\sum_{i=1}^d \lambda_i h_i, h_i\in\mathcal{H}\}$ and fix $\delta\in(0,1]$. Then*

$$\mathbb{P}\left(\forall f\in\mathcal{F}_d,\ \frac{\mathbb{E}\varphi_\delta - \bar{\varphi_\delta}}{\sqrt{\mathbb{E}\varphi_\delta}} \leq K\left(\sqrt{\frac{dV\log\frac{n}{\delta}}{n}} + \sqrt{\frac{t}{n}}\right)\right) \geq 1 - e^{-t}.$$

*Proof.* Denote

$$\varphi_\delta\left(y\mathcal{F}_d(x)\right) = \{\varphi_\delta\left(yf(x)\right), f\in\mathcal{F}_d\}.$$

Note that $\varphi_\delta\left(yf(x)\right) : \mathcal{X}\times\mathcal{Y}\mapsto[0,1]$.

For any $n$, take any possible points $(x_1,y_1),\ldots,(x_n,y_n)$. Since

$$|\varphi_\delta\left(s\right) - \varphi_\delta\left(t\right)| \leq \frac{1}{\delta}|s-t|,$$

we have

$$\begin{aligned}
d_{x,y}\left(\varphi_\delta\left(yf(x)\right),\varphi_\delta\left(yg(x)\right)\right) &= \left(\frac{1}{n}\sum_{i=1}^{n}\left(\varphi_\delta\left(y_if(x_i)\right)-\varphi_\delta\left(y_ig(x_i)\right)\right)^2\right)^{1/2}\\
&\le \left(\frac{1}{\delta^2}\frac{1}{n}\sum_{i=1}^{n}\left(y_if(x_i)-y_ig(x_i)\right)^2\right)^{1/2}\\
&= \frac{1}{\delta}\left(\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i)-g(x_i)\right)^2\right)^{1/2}\\
&= \frac{1}{\delta}d_x(f,g)
\end{aligned}$$

where $f,g\in\mathcal{F}_d$.

Choose $\varepsilon\cdot\delta$-packing of $\mathcal{F}_d$ so that

$$d_{x,y}\left(\varphi_\delta\left(yf(x)\right),\varphi_\delta\left(yg(x)\right)\right)\le\frac{1}{\delta}d_x(f,g)\le\varepsilon.$$

Hence,

$$\mathcal{N}(\varphi_\delta\left(y\mathcal{F}_d(x)\right),\varepsilon,d_{x,y})\le\mathcal{D}(\mathcal{F}_d,\varepsilon\delta,d_x)$$

and

$$\log\mathcal{N}(\varphi_\delta\left(y\mathcal{F}_d(x)\right),\varepsilon,d_{x,y})\le\log\mathcal{D}(\mathcal{F}_d,\varepsilon\delta,d_x)\le KdV\log\frac{2}{\varepsilon\delta}.$$

We get

$$\log\mathcal{D}(\varphi_\delta\left(y\mathcal{F}_d\right),\varepsilon/2,d_{x,y})\le KdV\log\frac{2}{\varepsilon\delta}.$$

So, we can choose $f_1,\ldots,f_D$, $D=\mathcal{D}(\mathcal{F}_d,\varepsilon\delta,d_x)$ such that for any $f\in\mathcal{F}_d$ there exists $f_i$, $d_x(f,f_i)\le\varepsilon\delta$. Hence,

$$d_{x,y}(\varphi_\delta\left(yf(x)\right),\varphi_\delta\left(yf_i(x)\right))\le\varepsilon$$

and $\varphi_\delta\left(yf_1(x)\right),\ldots,\varphi_\delta\left(yf_D(x)\right)$ is an $\varepsilon$-cover of $\varphi_\delta\left(y\mathcal{F}_d(x)\right)$. □

We continue to prove the lemma from Lecture 20:

**Lemma 21.1.** *Let* $\mathcal{F}_d = conv_d\ \mathcal{H} = \{\sum_{i=1}^d \lambda_i h_i, h_i \in \mathcal{H}\}$ *and fix* $\delta \in (0,1]$. *Then*

$$\mathbb{P}\left(\forall f \in \mathcal{F}_d,\ \frac{\mathbb{E}\varphi_\delta - \bar{\varphi}_\delta}{\sqrt{\mathbb{E}\varphi_\delta}} \leq K\left(\sqrt{\frac{dV\log\frac{n}{\delta}}{n}} + \sqrt{\frac{t}{n}}\right)\right) \geq 1 - e^{-t}.$$

*Proof.* We showed that

$$\log \mathcal{D}(\varphi_\delta\left(y\mathcal{F}_d\right), \varepsilon/2, d_{x,y}) \leq K dV \log\frac{2}{\varepsilon\delta}.$$

By the result of Lecture 16,

$$\mathbb{E}\varphi_\delta\left(yf(x)\right) - \frac{1}{n}\sum_{i=1}^n \varphi_\delta\left(y_i f(x_i)\right) \leq \frac{k}{\sqrt{n}}\int_0^{\sqrt{\mathbb{E}\varphi_\delta}} \log^{1/2}\mathcal{D}(\varphi_\delta\left(y\mathcal{F}_d(x)\right), \varepsilon)d\varepsilon + \sqrt{\frac{t\mathbb{E}\varphi_\delta}{n}}$$

with probability at least $1 - e^{-t}$. We have

$$\begin{aligned}
\frac{k}{\sqrt{n}}\int_0^{\sqrt{\mathbb{E}\varphi_\delta}} \log^{1/2}\mathcal{D}(\varphi_\delta\left(y\mathcal{F}_d(x)\right), \varepsilon)d\varepsilon &\leq \frac{k}{\sqrt{n}}\int_0^{\sqrt{\mathbb{E}\varphi_\delta}} \sqrt{dV\log\frac{2}{\varepsilon\delta}}d\varepsilon \\
&= \frac{k}{\sqrt{n}}\frac{2}{\delta}\int_0^{\delta\sqrt{\mathbb{E}\varphi_\delta}/2} \sqrt{dV}\sqrt{\log\frac{1}{x}}dx \\
&\leq \frac{k}{\sqrt{n}}\frac{2}{\delta}\sqrt{dV}2\frac{\delta}{2}\sqrt{\mathbb{E}\varphi_\delta}\sqrt{\log\frac{2}{\delta\sqrt{\mathbb{E}\varphi_\delta}}}
\end{aligned}$$

where we have made a change of variables $\frac{2}{\varepsilon\delta} = x$, $\varepsilon = \frac{2x}{\delta}$. Without loss of generality, assume $\mathbb{E}\varphi_\delta \geq 1/n$. Otherwise, we're doing better than in Lemma: $\frac{\mathbb{E}}{\sqrt{\mathbb{E}}} \leq \sqrt{\frac{\log n}{n}} \Rightarrow \mathbb{E} \leq \frac{\log n}{n}$. Hence,

$$\begin{aligned}
\frac{k}{\sqrt{n}}\int_0^{\sqrt{\mathbb{E}\varphi_\delta}} \log^{1/2}\mathcal{D}(\varphi_\delta\left(y\mathcal{F}_d(x)\right), \varepsilon)d\varepsilon &\leq K\sqrt{\frac{dV\mathbb{E}\varphi_\delta}{n}\log\frac{2\sqrt{n}}{\delta}} \\
&\leq K\sqrt{\frac{dV\mathbb{E}\varphi_\delta}{n}\log\frac{n}{\delta}}
\end{aligned}$$

So, with probability at least $1 - e^{-t}$,

$$\mathbb{E}\varphi_\delta\left(yf(x)\right) - \frac{1}{n}\sum_{i=1}^n \varphi_\delta\left(y_i f(x_i)\right) \leq K\sqrt{\frac{dV\mathbb{E}\varphi_\delta\left(yf(x)\right)}{n}\log\frac{n}{\delta}} + \sqrt{\frac{t\mathbb{E}\varphi_\delta\left(yf(x)\right)}{n}}$$

which concludes the proof. □

The above lemma gives a result for a fixed $d \geq 1$ and $\delta \in (0,1]$. To obtain a uniform result, it's enough to consider $\delta \in \Delta = \{2^{-k}, k \geq 1\}$ and $d \in \{1, 2, \ldots\}$. For a fixed $\delta$ and $d$, use the Lemma above with $t_{\delta,d}$ defined by $e^{-t_{\delta,d}} = e^{-t}\frac{6\delta}{d^2\pi^2}$. Then

$$\mathbb{P}\left(\forall f \in \mathcal{F}_d,\ \ldots + \sqrt{\frac{t_{\delta,d}}{n}}\right) \geq 1 - e^{-t_{\delta,d}} = 1 - e^{-t}\frac{6\delta}{d^2\pi^2}$$

and

$$\mathbb{P}\left(\bigcup_{d,\delta}\left\{\forall f \in \mathcal{F}_d,\ \ldots + \sqrt{\frac{t_{\delta,d}}{n}}\right\}\right) \geq 1 - \sum_{\delta,d} e^{-t}\frac{6\delta}{d^2\pi^2} = 1 - e^{-t}.$$

Since $t_{\delta,d} = t + \log \frac{d^2\pi^2}{6\delta}$,

$$\begin{aligned}
\forall f \in \mathcal{F}_d,\ \frac{\mathbb{E}\varphi_\delta - \bar{\varphi_\delta}}{\sqrt{\mathbb{E}\varphi_\delta}} &\leq K\left(\sqrt{\frac{dV\log\frac{n}{\delta}}{n}} + \sqrt{\frac{t+\log\frac{d^2\pi^2}{6\delta}}{n}}\right)\\
&\leq K\left(\sqrt{\frac{dV\log\frac{n}{\delta}}{n}} + \sqrt{\log\frac{d^2\pi^2}{6\delta}} + \sqrt{\frac{t}{n}}\right)\\
&\leq K'\left(\sqrt{\frac{dV\log\frac{n}{\delta}}{n}} + \sqrt{\frac{t}{n}}\right)
\end{aligned}$$

since $\log \frac{d^2\pi^2}{6\delta}$, the penalty for union-bound, is much smaller than $\sqrt{\frac{dV\log\frac{n}{\delta}}{n}}$.

Recall the bound on the misclassification error

$$\mathbb{P}\left(yf(x)\leq 0\right) \leq \frac{1}{n}\sum_{i=1}^{n} I(y_i f(x_i)\leq\delta) + \left(\mathbb{E}\varphi_\delta\left(yf(x)\right) - \frac{1}{n}\sum_{i=1}^{n}\varphi_\delta\left(y_i f(x_i)\right)\right).$$

If

$$\frac{\mathbb{E}\varphi_\delta - \frac{1}{n}\sum_{i=1}^{n}\varphi_\delta}{\sqrt{\mathbb{E}\varphi_\delta}} \leq \varepsilon,$$

then

$$\mathbb{E}\varphi_\delta - \varepsilon\sqrt{\mathbb{E}\varphi_\delta} - \frac{1}{n}\sum_{i=1}^{n}\varphi_\delta \leq 0.$$

Hence,

$$\begin{aligned}
\sqrt{\mathbb{E}\varphi_\delta} &\leq \frac{\varepsilon}{2} + \sqrt{\left(\frac{\varepsilon}{2}\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\varphi_\delta}\\
\mathbb{E}\varphi_\delta &\leq 2\left(\frac{\varepsilon}{2}\right)^2 + 2\frac{1}{n}\sum_{i=1}^{n}\varphi_\delta.
\end{aligned}$$

The bound becomes

$$\mathbb{P}\left(yf(x)\leq 0\right) \leq K\left(\frac{1}{n}\sum_{i=1}^{n} I(y_i f(x_i)\leq\delta) + \underbrace{\frac{dV}{n}\log\frac{n}{\delta}}_{(*)} + \frac{t}{n}\right)$$

where $K$ is a rough constant.

(*) not satisfactory because in boosting the bound should get better when the number of functions grows. We prove a better bound in the next lecture.

**Theorem 22.1.** *With probability at least $1-e^{-t}$, for any $T \geq 1$ and any $f = \sum_{i=1}^T \lambda_i h_i$,*

$$\mathbb{P}\left(yf(x) \leq 0\right) \leq \inf_{\delta \in (0,1)} \left(\varepsilon + \sqrt{\mathbb{P}_n\left(yf(x) \leq \delta\right) + \varepsilon^2}\right)^2$$

*where* $\varepsilon = \varepsilon(\delta) = K\left(\sqrt{\frac{V \min(T, (\log n)/\delta^2) \log \frac{n}{\delta}}{n}} + \sqrt{\frac{t}{n}}\right)$.

Here we used the notation $\mathbb{P}_n\left(C\right) = \frac{1}{n}\sum_{i=1}^n I(x_i \in C)$.

Remark:

$$\mathbb{P}\left(yf(x) \leq 0\right) \leq \inf_{\delta \in (0,1)} K\left(\underbrace{\mathbb{P}_n\left(yf(x) \leq \delta\right)}_{\text{inc. with } \delta} + \underbrace{\frac{V \min(T, (\log n)/\delta^2) \log \frac{n}{\delta}}{n}}_{\text{dec. with } \delta} + \frac{t}{n}\right).$$

*Proof.* Let $f = \sum_{i=1}^T \lambda_i h_i$, $g = \frac{1}{k}\sum_{j=1}^k Y_j$, where

$$\mathbb{P}\left(Y_j = h_i\right) = \lambda_i \quad\text{and}\quad \mathbb{P}\left(Y_j = 0\right) = 1 - \sum_{i=1}^T \lambda_i$$

as in Lecture 17. Then $\mathbb{E}Y_j(x) = f(x)$.

$$\begin{aligned}\mathbb{P}\left(yf(x) \leq 0\right) &= \mathbb{P}\left(yf(x) \leq 0, yg(x) \leq \delta\right) + \mathbb{P}\left(yf(x) \leq 0, yg(x) > \delta\right) \\ &\leq \mathbb{P}\left(yg(x) \leq \delta\right) + \mathbb{P}\left(yg(x) > \delta \mid yf(x) \leq 0\right)\end{aligned}$$

$$\mathbb{P}\left(yg(x) > \delta \;\middle|\; yf(x) \leq 0\right) = \mathbb{E}_x \mathbb{P}_Y\left(y\frac{1}{k}\sum_{j=1}^k Y_j(x) > \delta \;\middle|\; y\mathbb{E}_Y Y_j(x) \leq 0\right)$$

Shift $Y$'s to $[0,1]$ by defining $Y_j' = \frac{yY_j + 1}{2}$. Then

$$\begin{aligned}\mathbb{P}\left(yg(x) > \delta | yf(x) \leq 0\right) &= \mathbb{E}_x \mathbb{P}_Y\left(\frac{1}{k}\sum_{j=1}^k Y_j' \geq \frac{1}{2} + \frac{\delta}{2} \;\middle|\; \mathbb{E}Y_j' \leq \frac{1}{2}\right) \\ &\leq \mathbb{E}_x \mathbb{P}_Y\left(\frac{1}{k}\sum_{j=1}^k Y_j' \geq \mathbb{E}Y_1' + \frac{\delta}{2} \;\middle|\; \mathbb{E}Y_j' \leq \frac{1}{2}\right) \\ &\leq \text{(by Hoeffding's ineq.) } \mathbb{E}_x e^{-kD\left(\mathbb{E}Y_1' + \frac{\delta}{2}, \mathbb{E}Y_1'\right)} \\ &\leq \mathbb{E}_x e^{-k\delta^2/2} = e^{-k\delta^2/2}\end{aligned}$$

because $D(p,q) \geq 2(p-q)^2$ (KL-divergence for binomial variables, Homework 1) and, hence,

$$D\left(\mathbb{E}Y_1' + \frac{\delta}{2}, \mathbb{E}Y_1'\right) \geq 2\left(\frac{\delta}{2}\right)^2 = \delta^2/2.$$

We therefore obtain

$$\mathbb{P}\left(yf(x) \leq 0\right) \leq \mathbb{P}\left(yg(x) \leq \delta\right) + e^{-k\delta^2/2} \tag{22.1}$$

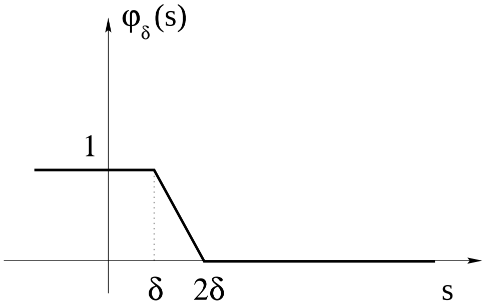
and the second term in the bound will be chosen to be equal to $1/n$.

Similarly, we can show

$$\mathbb{P}_n\left(yg(x) \le 2\delta\right) \le \mathbb{P}_n\left(yf(x) \le 3\delta\right) + e^{-k\delta^2/2}.$$

Choose $k$ such that $e^{-k\delta^2/2} = 1/n$, i.e. $k = \frac{2}{\delta^2}\log n$.

Now define $\varphi_\delta$ as follows:



Observe that

$$I(s \le \delta) \le \varphi_\delta\left(s\right) \le I(s \le 2\delta). \tag{22.2}$$

By the result of Lecture 21, with probability at least $1 - e^{-t}$, for all $k, \delta$ and any $g \in \mathcal{F}_k = \text{conv}_{\,k}(\mathcal{H})$,

$$\begin{aligned}\Phi\left(\mathbb{E}\varphi_\delta, \frac{1}{n}\sum_{i=1}^{n}\varphi_\delta\right) &= \frac{\mathbb{E}\varphi_\delta\left(yg(x)\right) - \frac{1}{n}\sum_{i=1}^{n}\varphi_\delta\left(y_i g(x_i)\right)}{\sqrt{\mathbb{E}\varphi_\delta\left(yg(x)\right)}} \\ &\le K\left(\sqrt{\frac{Vk\log\frac{n}{\delta}}{n}} + \sqrt{\frac{t}{n}}\right) \\ &= \varepsilon/2.\end{aligned}$$

Note that $\Phi(x, y) = \frac{x-y}{\sqrt{x}}$ is increasing with $x$ and decreasing with $y$.

By inequalities (22.1) and (22.2),

$$\mathbb{E}\varphi_\delta\left(yg(x)\right) \ge \mathbb{P}\left(yg(x) \le \delta\right) \ge \mathbb{P}\left(yf(x) \le 0\right) - \frac{1}{n}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\varphi_\delta\left(y_i g(x_i)\right) \le \mathbb{P}_n\left(yg(x) \le 2\delta\right) \le \mathbb{P}_n\left(yf(x) \le 3\delta\right) + \frac{1}{n}.$$

By decreasing $x$ and increasing $y$ in $\Phi(x, y)$, we decrease $\Phi(x, y)$. Hence,

$$\Phi\left(\underbrace{\mathbb{P}\left(yf(x) \le 0\right) - \frac{1}{n}}_{x}, \underbrace{\mathbb{P}_n\left(yf(x) \le 3\delta\right) + \frac{1}{n}}_{y}\right) \le K\left(\sqrt{\frac{Vk\log\frac{n}{\delta}}{n}} + \sqrt{\frac{t}{n}}\right)$$

where $k = \frac{2}{\delta^2}\log n$.

If $\frac{x-y}{\sqrt{x}} \leq \varepsilon$, we have

$$x \leq \left( \frac{\varepsilon}{2} + \sqrt{\left(\frac{\varepsilon}{2}\right)^2 + y} \right)^2$$

So,

$$\mathbb{P}\left(yf(x) \leq 0\right) - \frac{1}{n} \leq \left( \frac{\varepsilon}{2} + \sqrt{\left(\frac{\varepsilon}{2}\right)^2 + \mathbb{P}_n\left(yf(x) \leq 3\delta\right) + \frac{1}{n}} \right)^2.$$

□

Let $f = \sum_{i=1}^{T} \lambda_i h_i$, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_T \geq 0$. Rewrite $f$ as

$$f = \sum_{i=1}^{d} \lambda_i h_i + \sum_{i=d+1}^{T} \lambda_i h_i = \sum_{i=1}^{d} \lambda_i h_i + \gamma(d) \sum_{i=d+1}^{T} \lambda'_i h_i$$

where $\gamma(d) = \sum_{i=d+1}^{T} \lambda_i$ and $\lambda'_i = \lambda_i / \gamma(d)$.

Consider the following random approximation of $f$,

$$g = \sum_{i=1}^{d} \lambda_i h_i + \gamma(d) \frac{1}{k} \sum_{j=1}^{k} Y_j$$

where, as in the previous lectures,

$$\mathbb{P}\left(Y_j = h_i\right) = \lambda'_i, \quad i = d+1, \ldots, T$$

for any $j = 1, \ldots, k$. Recall that $\mathbb{E} Y_j = \sum_{i=d+1}^{T} \lambda'_i h_i$.

Then

$$\begin{aligned} \mathbb{P}\left(yf(x) \leq 0\right) &= \mathbb{P}\left(yf(x) \leq 0, yg(x) \leq \delta\right) + \mathbb{P}\left(yf(x) \leq 0, yg(x) > \delta\right) \\ &\leq \mathbb{P}\left(yg(x) \leq \delta\right) + \mathbb{E}\left[\mathbb{P}_Y\left(yf(x) \leq 0, yg(x) \geq \delta \;\middle|\; (x,y)\right)\right] \end{aligned}$$

Furthermore,

$$\begin{aligned} \mathbb{P}_Y\left(yf(x) \leq 0, yg(x) \geq \delta \;\middle|\; (x,y)\right) &\leq \mathbb{P}_Y\left(yg(x) - yf(x) > \delta \;\middle|\; (x,y)\right) \\ &= \mathbb{P}_Y\left(\gamma(d) y \left(\frac{1}{k}\sum_{j=1}^{k} Y_j(x) - \mathbb{E}Y_1\right) \geq \delta \;\middle|\; (x,y)\right). \end{aligned}$$

By renaming $Y'_j = \frac{yY_j + 1}{2} \in [0,1]$ and applying Hoeffding's inequality, we get

$$\begin{aligned} \mathbb{P}_Y\left(\gamma(d) y \left(\frac{1}{k}\sum_{j=1}^{k} Y_j(x) - \mathbb{E}Y\right) \geq \delta \;\middle|\; (x,y)\right) &= \mathbb{P}_Y\left(\frac{1}{k}\sum_{j=1}^{k} Y'_j(x) - \mathbb{E}Y'_1 \geq \frac{\delta}{2\gamma(d)} \;\middle|\; (x,y)\right) \\ &\leq e^{-\frac{k\delta^2}{2\gamma(d)^2}}. \end{aligned}$$

Hence,

$$\mathbb{P}\left(yf(x) \leq 0\right) \leq \mathbb{P}\left(yg(x) \leq \delta\right) + e^{-\frac{k\delta^2}{2\gamma^2(d)}}.$$

If we set $e^{-\frac{k\delta^2}{2\gamma(d)^2}} = \frac{1}{n}$, then $k = \frac{2\gamma^2(d)}{\delta^2} \log n$.

We have

$$g = \sum_{i=1}^{d} \lambda_i h_i + \gamma(d) \frac{1}{k} \sum_{j=1}^{k} Y_j \in \text{conv}_{d+k} \mathcal{H},$$

$d + k = d + \frac{2\gamma^2(d)}{\delta^2} \log n$.

Define the effective dimension of $f$ as

$$e(f,\delta) = \min_{0\le d\le T}\left(d + \frac{2\gamma^2(d)}{\delta^2}\log n\right).$$

Recall from the previous lectures that

$$\mathbb{P}_n\left(yg(x)\le 2\delta\right) \le \mathbb{P}_n\left(yf(x)\le 3\delta\right) + \frac{1}{n}.$$

Hence, we have the following *margin-sparsity bound*

**Theorem 23.1.** *For $\lambda_1 \ge \ldots \lambda_T \ge 0$, we define $\gamma(d,f) = \sum_{i=d+1}^{T}\lambda_i$. Then with probability at least $1-e^{-t}$,*

$$\mathbb{P}\left(yf(x)\le 0\right) \le \inf_{\delta\in(0,1)}\left(\varepsilon + \sqrt{\mathbb{P}_n\left(yf(x)\le\delta\right)+\varepsilon^2}\right)^2$$

*where*

$$\varepsilon = K\left(\sqrt{\frac{V\cdot e(f,\delta)}{n}\log\frac{n}{\delta}} + \sqrt{\frac{t}{n}}\right)$$

**Example 23.1.** Consider the zero-error case. Define

$$\delta^* = \sup\{\delta>0, \mathbb{P}_n\left(yf(x)\le\delta\right)=0\}.$$

Hence, $\mathbb{P}_n\left(yf(x)\le\delta^*\right)=0$ for confidence $\delta^*$. Then

$$\begin{aligned}\mathbb{P}\left(yf(x)\le 0\right) \le 4\varepsilon^2 &= K\left(\frac{V\cdot e(f,\delta^*)}{n}\log\frac{n}{\delta^*} + \frac{t}{n}\right)\\ &\le K\left(\frac{V\log n}{(\delta^*)^2 n}\log\frac{n}{\delta^*} + \frac{t}{n}\right)\end{aligned}$$

because $e(f,\delta)\le\frac{2}{\delta^2}\log n$ always.

Consider the polynomial weight decay: $\lambda_i \le Ki^{-\alpha}$, for some $\alpha>1$. Then

$$\gamma(d) = \sum_{i=d+1}^{T}\lambda_i \le K\sum_{i=d+1}^{T} i^{-\alpha} \le K\int_d^\infty x^{-\alpha}dx = K\frac{1}{(\alpha-1)d^{\alpha-1}} = \frac{K_\alpha}{d^{\alpha-1}}$$

Then

$$\begin{aligned}e(f,\delta) &= \min_d\left(d + \frac{2\gamma^2(d)}{\delta^2}\log n\right)\\ &\le \min_d\left(d + \frac{K'_\alpha}{\delta^2 d^{2(\alpha-1)}}\log n\right)\end{aligned}$$

Taking derivative with respect to $d$ and setting it to zero,

$$1 - \frac{K_\alpha\log n}{\delta^2 d^{2\alpha-1}} = 0$$

we get

$$d = K_\alpha\cdot\frac{\log^{1/(2\alpha-1)} n}{\delta^{2/(2\alpha-1)}} \le K\frac{\log n}{\delta^{2/(2\alpha-1)}}.$$

Hence,

$$e(f,\delta) \leq K \frac{\log n}{\delta^{2/(2\alpha-1)}}$$

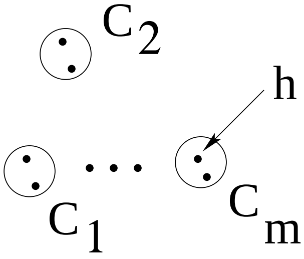
Plugging in,

$$\mathbb{P}\left(yf(x) \leq 0\right) \leq K \left( \frac{V \log n}{n(\delta^*)^{2/(2\alpha-1)}} \log \frac{n}{\delta^*} + \frac{t}{n} \right).$$

As $\alpha \to \infty$, the bound behaves like

$$\frac{V \log n}{n} \log \frac{n}{\delta^*}.$$

In this lecture, we give another example of margin-sparsity bound involved with mixture-of-experts type of models. Let $\mathcal{H}$ be a set of functions $h_i : \mathcal{X} \to [-1, +1]$ with finite VC dimension. Let $C_1, \cdots, C_m$ be partitions of $\mathcal{H}$ into $m$ clusters $\mathcal{H} = \bigcup_{i=1}^m C_i$. The elements in the convex hull conv$\mathcal{H}$ takes the form $f = \sum_{i=1}^T \lambda_i h_i = \sum_{c\in\{C_1,\cdots,C_m\}} \alpha_c \sum_{h\in c} \lambda_h^c \cdot h$, where $T \gg m$, $\sum_i \lambda_i = 1$, $\alpha_c = \sum_{h\in c} \lambda_h$, and $\lambda_h^c = \lambda_h/\alpha_c$ for $h \in c$. We can approximate $f$ by $g$ as follows. For each cluster $c$, let $\{Y_k^c\}_{k=1,\cdots,N}$ be random variables such that $\forall h \in c \subset \mathcal{H}$, we have $\mathbb{P}(Y_k^c = h) = \lambda_h^c$. Then $\mathbb{E}Y_k^c = \sum_{h\in c} \lambda_h^c \cdot h$. Let $Z_k = \sum_c \alpha_c Y_k^c$ and $g = \sum_c \alpha_c \frac{1}{N} \sum_{k=1}^N Y_k^c = \frac{1}{N}\sum_{k=1}^N Z_k$. Then $\mathbb{E}Z_k = \mathbb{E}g = f$. We define $\sigma_c^2 \overset{\triangle}{=} \text{var}(Z_k) = \sum_c \alpha_c^2 \text{var}(Y_k^c)$, where $\text{var}(Y_k^c) = \|Y_k^c - \mathbb{E}Y_k^c\|^2 = \sum_{h\in c} \lambda_h^c (h - \mathbb{E}Y_h^c)^2$. (If we define $\{Y_k\}_{k=1,\cdots,N}$ be random variables such that $\forall h \in \mathcal{H}$, $\mathbb{P}(Y_k = h) = \lambda_h$, and define $g = \frac{1}{N}\sum_{k=1}^N Y_k$, we might get much larger $\text{var}(Y_k)$).



Recall that a classifier takes the form $y = \text{sign}(f(x))$ and a classification error corresponds to $yf(x) < 0$. We can bound the error by

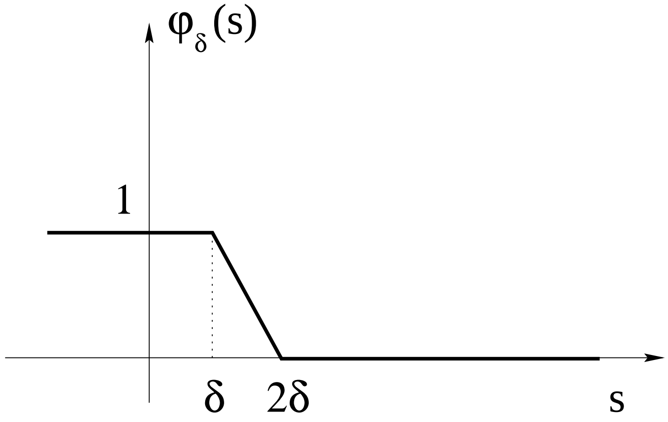
$$\mathbb{P}(yf(x) < 0) \le \mathbb{P}(yg \le \delta) + \mathbb{P}(\sigma_c^2 > r) + \mathbb{P}(yg > \delta | yf(x) \le 0, \sigma_c^2 < r). \tag{24.1}$$

The third term on the right side of inequality 24.1 can be bounded in the following way,

$$\begin{aligned}
\mathbb{P}(yg > \delta | yf(x) \le 0, \sigma_c^2 < r) &= \mathbb{P}\left(\frac{1}{N}\sum_{k=1}^N (yZ_k - \mathbb{E}yZ_k) > \delta - yf(x) | yf(x) \le 0, \sigma_c^2 < r\right) \\
&\le \mathbb{P}\left(\frac{1}{N}\sum_{k=1}^N (yZ_k - \mathbb{E}yZ_k) > \delta | yf(x) \le 0, \sigma_c^2 < r\right) \\
&\le \exp\left(-\frac{N^2\delta^2}{2N\sigma_c^2 + \frac{2}{3}N\delta\cdot 2}\right) \text{ ,Bernstein's inequality} \\
&\le \exp\left(-\min\left(\frac{N^2\delta^2}{4N\sigma_c^2}, \frac{N^2\delta^2}{\frac{8}{3}N\delta}\right)\right) \\
&\le \exp\left(-\frac{N\delta^2}{4r}\right) \text{ , for } r \text{ small enough} \\
&\overset{set}{\le} \frac{1}{n}.
\end{aligned} \tag{24.2}$$

As a result, $\forall N \ge \frac{4\cdot r}{\delta^2}\log n$, inequality 24.2 is satisfied.

To bound the first term on the right side of inequality 24.1, we note that $\mathbb{E}_{Y_1,\cdots,Y_N}\mathbb{P}(yg \le \delta) \le \mathbb{E}_{Y_1,\cdots,Y_N}\mathbb{E}\phi_\delta(yg)$ and $\mathbb{E}_n\phi_\delta(yg) \le \mathbb{P}_n(yg \le 2\delta)$ for some $\phi_\delta$:



Any realization of $g = \sum_{k=1}^{N_m} Z_k$, where $N_m$ depends on the number of clusters $(C_1, \cdots, C_m)$, is a linear combination of $h \in \mathcal{H}$, and $g \in \text{conv}_{N_m}\mathcal{H}$. According to lemma 20.2,

$$\left(\mathbb{E}\phi_\delta(yg) - \mathbb{E}_n\phi_\delta(yg)\right)/\sqrt{\mathbb{E}\phi_\delta(yg)} \quad \le \quad K\left(\sqrt{VN_m \log\frac{n}{\delta}/n} + \sqrt{u/n}\right)$$

with probability at least $1 - e^{-u}$. Using a technique developed earlier in this course, and taking the union bound over all $m$, $\delta$, we get, with probability at least $1 - Ke^{-u}$,

$$\mathbb{P}(yg \le \delta) \quad \le \quad K \inf_{m,\delta}\left(\mathbb{P}_n(yg \le 2\delta) + \frac{V \cdot N_m}{n}\log\frac{n}{\delta} + \frac{u}{n}\right).$$

(Since $\mathbb{E}\mathbb{P}_n(yg \le 2\delta) \le \mathbb{E}\mathbb{P}_n(yf(x) \le 3\delta) + \mathbb{E}\mathbb{P}_n(\sigma_c^2 \ge r) + \frac{1}{n}$ with appropriate choice of $N$, based on the same reasoning as inequality 24.1, we can also control $\mathbb{P}_n(yg \le 2\delta)$ by $\mathbb{P}_n(yf \le 3\delta)$ and $\mathbb{P}_n(\sigma_c^2 \ge r)$ probabilistically).

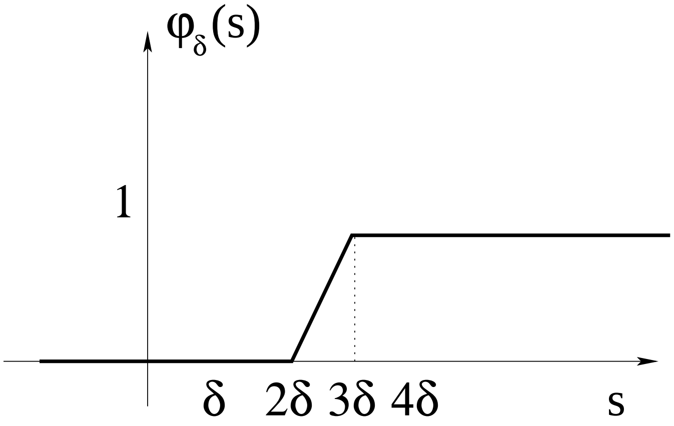
To bound the second term on the right side of inequality 24.1, we approximate $\sigma_c^2$ by
$\sigma_N^2 = \frac{1}{N}\sum_{k=1}^N \frac{1}{2}\left(Z_k^{(1)} - Z_k^{(2)}\right)^2$ where $Z_k^{(1)}$ and $Z_k^{(2)}$ are independent copies of $Z_k$ . We have

$$\begin{aligned}
\mathbb{E}_{Y^{(1,2)}_{1,\cdots,N}}\sigma_N^2 &= \sigma_c^2 \\
\text{var}_{Y^{(1,2)}_{1,\cdots,N}}\frac{1}{2}\left(Z_k^{(1)} - Z_k^{(2)}\right)^2 &= \frac{1}{4}\text{var}\left(Z_k^{(1)} - Z_k^{(2)}\right)^2 \\
&\le \frac{1}{4}\mathbb{E}\left(Z_k^{(1)} - Z_k^{(2)}\right)^4 \\
&\quad \left(-1 \le Z_k^{(1)}, Z_k^{(2)} \le 1 \text{ ,and } \left(Z_k^{(1)} - Z_k^{(2)}\right)^2 \le 4\right) \\
&\le \mathbb{E}\left(Z_k^{(1)} - Z_k^{(2)}\right)^2 \\
&= 2\sigma_c^2 \\
\text{var}_{Y^{(1,2)}_{1,\cdots,N}}\sigma_N^2 &\le 2\cdot\sigma_c^2.
\end{aligned}$$

We start with

$$\begin{aligned}\mathbb{P}_{Y_{1,\cdots,N}}(\sigma_c^2 \geq 4r) &\leq \mathbb{P}_{Y^{(1,2)}_{1,\cdots,N}}(\sigma_N^2 \geq 3r) + \mathbb{P}_{Y^{(1,2)}_{1,\cdots,N}}(\sigma_c^2 \geq 4r | \sigma_N^2 \leq 3r)\\ &\leq \mathbb{E}_{Y^{(1,2)}_{1,\cdots,N}}\phi_r\left(\sigma_N^2 \geq 3r\right) + \frac{1}{n}\end{aligned}$$

with appropriate choice of $N$, following the same line of reasoning as in inequality 24.1. We note that $\mathbb{P}_{Y_1,\cdots,Y_N}(\sigma_N^2 \geq 3r) \leq \mathbb{E}_{Y_1,\cdots,Y_N}\phi_r(\sigma_N^2)$, and $\mathbb{E}_n\phi_\delta(\sigma_N^2) \leq \mathbb{P}_n(\sigma_N^2 \geq 2r)$ for some $\phi_\delta$.



Since

$$\sigma_N^2 \in \{\frac{1}{2N}\sum_{k=1}^{N}\left(\sum_c \alpha_c\left(h_{k,c}^{(1)} - h_{k,c}^{(2)}\right)\right)^2 : h_{k,c}^{(1)}, h_{k,c}^{(2)} \in \mathcal{H}\} \subset \text{conv}_{N_m}\{h_i \cdot h_j : h_i, h_j \in \mathcal{H}\},$$

and $\log D(\{h_i \cdot h_j : h_i, h_j \in \mathcal{H}\}, \epsilon) \leq KV \log\frac{2}{\epsilon}$ by the assumption of our problem, we have $\log D(\text{conv}_{N_m}\{h_i \cdot h_j : h_i, h_j \in \mathcal{H}\}, \epsilon) \leq KV \cdot N_m \cdot \log\frac{2}{\epsilon}$ by the VC inequality, and

$$\left(\mathbb{E}\phi_r(\sigma_N^2) - \mathbb{E}_n\phi_r(\sigma_N^2)\right)/\sqrt{\mathbb{E}\phi_r(\sigma_N^2)} \leq K\left(\sqrt{V \cdot N_m \log\frac{n}{r}/n} + \sqrt{u/n}\right)$$

with probability at least $1 - e^{-u}$. Using a technique developed earlier in this course, and taking the union bound over all $m$, $\delta$, $r$, with probability at least $1 - Ke^{-u}$,

$$\mathbb{P}(\sigma_c^2 \geq 4r) \leq K\inf_{m,\delta,r}\left(\mathbb{P}_n(\sigma_N^2 \geq 2r) + \frac{1}{n} + \frac{V \cdot N_m}{n}\log\frac{n}{\delta} + \frac{u}{n}\right).$$

As a result, with probability at least $1 - Ke^{-u}$, we have

$$\mathbb{P}(yf(x) \leq 0) \leq K \cdot \inf_{r,\delta,m}\left(\mathbb{P}_n(yg \leq 2\cdot\delta) + \mathbb{P}_n(\sigma_N^2 \geq r) + \frac{V \cdot \min(r_m/\delta^2, N_m)}{n}\log\frac{n}{\delta}\log n + \frac{u}{n}\right)$$

for all $f \in \text{conv}\mathcal{H}$.

Let $Z(x_1,\ldots,x_n) : \mathcal{X}^n \mapsto \mathbb{R}$. We would like to bound $Z - \mathbb{E}Z$. We will be able to answer this question if for any $x_1,\ldots,x_n,x_1',\ldots,x_n'$,

$$|Z(x_1,\ldots,x_n) - Z(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)| \le c_i. \tag{25.1}$$

Decompose $Z - \mathbb{E}Z$ as follows

$$\begin{aligned}
Z(x_1,\ldots,x_n) - \mathbb{E}_{x'}Z(x_1',\ldots,x_n') &= (Z(x_1,\ldots,x_n) - \mathbb{E}_{x'}Z(x_1',x_2,\ldots,x_n)) \\
&+ (\mathbb{E}_{x'}Z(x_1',x_2,\ldots,x_n) - \mathbb{E}_{x'}Z(x_1',x_2',x_3,\ldots,x_n)) \\
&\ldots \\
&+ \left(\mathbb{E}_{x'}Z(x_1',\ldots,x_{n-1}',x_n) - \mathbb{E}_{x'}Z(x_1',\ldots,x_n')\right) \\
&= Z_1 + Z_2 + \ldots + Z_n
\end{aligned}$$

where

$$Z_i = \mathbb{E}_{x'}Z(x_1',\ldots,x_{i-1}',x_i,\ldots,x_n) - \mathbb{E}_{x'}Z(x_1',\ldots,x_i',x_{i+1},\ldots,x_n).$$

Assume

(1) $|Z_i| \le c_i$
(2) $\mathbb{E}_{X_i}Z_i = 0$
(3) $Z_i = Z_i(x_i,\ldots,x_n)$

**Lemma 25.1.** *For any $\lambda \in \mathbb{R}$,*

$$\mathbb{E}_{x_i}e^{\lambda Z_i} \le e^{\lambda^2 c_i^2/2}.$$

*Proof.* Take any $-1 \le s \le 1$. With respect to $\lambda$, function $e^{\lambda s}$ is convex and

$$e^{\lambda s} = e^{\lambda\left(\frac{1+s}{2}\right) + (-\lambda)\left(\frac{1-s}{2}\right)}.$$

Then $0 \le \frac{1+s}{2}, \frac{1-s}{2} \le 1$ and $\frac{1+s}{2} + \frac{1-s}{2} = 1$ and therefore

$$e^{\lambda s} \le \frac{1+s}{2}e^{\lambda} + \frac{1-s}{2}e^{-\lambda} = \frac{e^{\lambda}+e^{-\lambda}}{2} + s\frac{e^{\lambda}-e^{-\lambda}}{2} \le e^{\lambda^2/2} + s \cdot \mathrm{sh}(x)$$

using Taylor expansion. Now use $\frac{Z_i}{c_i} = s$, where, by assumption, $-1 \le \frac{Z_i}{c_i} \le 1$. Then

$$e^{\lambda Z_i} = e^{\lambda c_i \cdot \frac{Z_i}{c_i}} \le e^{\lambda^2 c_i^2/2} + \frac{Z_i}{c_i}\mathrm{sh}(\lambda c_i).$$

Since $\mathbb{E}_{x_i}Z_i = 0$,

$$\mathbb{E}_{x_i}e^{\lambda Z_i} \le e^{\lambda^2 c_i^2/2}.$$

□

We now prove McDiarmid's inequality

**Theorem 25.1.** *If condition (25.1) is satisfied,*

$$\mathbb{P}\left(Z - \mathbb{E}Z > t\right) \le e^{-\frac{t^2}{2\sum_{i=1}^n c_i^2}}.$$

*Proof.* For any $\lambda > 0$

$$\mathbb{P}\left(Z - \mathbb{E}Z > t\right) = \mathbb{P}\left(e^{\lambda(Z-\mathbb{E}Z)} > e^{\lambda t}\right) \le \frac{\mathbb{E}e^{\lambda(Z-\mathbb{E}Z)}}{e^{\lambda t}}.$$

Furthermore,

$$\begin{aligned}
\mathbb{E}e^{\lambda(Z-\mathbb{E}Z)} &= \mathbb{E}e^{\lambda(Z_1+\ldots+Z_n)} \\
&= \mathbb{E}\mathbb{E}_{x_1}e^{\lambda(Z_1+\ldots+Z_n)} \\
&= \mathbb{E}\left[e^{\lambda(Z_2+\ldots+Z_n)}\mathbb{E}_{x_1}e^{\lambda Z_1}\right] \\
&\le \mathbb{E}\left[e^{\lambda(Z_2+\ldots+Z_n)}e^{\lambda^2c_1^2/2}\right] \\
&= e^{\lambda^2c_1^2/2}\mathbb{E}\mathbb{E}_{x_2}\left[e^{\lambda(Z_2+\ldots+Z_n)}\right] \\
&= e^{\lambda^2c_1^2/2}\mathbb{E}\left[e^{\lambda(Z_3+\ldots+Z_n)}\mathbb{E}_{x_2}e^{\lambda Z_2}\right] \\
&\le e^{\lambda^2(c_1^2+c_2^2)/2}\mathbb{E}e^{\lambda(Z_3+\ldots+Z_n)} \\
&\le e^{\lambda^2\sum_{i=1}^n c_i^2/2}
\end{aligned}$$

Hence,

$$\mathbb{P}\left(Z - \mathbb{E}Z > t\right) \le e^{-\lambda t + \lambda^2\sum_{i=1}^n c_i^2/2}$$

and we minimize over $\lambda > 0$ to get the result of the theorem. □

**Example 25.1.** Let $\mathcal{F}$ be a class of functions: $\mathcal{X} \mapsto [a,b]$. Define the empirical process

$$Z(x_1,\ldots,x_n) = \sup_{f\in\mathcal{F}}\left|\mathbb{E}f - \frac{1}{n}\sum_{i=1}^n f(x_i)\right|.$$

Then, for any $i$,

$$\begin{aligned}
&|Z(x_1,\ldots,x_i',\ldots,x_n) - Z(x_1,\ldots,x_i,\ldots,x_n)| \\
&= \left|\sup_f\left|\mathbb{E}f - \frac{1}{n}\left(f(x_1)+\ldots+f(x_i')+\ldots+f(x_n)\right)\right|\right. \\
&\quad \left. - \sup_f\left|\mathbb{E}f - \frac{1}{n}\left(f(x_1)+\ldots+f(x_i)+\ldots+f(x_n)\right)\right|\right| \\
&\le \sup_{f\in\mathcal{F}}\frac{1}{n}|f(x_i) - f(x_i')| \le \frac{b-a}{n} = c_i
\end{aligned}$$

because

$$\sup_t f(t) - \sup_t g(t) \le \sup_t(f(t) - g(t))$$

and

$$|c| - |d| \leq |c - d|.$$

Thus, if $a \leq f(x) \leq b$ for all $f$ and $x$, then, setting $c_i = \frac{b-a}{n}$ for all $i$,

$$\mathbb{P}\left(Z - \mathbb{E}Z > t\right) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \frac{(b-a)^2}{n^2}}\right) = e^{-\frac{nt^2}{2(b-a)^2}}.$$

By setting $t = \sqrt{\frac{2u}{n}}(b-a)$, we get

$$\mathbb{P}\left(Z - \mathbb{E}Z > \sqrt{\frac{2u}{n}}(b-a)\right) \leq e^{-u}.$$

Let $\varepsilon_1, \ldots, \varepsilon_n$ be i.i.d. such that $\mathbb{P}\left(\varepsilon = \pm 1\right) = \frac{1}{2}$. Define

$$Z((\varepsilon_1, x_1), \ldots, (\varepsilon_n, x_n)) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right|.$$

Then, for any $i$,

$$\begin{aligned}
&|Z((\varepsilon_1, x_1), \ldots, (\varepsilon_i', x_i'), \ldots, (\varepsilon_n, x_n)) - Z((\varepsilon_1, x_1), \ldots, (\varepsilon_i, x_i), \ldots, (\varepsilon_n, x_n))| \\
&\leq \sup_{f \in \mathcal{F}} \left| \frac{1}{n}(\varepsilon_i' f(x_i') - \varepsilon_i f(x_i)) \right| \leq \frac{2M}{n} = c_i
\end{aligned}$$

where $-M \leq f(x) \leq M$ for all $f$ and $x$.

Hence,

$$\mathbb{P}\left(Z - \mathbb{E}Z > t\right) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \frac{(2M)^2}{n^2}}\right) = e^{-\frac{nt^2}{8M^2}}.$$

By setting $t = \sqrt{\frac{8u}{n}}M$, we get

$$\mathbb{P}\left(Z - \mathbb{E}Z > \sqrt{\frac{8u}{n}}M\right) \leq e^{-u}.$$

Similarly,

$$\mathbb{P}\left(\mathbb{E}Z - Z > \sqrt{\frac{8u}{n}}M\right) \leq e^{-u}.$$

Define the following processes:

$$Z(x) = \sup_{f\in\mathcal{F}} \left( \mathbb{E}f - \frac{1}{n}\sum_{i=1}^{n} f(x_i) \right)$$

and

$$R(x) = \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(x_i).$$

Assume $a \le f(x) \le b$ for all $f, x$. In the last lecture we proved $Z$ is concentrated around its expectation: with probability at least $1 - e^{-t}$,

$$Z < \mathbb{E}Z + (b-a)\sqrt{\frac{2t}{n}}.$$

Furthermore,

$$\begin{aligned}
\mathbb{E}Z(x) &= \mathbb{E}\sup_{f\in\mathcal{F}} \left( \mathbb{E}f - \frac{1}{n}\sum_{i=1}^{n} f(x_i) \right) \\
&= \mathbb{E}\sup_{f\in\mathcal{F}} \left( \mathbb{E}\left[ \frac{1}{n}\sum_{i=1}^{n} f(x_i') \right] - \frac{1}{n}\sum_{i=1}^{n} f(x_i) \right) \\
&\le \mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} (f(x_i') - f(x_i)) \\
&= \mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i (f(x_i') - f(x_i)) \\
&\le \mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(x_i') + \sup_{f\in\mathcal{F}} \left( -\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(x_i) \right) \\
&\le 2\mathbb{E}R(x).
\end{aligned}$$

Hence, with probability at least $1 - e^{-t}$,

$$Z < 2\mathbb{E}R + (b-a)\sqrt{\frac{2t}{n}}.$$

It can be shown that $R$ is also concentrated around its expectation: if $-M \le f(x) \le M$ for all $f, x$, then with probability at least $1 - e^{-t}$,

$$\mathbb{E}R \le R + M\sqrt{\frac{2t}{n}}.$$

Hence, with high probability,

$$Z(x) \le 2R(x) + 4M\sqrt{\frac{2t}{n}}.$$

**Theorem 26.1.** *If $-1 \le f \le 1$, then*

$$\mathbb{P}\left( Z(x) \le 2\mathbb{E}R(x) + 2\sqrt{\frac{2t}{n}} \right) \ge 1 - e^{-t}.$$

*If* $0 \le f \le 1$, *then*

$$\mathbb{P}\left(Z(x) \le 2\mathbb{E}R(x) + \sqrt{\frac{2t}{n}}\right) \ge 1 - e^{-t}.$$

Consider $\mathbb{E}_\varepsilon R(x) = \mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)$. Since $x_i$ are fixed, $f(x_i)$ are just vectors. Let $F \subseteq \mathbb{R}^n$, $f \in F$, where $f = (f_1, \dots, f_n)$.

Define *contraction* $\varphi_i : \mathbb{R} \mapsto \mathbb{R}$ for $i = 1, \dots, n$ such that $\varphi_i(0) = 0$ and $|\varphi_i(s) - \varphi_i(t)| \le |s - t|$.

Let $G : \mathbb{R} \mapsto \mathbb{R}$ be convex and non-decreasing.

The following theorem is called *Comparison inequality for Rademacher process.*

**Theorem 26.2.**

$$\mathbb{E}_\varepsilon G\left(\sup_{f\in F}\sum \varepsilon_i\varphi_i(f_i)\right) \le \mathbb{E}_\varepsilon G\left(\sup_{f\in F}\sum \varepsilon_i f_i\right).$$

*Proof.* It is enough to show that for $T \subseteq \mathbb{R}^2$, $t = (t_1, t_2) \in T$

$$\mathbb{E}_\varepsilon G\left(\sup_{t\in T} t_1 + \varepsilon\varphi(t_2)\right) \le \mathbb{E}_\varepsilon G\left(\sup_{t\in T} t_1 + \varepsilon t_2\right),$$

i.e. enough to show that we can erase contraction for 1 coordinate while fixing all others.

Since $\mathbb{P}(\varepsilon = \pm 1) = 1/2$, we need to prove

$$\frac{1}{2}G\left(\sup_{t\in T} t_1 + \varphi(t_2)\right) + \frac{1}{2}G\left(\sup_{t\in T} t_1 - \varphi(t_2)\right) \le \frac{1}{2}G\left(\sup_{t\in T} t_1 + t_2\right) + \frac{1}{2}G\left(\sup_{t\in T} t_1 - t_2\right).$$

Assume $\sup_{t\in T} t_1 + \varphi(t_2)$ is attained on $(t_1, t_2)$ and $\sup_{t\in T} t_1 - \varphi(t_2)$ is attained on $(s_1, s_2)$. Then

$$t_1 + \varphi(t_2) \ge s_1 + \varphi(s_2)$$

and

$$s_1 - \varphi(s_2) \ge t_1 - \varphi(t_2).$$

Again, we want to show

$$\Sigma = G(t_1 + \varphi(t_2)) + G(s_1 - \varphi(s_2)) \le G(t_1 + t_2) + G(t_1 - t_2).$$

**Case 1:** $t_2 \le 0, s_2 \ge 0$

Since $\varphi$ is a contraction, $\varphi(t_2) \le |t_2| \le -t_2$, $-\varphi(s_2) \le s_2$.

$$\begin{aligned}\Sigma = G(t_1 + \varphi(t_2)) + G(s_1 - \varphi(s_2)) &\le G(t_1 - t_2) + G(s_1 + s_2)\\ &\le G\left(\sup_{t\in T} t_1 - t_2\right) + G\left(\sup_{t\in T} t_1 + t_2\right).\end{aligned}$$

**Case 2:** $t_2 \ge 0, s_2 \le 0$

Then $\varphi(t_2) \le t_2$ and $-\varphi(s_2) \le -s_2$. Hence

$$\Sigma \le G(t_1 + t_2) + G(s_1 - s_2) \le G\left(\sup_{t\in T} t_1 + t_2\right) + G\left(\sup_{t\in T} t_1 - t_2\right).$$

**Case 3:** $t_2 \geq 0, s_2 \geq 0$

**Case 3a:** $s_2 \leq t_2$

It is enough to prove

$$G(t_1 + \varphi(t_2)) + G(s_1 - \varphi(s_2)) \leq G(t_1 + t_2) + G(s_1 - s_2).$$

Note that $s_2 - \varphi(s_2) \geq 0$ since $s_2 \geq 0$ and $\varphi$ – contraction. Since $|\varphi(s)| \leq |s|$,

$$s_1 - s_2 \leq s_1 + \varphi(s_2) \leq t_1 + \varphi(t_2),$$

where we use the fact that $t_1, t_2$ attain maximum.

Furthermore,

$$G\Big(\underbrace{(s_1 - s_2)}_{u} + \underbrace{(s_2 - \varphi(s_2))}_{x}\Big) - G\Big(s_1 - s_2\Big) \leq G\Big((t_1 + \varphi(t_2)) + (s_2 - \varphi(s_2))\Big) - G\Big(t_1 + \varphi(t_2)\Big)$$

Indeed, $\Psi(u) = G(u+x) - G(u)$ is non-decreasing for $x \geq 0$ since $\Psi'(u) = G'(u+x) - G'(u) > 0$ by convexity of $G$.

Now,

$$(t_1 + \varphi(t_2)) + (s_2 - \varphi(s_2)) \leq t_1 + t_2$$

since

$$\varphi(t_2) - \varphi(s_2) \leq |t_2 - s_2| = t_2 - s_2.$$

Hence,

$$\begin{aligned} G\Big(s_1 - \varphi(s_2)\Big) - G\Big(s_1 - s_2\Big) &= G\Big((s_1 - s_2) + (s_2 - \varphi(s_2))\Big) - G\Big(s_1 - s_2\Big) \\ &\leq G\Big(t_1 + t_2\Big) - G\Big(t_1 + \varphi(t_2)\Big). \end{aligned}$$

**Case 3a:** $t_2 \leq s_2$

$$\Sigma \leq G(s_1 + s_2) + G(t_1 - t_2)$$

Again, it's enough to show

$$G(t_1 + \varphi(t_2)) - G(t_1 - t_2) \leq G(s_1 + s_2) - G(s_1 - \varphi(s_2))$$

We have

$$t_1 - t_2 \leq t_1 - \varphi(t_2) \leq s_1 - \varphi(s_2)$$

since $s_1, s_2$ achieves maximum and since $t_2 + \varphi(t_2) \geq 0$ ($\varphi$ is a contraction and $t_2 \geq 0$).

Hence,

$$G\Big(\underbrace{(t_1 - t_2)}_{u} + \underbrace{(t_2 + \varphi(t_2))}_{x}\Big) - G\Big(t_1 - t_2\Big) \leq G\Big((s_1 - \varphi(s_2)) + (t_2 + \varphi(t_2))\Big) - G\Big(s_1 - \varphi(s_2)\Big)$$

Since

$$\varphi(t_2) - \varphi(s_2) \le |t_2 - s_2| = s_2 - t_2,$$

we get

$$\varphi(t_2) - \varphi(s_2) \le s_2 - t_2.$$

Therefore,

$$s_1 - \varphi(s_2) + (t_2 + \varphi(t_2) \le s_1 + s_2$$

and so

$$G(t_1 + \varphi(t_2)) - G(t_1 - t_2) \le G(s_1 + s_2) - G(s_1 - \varphi(s_2))$$

**Case 4:** $t_2 \le 0, s_2 \le 0$

Proved in the same way as Case 3. □

We now apply the theorem with $G(s) = (s)^+$.

**Lemma 26.1.**

$$\mathbb{E}\sup_{t\in T}\left|\sum_{i=1}^n \varepsilon_i \varphi_i(t_i)\right| \le 2\mathbb{E}\sup_{t\in T}\left|\sum_{i=1}^n \varepsilon_i t_i\right|$$

*Proof.* Note that

$$|x| = (x)^+ + (x)^- = (x)^+ + (-x)^+.$$

We apply the Contraction Inequality for Rademacher processes with $G(s) = (s)^+$.

$$\begin{aligned}\mathbb{E}\sup_{t\in T}\left|\sum_{i=1}^n \varepsilon_i \varphi_i(t_i)\right| &= \mathbb{E}\sup_{t\in T}\left(\left(\sum_{i=1}^n \varepsilon_i \varphi_i(t_i)\right)^+ + \left(\sum_{i=1}^n (-\varepsilon_i) \varphi_i(t_i)\right)^+\right)\\ &\le 2\mathbb{E}\sup_{t\in T}\left(\sum_{i=1}^n \varepsilon_i \varphi_i(t_i)\right)^+\\ &\le 2\mathbb{E}\sup_{t\in T}\left(\sum_{i=1}^n \varepsilon_i t_i\right)^+ \le 2\mathbb{E}\sup_{t\in T}\left|\sum_{i=1}^n \varepsilon_i t_i\right|.\end{aligned}$$

□

Let $\mathcal{F} \subset \{f \in [0,1]\}$ be a class of $[0,1]$ valued functions, $Z = \sup_f \left(\mathbb{E}f - \frac{1}{n}\sum_{i=1}^n f(x_i)\right)$, and $R = \sup_f \sum_f \frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)$ for any given $x_i, \cdots, x_n$ where $\epsilon_1, \cdots \epsilon_n$ are Rademarcher random variables. For any $f \in \mathcal{F}$ unknown and to be estimated, the empirical error $Z$ can be probabilistically bounded by $R$ in the following way. Using the fact that $Z \le 2R$ and by Martingale inequality, $\mathbb{P}\left(Z \le \mathbb{E}Z + \sqrt{\frac{2u}{n}}\right) \ge 1 - e^{-u}$, and $\mathbb{P}\left(\mathbb{E}R \le R + 2\sqrt{\frac{2u}{n}}\right) \ge 1 - e^{-u}$. Taking union bound, $\mathbb{P}\left(Z \le R + 5\sqrt{\frac{2u}{n}}\right) \ge 1 - 2e^{-u}$. Taking union bound again over all $(n_k)_{k>1}$ and let $\epsilon = 5\sqrt{\frac{2u}{n}}$, $\mathbb{P}\left(\forall n \in (n_k)_{k\ge 1} \forall f \in \mathcal{F}, Z \le 2R + \epsilon\right) \ge 1 - \exp\left(-\sum_k \frac{n_k \epsilon^2}{50}\right) \overset{\text{set}}{\ge} 1 - \delta$. Using big O notation, $n_k = \mathcal{O}\left(\frac{1}{\epsilon^2}\log\frac{1}{\delta^2}\right)$.

For voting algorithms, the candidate function to be estimated is a symmetric convex combination of some base functions $\mathcal{F} = \text{conv}\mathcal{H}$, where $\mathcal{H} \subset \{h \in [0,1]\}$. The trained classifier is $\text{sign}(f(x))$ where $f \in \mathcal{F}$ is our estimation, and the training error is $\mathbb{P}(yf(x))$. The training error can be bounded as the following,

$$\begin{aligned}
\mathbb{P}(yf(x) < 0) \quad & \le \quad \mathbb{E}\phi_\delta(yf(x)) \\
& \le \quad \mathbb{E}_n\phi_\delta(yf(x)) + \underbrace{\sup_{f\in\mathcal{F}}(\mathbb{E}\phi_\delta(yf(x)) - \frac{1}{n}\sum_{i=1}^n \phi_\delta(y_i f(x_i)))}_{Z} \\
& \underbrace{\le}_{\text{with probability } 1-e^{-u}} \quad \mathbb{E}_n\phi_\delta(yf(x)) + 2\cdot\mathbb{E}\underbrace{\sup_{f\in\mathcal{F}}(\frac{1}{n}\sum_{i=1}^n \epsilon_i\phi_\delta(y_i f(x_i)))}_{R} + \sqrt{\frac{2u}{n}} \\
& \underbrace{\le}_{\text{contraction}} \quad \mathbb{E}_n\phi_\delta(yf(x)) + \frac{2}{\delta}\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \epsilon_i y_i f(x_i) + \sqrt{\frac{2u}{n}} \\
& = \quad \mathbb{E}_n\phi_\delta(yf(x)) + \frac{2}{\delta}\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i) + \sqrt{\frac{2u}{n}} \\
& \le \quad \mathbb{P}_n(yf(x) < 0) + \frac{2}{\delta}\mathbb{E}\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n \epsilon_i h(x_i) + \sqrt{\frac{2u}{n}}.
\end{aligned}$$

To bound the second term $(\mathbb{E}\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n \epsilon_i h(x_i))$ above, we will use the following fact.

**Fact 27.1.** *If $\mathbb{P}(\xi \ge a + b\cdot t) \le \exp(-t^2)$, then $\mathbb{E}\xi \le K\cdot(a+b)$ for some constant $K$.*

If $\mathcal{H}$ is a VC-subgraph class and $V$ is its VC dimension, $D(\mathcal{H}, \epsilon, d_x) \le K\left(\frac{1}{\epsilon}\right)^{2\cdot V}$ by D. Haussler. By Kolmogorov's chaining method (Lecture 14),

$$\begin{aligned}
& = \quad \mathbb{P}\left(\sup_h \frac{1}{n}\sum_{i=1}^n \epsilon_i h(x_i) \le K\left(\frac{1}{n}\int_0^1 \log^{1/2} D(\mathcal{H}, \epsilon, d_x)d\epsilon + \sqrt{\frac{u}{n}}\right)\right) \\
& = \quad \mathbb{P}\left(\sup_h \frac{1}{n}\sum_{i=1}^n \epsilon_i h(x_i) \le K\left(\frac{1}{n}\int_0^1 \sqrt{V\log\frac{1}{\epsilon}}d\epsilon + \sqrt{\frac{u}{n}}\right)\right) \\
& \ge \quad 1 - e^{-u}.
\end{aligned}$$

Thus $\mathbb{E}\sup \frac{1}{n}\sum \epsilon_i h(x_i) \leq K\left(\sqrt{\frac{V}{n}} + \sqrt{\frac{1}{n}}\right) \leq K\sqrt{\frac{V}{n}}$, and

$$\mathbb{P}\left(\mathbb{P}(yf(x) < 0) \leq \mathbb{P}_n(yf(x) < 0) + K\frac{1}{\delta}\sqrt{\frac{V}{n}} + \sqrt{\frac{2u}{n}}\right) \quad \geq \quad 1 - e^{-u}.$$

Recall our set up for Martingale inequalities. Let $Z = Z(x_1, \cdots, x_n)$ where $x_1, \cdots, x_n$ are independent random variables. We need to bound $Z - \mathbb{E}Z$. Since $Z$ is not a sum of independent random variables, certain classical concentration inequalities is not applicable. But we can try to bound $Z - \mathbb{E}Z$ with certain form of Martingale inequalities.

$$\begin{aligned} Z - \mathbb{E}Z \quad = \quad & \underbrace{Z - \mathbb{E}_{x_1}(Z|x_2, \cdots, x_n)}_{d_1(x_1,\cdots,x_n)} + \underbrace{\mathbb{E}_{x_1}(Z|x_2, \cdots, x_n) - \mathbb{E}_{x_1,x_2}(Z|x_3, \cdots, x_n)}_{d_2(x_2,\cdots,x_n)} + \\ & \cdots + \underbrace{\mathbb{E}_{x_1,\cdots,x_{n-1}}(Z|x_n) - \mathbb{E}_{x_1,\cdots,x_n}(Z)}_{d_n(x_n)} \end{aligned}$$

with the assumptions that $\mathbb{E}_{x_i} d_i = 0$, and $\|d_i\|_\infty \leq c_i$.

We will give a generalized martingale inequality below. $\sum_{i=1}^n d_i = Z - \mathbb{E}Z$ where $d_i = d_i(x_i, \cdots, x_n)$, $\max_i \|d_i\|_\infty \leq C$, $\sigma_i^2 = \sigma_i^2(x_{i+1}, \cdots, x_n) = \text{var}(d_i)$, and $\mathbb{E}d_i = 0$. Take $\epsilon > 0$,

$$\begin{aligned} & \mathbb{P}(\sum_{i=1}^n d_i - \epsilon\sum_{i=1}^n \sigma_i^2 \geq t) \\ \leq \quad & e^{-\lambda t}\mathbb{E}\exp(\sum_{i=1}^n \lambda(d_i - \epsilon\sigma_i^2)) \\ = \quad & e^{-\lambda t}\mathbb{E}\exp(\sum_{i=1}^{n-1} \lambda(d_i - \epsilon\sigma_i^2) \cdot \mathbb{E}\exp(\lambda d_n) \cdot \exp(\lambda\epsilon\sigma_n^2) \end{aligned}$$

The term $\exp(\lambda d_n)$ can be bounded in the following way.

$$\begin{aligned} & \mathbb{E}\exp(\lambda d_n) \\ \underbrace{=}_{\text{Taylor expansion}} \quad & \mathbb{E}\left(1 + \lambda d_n + \frac{\lambda^2}{2!}d_n^2 + \frac{\lambda^3}{3!}d_n^3 + \cdots\right) \\ \leq \quad & 1 + \frac{\lambda^2}{2}\sigma_n^2 \cdot \left(1 + \frac{\lambda C}{3} + \frac{\lambda^2 C^2}{3 \cdot 4} + \cdots\right) \\ \leq \quad & \exp\left(\frac{\lambda^2 \cdot \sigma_n^2}{2} \cdot \frac{1}{(1 - \lambda C)}\right). \end{aligned}$$

Choose $\lambda$ such that $\frac{\lambda^2}{2\cdot(1-\lambda C)} \leq \lambda\epsilon$, we get $\lambda \leq \frac{2\epsilon}{1+2\epsilon C}$, and $\mathbb{E}_{d_n}\exp(\lambda d_n) \cdot \exp(\lambda\epsilon\sigma_n^2) \leq 1$. Iterate over $i = n, \cdots, 1$, we get

$$\mathbb{P}\left(\sum_{i=1}^n d_i - \epsilon\sum_{i=1}^n \sigma_i^2 \geq t\right) \quad \leq \quad \exp(-\lambda \cdot t)$$

. Take $t = u/\lambda$, we get

$$\mathbb{P}\left(\sum_{i=1}^{n} d_i \geq \epsilon \sum_{i=1}^{n} \sigma_i^2 + \frac{u}{2\epsilon}(1+2\epsilon C)\right) \quad \leq \quad \exp(-u)$$

To minimize the sum $\epsilon \sum_{i=1}^{n} \sigma_i^2 + \frac{u}{2\epsilon}(1+2\epsilon C)$, we set its derivative to 0, and get $\epsilon = \sqrt{\frac{u}{2\sum \sigma_i^2}}$. Thus

$$\mathbb{P}\left(\sum d_i \geq 3\sqrt{u\sum_i \sigma_i^2/2} + Cu\right) \quad \leq \quad e^{-u}$$

. This inequality takes the form of the Bernstein's inequality.

Let $\mathcal{H}$ be a class of "simple" functions (VC-subgraph, perceptrons). Define recursively

$$\mathcal{H}_{i+1} = \left\{\sigma\left(\sum \alpha_j h_j\right) : \ h_j \in \mathcal{H}_i, \ \alpha_j \in \mathbb{R}\right\}$$

where $\sigma$ is sigmoid function such that $\sigma(0) = 0$ and $|\sigma(s) - \sigma(t)| \leq L|s-t|$, $-1 \leq \sigma \leq 1$.

Example:

$$\sigma(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}.$$

Assume we have data $(x_1, y_1), \ldots, (x_n, y_n)$, $-1 \leq y_i \leq 1$. We can minimize

$$\frac{1}{n}\sum_{i=1}^n (y_i - h(x_i))^2$$

over $\mathcal{H}_k$, where $k$ is the number of layers.

Define $\mathcal{L}(y, h(x)) = (y - h(x))^2$, $0 \leq \mathcal{L}(y, h(x)) \leq 4$. We want to bound $\mathbb{E}\mathcal{L}(y, h(x))$.

From the previous lectures,

$$\sup \left| \mathbb{E}\mathcal{L}(y, h(x)) - \frac{1}{n}\sum_{i=1}^n \mathcal{L}(y_i, h(x_i))\right| \leq 2\mathbb{E}\sup\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathcal{L}(y_i, h(x_i))\right| + 4\sqrt{\frac{2t}{n}}$$

with probability at least $1 - e^{-t}$.

Define

$$\mathcal{H}_{i+1}(A_1, \ldots, A_{i+1}) = \left\{\sigma\left(\sum \alpha_j h_j\right) : \ \sum |\alpha_j| \leq A_{i+1}, \ h_j \in \mathcal{H}_i\right\}.$$

For now, assume bounds $A_i$ on sum of weights (although this is not true in practice, so we will take union bound later).

**Theorem 28.1.**

$$\mathbb{E}\sup_{h\in\mathcal{H}_k(A_1,\ldots,A_k)}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathcal{L}(y_i, h(x_i))\right| \leq 8\prod_{j=1}^k (2L\cdot A_j)\cdot \mathbb{E}\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i h(x_i)\right| + \frac{8}{\sqrt{n}}.$$

*Proof.* Since $-2 \leq y - h(x) \leq 2$, $\frac{(y-h(x))^2}{4} : [-2, 2] \mapsto \mathbb{R}$ is a contraction because largest derivative of $s^2$ on $[-2, 2]$ is 4. Hence,

$$\begin{aligned}
\mathbb{E}\sup_{h\in\mathcal{H}_k(A_1,\ldots,A_k)}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i (y_i - h(x_i))^2\right| &= \mathbb{E}\mathbb{E}_\varepsilon \sup_{h\in\mathcal{H}_k(A_1,\ldots,A_k)}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i (y_i - h(x_i))^2\right| \\
&= 4\mathbb{E}\mathbb{E}_\varepsilon \sup_{h\in\mathcal{H}_k(A_1,\ldots,A_k)}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \frac{(y_i - h(x_i))^2}{4}\right| \\
&\leq 8\mathbb{E}\mathbb{E}_\varepsilon \sup_{h\in\mathcal{H}_k(A_1,\ldots,A_k)}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i (y_i - h(x_i))\right| \\
&\leq 8\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i y_i\right| + 8\mathbb{E}\sup_{h\in\mathcal{H}_k(A_1,\ldots,A_k)}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i h(x_i)\right|
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i y_i\right| &\le \left(\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i y_i\right)^2\right)^{1/2} \\
&= \left(\mathbb{E}\sum_{i=1}^{n}\frac{1}{n^2}\varepsilon_i^2 y_i^2\right)^{1/2} \\
&= \left(\frac{1}{n}\mathbb{E}y_1^2\right)^{1/2} \le \sqrt{\frac{1}{n}}
\end{aligned}$$

Using the fact that $\sigma/L$ is a contraction,

$$\begin{aligned}
\mathbb{E}_\varepsilon \sup_{h\in\mathcal{H}_k(A_1,\ldots,A_k)}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\sigma\left(\sum\alpha_j h_j(x_i)\right)\right| &= L\mathbb{E}_\varepsilon\sup_h\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\frac{\sigma}{L}\left(\sum\alpha_j h_j(x_i)\right)\right| \\
&\le 2L\mathbb{E}_\varepsilon\sup_h\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left(\sum\alpha_j h_j(x_i)\right)\right| \\
&= 2L\mathbb{E}_\varepsilon\sup_h\left|\frac{1}{n}\sum_j\alpha_j\left(\sum_{i=1}^{n}\varepsilon_i h_j(x_i)\right)\right| \\
&= 2L\mathbb{E}_\varepsilon\sup_h\left|\frac{\sum|\alpha_j|}{n}\sum_j\alpha_j'\left(\sum_{i=1}^{n}\varepsilon_i h_j(x_i)\right)\right|
\end{aligned}$$

where $\alpha_j' = \frac{\alpha_j}{\sum_j|\alpha_j|}$. Since $\sum_j|\alpha_j| \le A_k$ for the layer $k$,

$$\begin{aligned}
&2L\mathbb{E}_\varepsilon\sup_{h\in\mathcal{H}_k(A_1,\ldots,A_k)}\left|\frac{\sum|\alpha_j|}{n}\sum_j\alpha_j'\left(\sum_{i=1}^{n}\varepsilon_i h_j(x_i)\right)\right| \\
&\le 2LA_k\mathbb{E}_\varepsilon\sup_{h\in\mathcal{H}_k(A_1,\ldots,A_k)}\left|\frac{1}{n}\sum_j\alpha_j'\left(\sum_{i=1}^{n}\varepsilon_i h_j(x_i)\right)\right| \\
&= 2LA_k\mathbb{E}_\varepsilon\sup_{h\in\mathcal{H}_{k-1}(A_1,\ldots,A_{k-1})}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i h_j(x_i)\right|
\end{aligned}$$

The last equality holds because $\sup|\sum\lambda_j s_j| = \max_j|s_j|$, i.e. max is attained at one of the vertices.

By induction,

$$\mathbb{E}\sup_{h\in\mathcal{H}_k(A_1,\ldots,A_k)}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(y_i-h(x_i))^2\right| \le 8\prod_{j=1}^{k}(2LA_j)\cdot\mathbb{E}\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i h(x_i)\right| + \frac{8}{\sqrt{n}},$$

where $\mathcal{H}$ is the class of simple classifiers. □

In Lecture 28 we proved

$$\mathbb{E} \sup_{h \in \mathcal{H}_k(A_1,\ldots,A_k)} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i (y_i - h(x_i))^2 \right| \leq 8 \prod_{j=1}^k (2LA_j) \cdot \mathbb{E} \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i h(x_i) \right| + \frac{8}{\sqrt{n}}$$

Hence,

$$\begin{aligned} Z\left(\mathcal{H}_k(A_1,\ldots,A_k)\right) := & \sup_{h \in \mathcal{H}_k(A_1,\ldots,A_k)} \left| \mathbb{E}\mathcal{L}(y, h(x)) - \frac{1}{n} \sum_{i=1}^n \mathcal{L}(y_i, h(x_i)) \right| \\ & \leq 8 \prod_{j=1}^k (2LA_j) \cdot \mathbb{E} \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i h(x_i) \right| + \frac{8}{\sqrt{n}} + 8\sqrt{\frac{t}{n}} \end{aligned}$$

with probability at least $1 - e^{-t}$.

Assume $\mathcal{H}$ is a VC-subgraph class, $-1 \leq h \leq 1$.

We had the following result:

$$\mathbb{P}_\varepsilon \left( \forall h \in \mathcal{H},\ \frac{1}{n} \sum_{i=1}^n \varepsilon_i h(x_i) \leq \frac{K}{\sqrt{n}} \int_0^{\sqrt{\frac{1}{n}\sum_{i=1}^n h^2(x_i)}} \log^{1/2} \mathcal{D}(\mathcal{H}, \varepsilon, d_x) d\varepsilon + K \sqrt{\frac{t}{n} \left( \frac{1}{n} \sum_{i=1}^n h^2(x_i) \right)} \right) \geq 1 - e^{-t},$$

where

$$d_x(f,g) = \left( \frac{1}{n} \sum_{i=1}^n (f(x_i) - g(x_i))^2 \right)^{1/2}.$$

Furthermore,

$$\mathbb{P}_\varepsilon \left( \forall h \in \mathcal{H},\ \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i h(x_i) \right| \leq \frac{K}{\sqrt{n}} \int_0^{\sqrt{\frac{1}{n}\sum_{i=1}^n h^2(x_i)}} \log^{1/2} \mathcal{D}(\mathcal{H}, \varepsilon, d_x) d\varepsilon + K \sqrt{\frac{t}{n} \left( \frac{1}{n} \sum_{i=1}^n h^2(x_i) \right)} \right). \geq 1 - 2e^{-t},$$

Since $-1 \leq h \leq 1$ for all $h \in \mathcal{H}$,

$$\mathbb{P}_\varepsilon \left( \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i h(x_i) \right| \leq \frac{K}{\sqrt{n}} \int_0^1 \log^{1/2} \mathcal{D}(\mathcal{H}, \varepsilon, d_x) d\varepsilon + K \sqrt{\frac{t}{n}} \right) \geq 1 - 2e^{-t},$$

Since $\mathcal{H}$ is a VC-subgraph class with $VC(\mathcal{H}) = V$,

$$\log \mathcal{D}(\mathcal{H}, \varepsilon, d_x) \leq KV \log \frac{2}{\varepsilon}.$$

Hence,

$$\int_0^1 \log^{1/2} \mathcal{D}(\mathcal{H}, \varepsilon, d_x) d\varepsilon \leq \int_0^1 \sqrt{KV \log \frac{2}{\varepsilon}} d\varepsilon$$
$$\leq K\sqrt{V} \int_0^1 \sqrt{\log \frac{2}{\varepsilon}} d\varepsilon \leq K\sqrt{V}$$

Let $\xi \geq 0$ be a random variable. Then

$$\mathbb{E}\xi = \int_0^\infty \mathbb{P}(\xi \geq t)\, dt = \int_0^a \mathbb{P}(\xi \geq t)\, dt + \int_a^\infty \mathbb{P}(\xi \geq t)\, dt$$
$$\leq a + \int_a^\infty \mathbb{P}(\xi \geq t)\, dt = a + \int_0^\infty \mathbb{P}(\xi \geq a + u)\, du$$

Let $K\sqrt{\frac{V}{n}} = a$ and $K\sqrt{\frac{t}{n}} = u$. Then $e^{-t} = e^{-\frac{nu^2}{K^2}}$. Hence, we have

$$\mathbb{E}_\varepsilon \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i h(x_i) \right| \leq K\sqrt{\frac{V}{n}} + \int_0^\infty 2e^{-\frac{nu^2}{K^2}} du$$
$$= K\sqrt{\frac{V}{n}} + \int_0^\infty \frac{K}{\sqrt{n}} e^{-x^2} dx$$
$$\leq K\sqrt{\frac{V}{n}} + \frac{K}{\sqrt{n}} \leq K\sqrt{\frac{V}{n}}$$

for $V \geq 2$. We made a change of variable so that $x^2 = \frac{nu^2}{K^2}$. Constants $K$ change their values from line to line.

We obtain,

$$Z\left(\mathcal{H}_k(A_1, \ldots, A_k)\right) \leq K \prod_{j=1}^k (2LA_j) \cdot \sqrt{\frac{V}{n}} + \frac{8}{\sqrt{n}} + 8\sqrt{\frac{t}{n}}$$

with probability at least $1 - e^{-t}$.

Assume that for any $j$, $A_j \in (2^{-\ell_j - 1}, 2^{-\ell_j}]$. This defines $\ell_j$. Let

$$\mathcal{H}_k(\ell_1, \ldots, \ell_k) = \bigcup \left\{ \mathcal{H}_k(A_1, \ldots, A_k) : A_j \in (2^{-\ell_j - 1}, 2^{-\ell_j}] \right\}.$$

Then the empirical process

$$Z\left(\mathcal{H}_k(\ell_1, \ldots, \ell_k)\right) \leq K \prod_{j=1}^k (2L \cdot 2^{-\ell_j}) \cdot \sqrt{\frac{V}{n}} + \frac{8}{\sqrt{n}} + 8\sqrt{\frac{t}{n}}$$

with probability at least $1 - e^{-t}$.

For a given sequence $(\ell_1, \ldots, \ell_k)$, redefine $t$ as $t + 2\sum_{j=1}^k \log |w_j|$ where $w_j = \ell_j$ if $\ell_j \neq 0$ and $w_j = 1$ if $\ell_j = 0$.

With this $t$,

$$Z\left(\mathcal{H}_k(\ell_1, \ldots, \ell_k)\right) \leq K \prod_{j=1}^k (2L \cdot 2^{-\ell_j}) \cdot \sqrt{\frac{V}{n}} + \frac{8}{\sqrt{n}} + 8\sqrt{\frac{t + 2\sum_{j=1}^k \log |w_j|}{n}}$$

with probability at least

$$1 - e^{-t-2\sum_{j=1}^k \log|w_j|} = 1 - \prod_{j=1}^k \frac{1}{|w_j|^2} e^{-t}.$$

By union bound, the above holds for all $\ell_1, \dots, \ell_k \in \mathcal{Z}$ with probability at least

$$\begin{aligned} 1 - \sum_{\ell_1,\dots,\ell_k \in \mathcal{Z}} \prod_{j=1}^k \frac{1}{|w_j|^2} e^{-t} &= 1 - \left( \sum_{\ell_1 \in \mathcal{Z}} \frac{1}{|w_1|^2} \right)^k e^{-t} \\ &= 1 - \left(1 + 2\frac{\pi^2}{6}\right)^k e^{-t} \geq 1 - 5^k e^{-t} = 1 - e^{-u} \end{aligned}$$

for $t = u + k \log 5$.

Hence, with probability at least $1 - e^{-u}$,

$$\forall(\ell_1, \dots, \ell_k),\ Z\left(\mathcal{H}_k(\ell_1, \dots, \ell_k)\right) \leq K \prod_{j=1}^k (2L \cdot 2^{-\ell_j}) \cdot \sqrt{\frac{V}{n}} + \frac{8}{\sqrt{n}} + 8\sqrt{\frac{2\sum_{j=1}^k \log|w_j| + k\log 5 + u}{n}}.$$

If $A_j \in (2^{-\ell_j - 1}, 2^{-\ell_j}]$, then $-\ell_j - 1 \leq \log A_j \leq \ell_j$ and $|\ell_j| \leq |\log A_j| + 1$. Hence, $|w_j| \leq |\log A_j| + 1$. Therefore, with probability at least $1 - e^{-u}$,

$$\begin{aligned} \forall(\mathcal{A}_1, \dots, \mathcal{A}_k),\ Z\left(\mathcal{H}_k(A_1, \dots, A_k)\right) \leq K \prod_{j=1}^k (4L \cdot A_j) \cdot \sqrt{\frac{V}{n}} + \frac{8}{\sqrt{n}} \\ +8\sqrt{\frac{2\sum_{j=1}^k \log\left(|\log A_j| + 1\right) + k\log 5 + u}{n}}. \end{aligned}$$

Notice that $\log\left(|\log A_j| + 1\right)$ is large when $A_j$ is very large or very small. This is penalty and we want the product term to be dominating. But $\log\log A_j \leq 5$ for most practical applications.

Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact subset. Assume $x_1, \ldots, x_n$ are i.i.d. and $y_1, \ldots, y_n = \pm 1$ for classification and $[-1, 1]$ for regression. Assume we have a kernel $K(x, y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(y)$, $\lambda_i > 0$.

Consider a map

$$x \in \mathcal{X} \mapsto \phi(x) = (\sqrt{\lambda_1}\phi_1(x), \ldots, \sqrt{\lambda_k}\phi_k(x), \ldots) = (\sqrt{\lambda_k}\phi_k(x))_{k \geq 1} \in \mathcal{H}$$

where $\mathcal{H}$ is a Hilbert space.

Consider the scalar product in $\mathcal{H}$: $(u, v)_{\mathcal{H}} = \sum_{i=1}^{\infty} u_i v_i$ and $\|u\|_{\mathcal{H}} = \sqrt{(u, v)_{\mathcal{H}}}$.

For $x, y \in \mathcal{X}$,

$$(\phi(x), \phi(y))_{\mathcal{H}} = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(y) = K(x, y).$$

Function $\phi$ is called *feature map*.

Family of classifiers:

$$\mathcal{F}_{\mathcal{H}} = \{(w, z)_{\mathcal{H}} : \|w\|_{\mathcal{H}} \leq 1\}.$$

$$\mathcal{F} = \{(w, \phi(x))_{\mathcal{H}} : \|w\|_{\mathcal{H}} \leq 1\} \ni f : \mathcal{X} \mapsto \mathbb{R}.$$

**Algorithms:**

(1) **SVMs**

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) = (\underbrace{\sum_{i=1}^{n} \alpha_i \phi(x_i)}_{w}, \phi(x))_{\mathcal{H}}$$

Here, instead of taking any $w$, we only take $w$ as a linear combination of images of data points. We have a choice of Loss function $\mathcal{L}$:

- $\mathcal{L}(y, f(x)) = I(yf(x) \leq 0)$ – classification
- $\mathcal{L}(y, f(x)) = (y - f(x))^2$ – regression

(2) **Square-loss regularization**

Assume an algorithm outputs a classifier from $\mathcal{F}$ (or $\mathcal{F}_{\mathcal{H}}$), $f(x) = (w, \phi(x))_{\mathcal{H}}$. Then, as in Lecture 20,

$$\begin{aligned}\mathbb{P}\left(yf(x) \leq 0\right) \leq \mathbb{E}\varphi_\delta\left(yf(x)\right) &= \frac{1}{n}\sum_{i=1}^{n} \varphi_\delta\left(y_i f(x_i)\right) + \left(\mathbb{E}\varphi_\delta\left(yf(x)\right) - \frac{1}{n}\sum_{i=1}^{n}\varphi_\delta\left(y_i f(x_i)\right)\right) \\ &\leq \frac{1}{n}\sum_{i=1}^{n} I(y_i f(x_i) \leq \delta) + \sup_{f \in \mathcal{F}}\left(\mathbb{E}\varphi_\delta\left(yf(x)\right) - \frac{1}{n}\sum_{i=1}^{n}\varphi_\delta\left(y_i f(x_i)\right)\right)\end{aligned}$$

By McDiarmid's inequality, with probability at least $1 - e^{-t}$

$$\sup_{f \in \mathcal{F}}\left(\mathbb{E}\varphi_\delta\left(yf(x)\right) - \frac{1}{n}\sum_{i=1}^{n}\varphi_\delta\left(y_i f(x_i)\right)\right) \leq \mathbb{E}\sup_{f \in \mathcal{F}}\left(\mathbb{E}\varphi_\delta\left(yf(x)\right) - \frac{1}{n}\sum_{i=1}^{n}\varphi_\delta\left(y_i f(x_i)\right)\right) + \sqrt{\frac{2t}{n}}$$

Using the symmetrization technique,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left(\mathbb{E}(\varphi_\delta\left(yf(x)\right)-1)-\frac{1}{n}\sum_{i=1}^{n}(\varphi_\delta\left(y_if(x_i)\right)-1)\right)\leq 2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\big(\varphi_\delta\left(y_if(x_i)\right)-1\big)\right|.$$

Since $\delta\cdot(\varphi_\delta-1)$ is a contraction,

$$\begin{aligned}
2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\big(\varphi_\delta\left(y_if(x_i)\right)-1\big)\right| &\leq \frac{2}{\delta}2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_iy_if(x_i)\right|\\
&= \frac{4}{\delta}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_if(x_i)\right| = \frac{4}{\delta}\mathbb{E}\sup_{\|w\|\leq 1}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(w,\phi(x_i))_{\mathcal{H}}\right|\\
&= \frac{4}{\delta n}\mathbb{E}\sup_{\|w\|\leq 1}\left|(w,\sum_{i=1}^{n}\varepsilon_i\phi(x_i))_{\mathcal{H}}\right| = \frac{4}{\delta n}\mathbb{E}\sup_{\|w\|\leq 1}\left\|\sum_{i=1}^{n}\varepsilon_i\phi(x_i)\right\|_{\mathcal{H}}\\
&= \frac{4}{\delta n}\mathbb{E}\sqrt{\left(\sum_{i=1}^{n}\varepsilon_i\phi(x_i),\sum_{i=1}^{n}\varepsilon_i\phi(x_i)\right)_{\mathcal{H}}} = \frac{4}{\delta n}\mathbb{E}\sqrt{\sum_{i,j}\varepsilon_i\varepsilon_j(\phi(x_i),\phi(x_i))_{\mathcal{H}}}\\
&= \frac{4}{\delta n}\mathbb{E}\sqrt{\sum_{i,j}\varepsilon_i\varepsilon_jK(x_i,x_j)} \leq \frac{4}{\delta n}\sqrt{\mathbb{E}\sum_{i,j}\varepsilon_i\varepsilon_jK(x_i,x_j)}\\
&= \frac{4}{\delta n}\sqrt{\sum_{i=1}^{n}\mathbb{E}K(x_i,x_i)} = \frac{4}{\delta}\sqrt{\frac{\mathbb{E}K(x_1,x_1)}{n}}
\end{aligned}$$

Putting everything together, with probability at least $1-e^{-t}$,

$$\mathbb{P}\left(yf(x)\leq 0\right)\leq\frac{1}{n}\sum_{i=1}^{n}I(y_if(x_i)\leq\delta)+\frac{4}{\delta}\sqrt{\frac{\mathbb{E}K(x_1,x_1)}{n}}+\sqrt{\frac{2t}{n}}.$$

Before the contraction step, we could have used Martingale method again to have $\mathbb{E}_\varepsilon$ only. Then $\mathbb{E}K(x_1,x_1)$ in the above bound will become $\frac{1}{n}\sum_{i=1}^{n}K(x_i,x_i)$.

As in the previous lecture, let $\mathcal{F} = \{(w, \phi(x))_{\mathcal{H}}, \|w\| \leq 1\}$, where $\phi(x) = (\sqrt{\lambda_i}\phi_i(x))_{i\geq 1}$, $\mathcal{X} \subset \mathbb{R}^d$.

Define $d(f,g) = \|f-g\|_\infty = \sup_{x\in\mathcal{X}} |f(x)-g(x)|$.

The following theorem appears in Cucker & Smale:

**Theorem 31.1.** *$\forall h \geq d$,*

$$\log \mathcal{N}(\mathcal{F}, \varepsilon, d) \leq \left(\frac{C_h}{\varepsilon}\right)^{\frac{2d}{h}}$$

*where $C_h$ is a constant.*

Note that for any $x_1, \ldots, x_n$,

$$d_x(f,g) = \left(\frac{1}{n}\sum_{i=1}^n (f(x_i)-g(x_i))^2\right)^{1/2} \leq d(f,g) = \sup_x |f(x)-g(x)| \leq \varepsilon.$$

Hence,

$$\mathcal{N}(\mathcal{F}, \varepsilon, d_x) \leq \mathcal{N}(\mathcal{F}, \varepsilon, d).$$

Assume the loss function $\mathcal{L}(y, f(x)) = (y-f(x))^2$. The *loss class* is defined as

$$\mathcal{L}(y, F) = \{(y-f(x))^2, f \in \mathcal{F}\}.$$

Suppose $|y - f(x)| \leq M$. Then

$$|(y-f(x))^2 - (y-g(x))^2| \leq 2M|f(x)-g(x)| \leq \varepsilon.$$

So,

$$\mathcal{N}(\mathcal{L}(y,\mathcal{F}), \varepsilon, d_x) \leq \mathcal{N}\left(\mathcal{F}, \frac{\varepsilon}{2M}, d_x\right)$$

and

$$\log \mathcal{N}(\mathcal{L}(y,\mathcal{F}), \varepsilon, d_x) \leq \left(\frac{2MC_h}{\varepsilon}\right)^{\frac{2d}{h}} = \left(\frac{2MC_h}{\varepsilon}\right)^{\alpha}$$

$\alpha = \frac{2d}{h} < 2$ (see Homework 2, problem 4).

Now, we would like to use specific form of solution for SVM: $f(x) = \sum_{i=1}^n \alpha_i K(x_i, x)$, i.e. $f$ belongs to a random subclass. We now prove a VC inequality for random collection of sets.

Let's consider $\mathcal{C}(x_1,\ldots,x_n) = \{C : C \subseteq \mathcal{X}\}$ - random collection of sets. Assume that $\mathcal{C}(x_1,\ldots,x_n)$ satisfies:

(1) $C(x_1,\ldots,x_n) \subseteq C(x_1,\ldots,x_n,x_{n+1})$

(2) $C(\pi(x_1,\ldots,x_n)) = C(x_1,\ldots,x_n)$ for any permutation $\pi$.

Let

$$\triangle_{\mathcal{C}}(x_1,\ldots,x_n) = \text{card } \{C \cap \{x_1,\ldots,x_n\}; C \in \mathcal{C}\}$$

and

$$G(n) = \mathbb{E}\triangle_{\mathcal{C}(x_1,\ldots,x_n)}(x_1,\ldots,x_n).$$

**Theorem 31.2.**

$$\mathbb{P}\left(\sup_{C\in\mathcal{C}(x_1,\ldots,x_n)} \frac{\mathbb{P}\left(C\right)-\frac{1}{n}\sum_{i=1}^n I(x_i\in C)}{\sqrt{\mathbb{P}\left(C\right)}} \geq t\right) \leq 4G(2n)e^{-\frac{nt^2}{4}}$$

Consider event

$$A_x = \left\{x=(x_1,\ldots,x_n) : \sup_{C\in\mathcal{C}(x_1,\ldots,x_n)} \frac{\mathbb{P}\left(C\right)-\frac{1}{n}\sum_{i=1}^n I(x_i\in C)}{\sqrt{\mathbb{P}\left(C\right)}} \geq t\right\}$$

So, there exists $C_x \in \mathcal{C}(x_1,\ldots,x_n)$ such that

$$\frac{\mathbb{P}\left(C_x\right)-\frac{1}{n}\sum_{i=1}^n I(x_i\in C_x)}{\sqrt{\mathbb{P}\left(C_x\right)}} \geq t.$$

For $x'_1,\ldots,x'_n$, an independent copy of $x$,

$$\mathbb{P}_{x'}\left(\mathbb{P}\left(C_x\right) \leq \frac{1}{n}\sum_{i=1}^n I(x'_i\in C_x)\right) \geq \frac{1}{4}$$

if $\mathbb{P}\left(C_x\right) \geq \frac{1}{n}$ (which we can assume without loss of generality).

Together,

$$\mathbb{P}\left(C_x\right) \leq \frac{1}{n}\sum_{i=1}^n I(x'_i\in C_x)$$

and

$$\frac{\mathbb{P}\left(C_x\right)-\frac{1}{n}\sum_{i=1}^n I(x_i\in C_x)}{\sqrt{\mathbb{P}\left(C_x\right)}} \geq t$$

imply

$$\frac{\frac{1}{n}\sum_{i=1}^n I(x'_i\in C_x)-\frac{1}{n}\sum_{i=1}^n I(x_i\in C_x)}{\sqrt{\frac{1}{2n}\sum_{i=1}^n (I(x'_i\in C_x)+I(x_i\in C_x))}} \geq t.$$

Indeed,

$$\begin{aligned}
0 < t &\leq \frac{\mathbb{P}\left(C_x\right)-\frac{1}{n}\sum_{i=1}^n I(x_i\in C_x)}{\sqrt{\mathbb{P}\left(C_x\right)}} \\
&\leq \frac{\mathbb{P}\left(C_x\right)-\frac{1}{n}\sum_{i=1}^n I(x_i\in C_x)}{\sqrt{\frac{1}{2}\left(\mathbb{P}\left(C_x\right)+\frac{1}{n}\sum_{i=1}^n I(x_i\in C_x)\right)}} \\
&\leq \frac{\frac{1}{n}\sum_{i=1}^n I(x'_i\in C_x)-\frac{1}{n}\sum_{i=1}^n I(x_i\in C_x)}{\sqrt{\frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^n I(x'_i\in C_x)+\frac{1}{n}\sum_{i=1}^n I(x_i\in C_x)\right)}}
\end{aligned}$$

Hence, multiplying by an indicator,

$$\begin{aligned}
\frac{1}{4}\cdot I(x\in A_x) &\le \mathbb{P}_{x'}\left(\mathbb{P}\left(C_x\right)\le \frac{1}{n}\sum_{i=1}^{n} I(x_i'\in C_x)\right)\cdot I(x\in A_x)\\
&\le \mathbb{P}_{x'}\left(\frac{\frac{1}{n}\sum_{i=1}^{n} I(x_i'\in C_x)-\frac{1}{n}\sum_{i=1}^{n} I(x_i\in C_x)}{\sqrt{\frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} I(x_i'\in C_x)+\frac{1}{n}\sum_{i=1}^{n} I(x_i\in C_x)\right)}}\ge t\right)\\
&\le \mathbb{P}_{x'}\left(\sup_{C\in\mathcal{C}(x_1,\ldots,x_n)}\frac{\frac{1}{n}\sum_{i=1}^{n} I(x_i'\in C_x)-\frac{1}{n}\sum_{i=1}^{n} I(x_i\in C_x)}{\sqrt{\frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} I(x_i'\in C_x)+\frac{1}{n}\sum_{i=1}^{n} I(x_i\in C_x)\right)}}\ge t\right)
\end{aligned}$$

Taking expectation with respect to $x$ on both sides,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{C\in\mathcal{C}(x_1,\ldots,x_n)}\frac{\mathbb{P}\left(C\right)-\frac{1}{n}\sum_{i=1}^{n} I(x_i\in C)}{\sqrt{\mathbb{P}\left(C\right)}}\ge t\right)\\
&\le 4\mathbb{P}\left(\sup_{C\in\mathcal{C}(x_1,\ldots,x_n)}\frac{\frac{1}{n}\sum_{i=1}^{n} I(x_i'\in C_x)-\frac{1}{n}\sum_{i=1}^{n} I(x_i\in C_x)}{\sqrt{\frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} I(x_i'\in C_x)+\frac{1}{n}\sum_{i=1}^{n} I(x_i\in C_x)\right)}}\ge t\right)\\
&\le 4\mathbb{P}\left(\sup_{C\in\mathcal{C}(x_1,\ldots,x_n,x_1',\ldots,x_n')}\frac{\frac{1}{n}\sum_{i=1}^{n} I(x_i'\in C_x)-\frac{1}{n}\sum_{i=1}^{n} I(x_i\in C_x)}{\sqrt{\frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} I(x_i'\in C_x)+\frac{1}{n}\sum_{i=1}^{n} I(x_i\in C_x)\right)}}\ge t\right)\\
&= 4\mathbb{P}\left(\sup_{C\in\mathcal{C}(x_1,\ldots,x_n,x_1',\ldots,x_n')}\frac{\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i(I(x_i'\in C_x)-I(x_i\in C_x))}{\sqrt{\frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} I(x_i'\in C_x)+\frac{1}{n}\sum_{i=1}^{n} I(x_i\in C_x)\right)}}\ge t\right)\\
&= 4\mathbb{E}\mathbb{P}_{\varepsilon}\left(\sup_{C\in\mathcal{C}(x_1,\ldots,x_n,x_1',\ldots,x_n')}\frac{\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i(I(x_i'\in C_x)-I(x_i\in C_x))}{\sqrt{\frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} I(x_i'\in C_x)+\frac{1}{n}\sum_{i=1}^{n} I(x_i\in C_x)\right)}}\ge t\right)
\end{aligned}$$

By Hoeffding,

$$\begin{aligned}
&4\mathbb{E}\mathbb{P}_{\varepsilon}\left(\sup_{C\in\mathcal{C}(x_1,\ldots,x_n,x_1',\ldots,x_n')}\frac{\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i(I(x_i'\in C_x)-I(x_i\in C_x))}{\sqrt{\frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} I(x_i'\in C_x)+\frac{1}{n}\sum_{i=1}^{n} I(x_i\in C_x)\right)}}\ge t\right)\\
&\le 4\mathbb{E}\triangle_{\mathcal{C}(x_1,\ldots,x_n,x_1',\ldots,x_n')}(x_1,\ldots,x_n,x_1',\ldots,x_n')\cdot\exp\left(-\frac{t^2}{2\sum\left(\frac{\frac{1}{n}(I(x_i'\in C_x)-I(x_i\in C_x))}{\sqrt{\frac{1}{2n}\sum_{i=1}^{n}(I(x_i'\in C_x)+I(x_i\in C_x))}}\right)^2}\right)\\
&\le 4\mathbb{E}\triangle_{\mathcal{C}(x_1,\ldots,x_n,x_1',\ldots,x_n')}(x_1,\ldots,x_n,x_1',\ldots,x_n')\cdot e^{-\frac{nt^2}{4}}\\
&= 4G(2n)e^{-\frac{nt^2}{4}}
\end{aligned}$$

Recall that the solution of SVM is $f(x) = \sum_{i=1}^n \alpha_i K(x_i, x)$, where $(x_1, y_1), \ldots, (x_n, y_n)$ – data, with $y_i \in \{-1, 1\}$. The label is predicted by $\text{sign}(f(x))$ and $\mathbb{P}(yf(x) \leq 0)$ is *misclassification error.*

Let $\mathcal{H} = \mathcal{H}((x_1, y_1), \ldots, (x_n, y_n))$ be random collection of functions, with card $\mathcal{H} \leq \mathcal{N}(n)$. Also, assume that for any $h \in \mathcal{H}$, $-h \in \mathcal{H}$ so that $\alpha$ can be positive.

Define

$$\mathcal{F} = \left\{ \sum_{i=1}^T \lambda_i h_i,\ T \geq 1,\ \lambda_i \geq 0,\ \sum_{i=1}^T \lambda_i = 1,\ h_i \in \mathcal{H} \right\}.$$

For SVM, $\mathcal{H} = \{\pm K(x_i, x) : i = 1, \ldots, n\}$ and card $\mathcal{H} \leq 2n$.

Recall margin-sparsity bound (voting classifiers): algorithm outputs $f = \sum_{i=1}^T \lambda_i h_i$. Take random approximation $g(x) = \frac{1}{k} \sum_{j=1}^k Y_j(x)$, where $Y_1, \ldots, Y_k$ i.i.d with $\mathbb{P}(Y_j = h_i) = \lambda_i$, $\mathbb{E} Y_j(x) = f(x)$.

Fix $\delta > 0$.

$$\begin{aligned}
\mathbb{P}(yf(x) \leq 0) &= \mathbb{P}(yf(x) \leq 0, yg(x) \leq \delta) + \mathbb{P}(yf(x) \leq 0, yg(x) > \delta) \\
&\leq \mathbb{P}(yg(x) \leq \delta) + \mathbb{E}_{x,y} \mathbb{P}_Y \left( y \frac{1}{k} \sum_{j=1}^k Y_j(x) > \delta,\ y \mathbb{E}_Y Y_1(x) \leq 0 \right) \\
&\leq \mathbb{P}(yg(x) \leq \delta) + \mathbb{E}_{x,y} \mathbb{P}_Y \left( \frac{1}{k} \sum_{j=1}^k (y Y_j(x) - \mathbb{E}(y Y_j(x))) \geq \delta \right) \\
&\leq \text{(by Hoeffding) } \mathbb{P}(yg(x) \leq \delta) + \mathbb{E}_{x,y} e^{-k\delta^2/2} \\
&= \mathbb{P}(yg(x) \leq \delta) + e^{-k\delta^2/2} \\
&= \mathbb{E}_Y \mathbb{P}_{x,y}(yg(x) \leq \delta) + e^{-k\delta^2/2}
\end{aligned}$$

Similarly to what we did before, on the data

$$\mathbb{E}_Y \left[ \frac{1}{n} \sum_{i=1}^n I(y_i g(x_i) \leq \delta) \right] \leq \frac{1}{n} \sum_{i=1}^n I(y_i f(x_i) \leq 2\delta) + e^{-k\delta^2/2}$$

Can we bound

$$\mathbb{P}_{x,y}(yg(x) \leq \delta) - \frac{1}{n} \sum_{i=1}^n I(y_i g(x_i) \leq \delta)$$

for any $g$?

Define

$$\mathcal{C} = \{\{yg(x) \leq \delta\},\ g \in \mathcal{F}_k,\ \delta \in [-1, 1]\}$$

where

$$\mathcal{F}_k = \left\{ \frac{1}{k} \sum_{j=1}^k h_j(x) :\ h_j \in \mathcal{H} \right\}$$

Note that $\mathcal{H}(x_1, \ldots, x_n) \subseteq \mathcal{H}(x_1, \ldots, x_n, x_{n+1})$ and $\mathcal{H}(\pi(x_1, \ldots, x_n)) = \mathcal{H}(x_1, \ldots, x_n)$.

In the last lecture, we proved

$$\mathbb{P}_{x,y}\left(\sup_{C\in\mathcal{C}}\frac{\mathbb{P}\left(C\right)-\frac{1}{n}\sum_{i=1}^{n}I(x_i\in C)}{\sqrt{\mathbb{P}\left(C\right)}}\geq t\right)\leq 4G(2n)e^{-\frac{nt^2}{2}}$$

where

$$G(n)=\mathbb{E}\triangle_{\mathcal{C}(x_1,\ldots,x_n)}(x_1,\ldots,x_n).$$

How many different $g$'s are there? At most card $\mathcal{F}_k\leq\mathcal{N}(n)^k$. For a fixed $g$,

$$\text{card}\,\{\{yg(x)\leq\delta\}\cap\{x_1,\ldots,x_n\},\ \delta\in[-1,1]\}\leq(n+1).$$

Indeed, we can order $y_1g(x_1),\ldots,y_ng(x_n)\to y_{i_1}g(x_{i_1})\leq\ldots\leq y_{i_n}g(x_{i_n})$ and level $\delta$ can be anywhere along this chain.

Hence,

$$\triangle_{\mathcal{C}(x_1,\ldots,x_n)}(x_1,\ldots,x_n)\leq\mathcal{N}(n)^k(n+1).$$

$$\begin{aligned}\mathbb{P}_{x,y}\left(\sup_{C\in\mathcal{C}}\frac{\mathbb{P}\left(C\right)-\frac{1}{n}\sum_{i=1}^{n}I(x_i\in C)}{\sqrt{\mathbb{P}\left(C\right)}}\geq t\right)&\leq 4G(2n)e^{-\frac{nt^2}{2}}\\&\leq 4\mathcal{N}(2n)^k(2n+1)e^{-\frac{nt^2}{2}}\end{aligned}$$

Setting the above bound to $e^{-u}$ and solving for $t$, we get

$$t=\sqrt{\frac{2}{n}(u+k\log\mathcal{N}(2n)+\log(8n+4))}$$

So, with probability at least $1-e^{-u}$, for all $C$

$$\frac{\left(\mathbb{P}\left(C\right)-\frac{1}{n}\sum_{i=1}^{n}I(x_i\in C)\right)^2}{\mathbb{P}\left(C\right)}\leq\frac{2}{n}\left(u+k\log\mathcal{N}(2n)+\log(8n+4)\right).$$

In particular,

$$\frac{\left(\mathbb{P}\left(yg(x)\leq\delta\right)-\frac{1}{n}\sum_{i=1}^{n}I(y_ig(x_i)\leq\delta)\right)^2}{\mathbb{P}\left(yg(x)\leq\delta\right)}\leq\frac{2}{n}\left(u+k\log\mathcal{N}(2n)+\log(8n+4)\right).$$

Since $\frac{(x-y)^2}{x}$ is convex with respect to $(x,y)$,

$$\begin{aligned}&\frac{\left(\mathbb{E}_Y\mathbb{P}_{x,y}\left(yg(x)\leq\delta\right)-\mathbb{E}_Y\frac{1}{n}\sum_{i=1}^{n}I(y_ig(x_i)\leq\delta)\right)^2}{\mathbb{E}_Y\mathbb{P}_{x,y}\left(yg(x)\leq\delta\right)}\\&\leq\mathbb{E}_Y\frac{\left(\mathbb{P}\left(yg(x)\leq\delta\right)-\frac{1}{n}\sum_{i=1}^{n}I(y_ig(x_i)\leq\delta)\right)^2}{\mathbb{P}\left(yg(x)\leq\delta\right)}\qquad(32.1)\\&\leq\frac{2}{n}\left(u+k\log\mathcal{N}(2n)+\log(8n+4)\right).\end{aligned}$$

Recall that

$$\mathbb{P}\left(yf(x)\leq 0\right)\leq\mathbb{E}_Y\mathbb{P}\left(yg(x)\leq\delta\right)+e^{-k\delta^2/2}\qquad(32.2)$$

and

$$\mathbb{E}_Y \frac{1}{n}\sum_{i=1}^{n} I(y_i g(x_i) \leq \delta) \leq \frac{1}{n}\sum_{i=1}^{n} I(y_i f(x_i) \leq 2\delta) + e^{-k\delta^2/2}. \tag{32.3}$$

Choose $k$ such that $e^{-k\delta^2/2} = \frac{1}{n}$, i.e. $k = \frac{2\log n}{\delta^2}$. Plug (32.2) and (32.3) into (32.1) (look at $\frac{(a-b)^2}{a}$). Hence,

$$\frac{\left(\mathbb{P}\left(yf(x) \leq 0\right) - \frac{2}{n} - \frac{1}{n}\sum_{i=1}^{n} I(y_i f(x_i) \leq 2\delta)\right)^2}{\mathbb{P}\left(yf(x) \leq 0\right) - \frac{2}{n}} \leq \frac{2}{n}\left(u + \frac{2\log n}{\delta^2}\log\mathcal{N}(2n) + \log(8n+4)\right)$$

with probability at least $1 - e^{-u}$.

Recall that for SVM, $\mathcal{N}(n) = \text{card}\left\{\pm K(x_i, x)\right\} \leq 2n$.

**Lemma 33.1.** *For* $0 \le r \le 1$,

$$\inf_{0\le\lambda\le1} e^{\frac{1}{4}(1-\lambda)^2} r^{-\lambda} \le 2-r.$$

*Proof.* Taking *log*, we need to show

$$\inf_{0\le\lambda\le1}\left(\frac{1}{4}(1-\lambda)^2-\lambda\log r-\log(2-r)\right)\le 0.$$

Taking derivative with respect to $\lambda$,

$$-\frac{1}{2}(1-\lambda)-\log r=0$$

$$\lambda = 1+2\log r \le 1$$

$$0\le\lambda = 1+2\log r$$

Hence,

$$e^{-1/2}\le r.$$

Take

$$\lambda=\begin{cases} 1+2\log r & e^{-1/2}\le r\\ 0 & e^{-1/2}\ge r\end{cases}$$

**Case a)**: $r\le e^{-1/2}$, $\lambda=0$

$\frac{1}{4}-\log(2-r)\le 0 \;\Leftrightarrow\; r\le 2-e^{\frac{1}{4}}$. $\quad e^{-1/2}\le 2-e^{\frac{1}{4}}$.

**Case a)**: $r\ge e^{-1/2}$, $\lambda=1+2\log r$

$$(\log r)^2-\log r-2(\log r)^2-\log(2-r)\le 0$$

Let

$$f(r)=\log(2-r)+\log r+(\log r)^2.$$

Is $f(r)\ge 0$? Enough to prove $f'(r)\le 0$. Is

$$f'(r)=-\frac{1}{2-r}+\frac{1}{r}+2\log r\cdot\frac{1}{r}\le 0.$$

$$rf'(r)=-\frac{r}{2-r}+1+2\log r\le 0.$$

Enough to show $(rf'(r))'\ge 0$:

$$(rf'(r))'=\frac{2}{r}-\frac{2-r+r}{(2-r)^2}=\frac{2}{r}-\frac{2}{(2-r)^2}.$$

□

Let $\mathcal{X}$ be a set (space of examples) and $P$ a probability measure on $\mathcal{X}$. Let $x_1, \ldots, x_n$ be i.i.d., $(x_1, \ldots, x_n) \in \mathcal{X}^n$, $P^n = P \times \ldots \times P$.

Consider a subset $A \in \mathcal{X}^n$. How can we define a distance from $x \in \mathcal{X}^n$ to $A$? Example: hamming distance between two points $d(x, y) = \sum I(x_i \neq y_1)$.

We now define *convex hull distance.*

**Definition 33.1.** *Define $V(A,x)$, $U(A,x)$, and $d(A,x)$ as follows:*

(1) $V(A,x) = \{(s_1, \ldots, s_n) : s_i \in \{0,1\}, \exists y \in A \text{ s.t. if } s_i = 0 \text{ then } x_i = y_i\}$

$$\begin{array}{rcccc} x = ( & x_1, & x_2, & \ldots, & x_n) \\ & = & \neq & \ldots & = \\ y = ( & y_1, & y_2, & \ldots, & y_n) \\ s = ( & 0, & 1, & \ldots, & 0) \end{array}$$

*Note that it can happen that $x_i = y_i$ but $s_i \neq 0$.*

(2) $U(A,x) = conv\, V(A,x) = \left\{\sum \lambda_i u^i,\ u^i = (u_1^i, \ldots, u_n^i) \in V(A,x),\ \lambda_i \geq 0,\ \sum \lambda_i = 1\right\}$

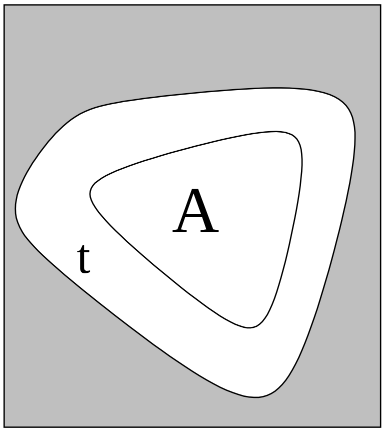
(3) $d(A,x) = \min_{u \in U(A,x)} |u|^2 = \min_{u \in U(A,x)} \sum u_i^2$

**Theorem 33.1.**

$$\mathbb{E} e^{\frac{1}{4} d(A,x)} = \int e^{\frac{1}{4} d(A,x)} dP^n(x) \leq \frac{1}{P^n(A)}$$

*and*

$$P^n(d(A,x) \geq t) \leq \frac{1}{P^n(A)} e^{-t/4}.$$



*Proof.* Proof is by induction on $n$.

$\mathbf{n = 1}$ :

$$d(A,x) = \begin{cases} 0, & x \in A \\ 1, & x \notin A \end{cases}$$

Hence,

$$\int e^{\frac{1}{4}d(A,x)}dP^n(x) = P(A)\cdot 1 + (1-P(A))e^{\frac{1}{4}} \le \frac{1}{P(A)}$$

because

$$e^{\frac{1}{4}} \le \frac{1+P(A)}{P(A)}.$$

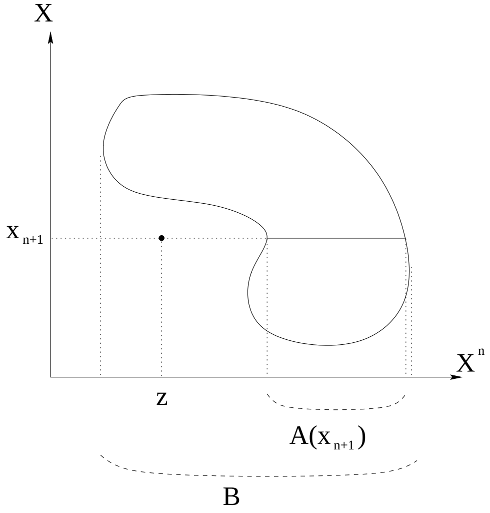
$\mathbf{n \to n+1}$ :

Let $x = (x_1, \ldots, x_n, x_{n+1}) = (z, x_{n+1})$. Define

$$A(x_{n+1}) = \{(y_1, \ldots, y_n) : \ (y_1, \ldots, y_n, x_{n+1}) \in A\}$$

and

$$B = \{(y_1, \ldots, y_n) : \ \exists y_{n+1}, \ (y_1, \ldots, y_n, y_{n+1}) \in A\}$$



One can verify that

$$s \in U(A(x_{n+1}, z)) \Rightarrow (s, 0) \in U(A, (z, x_{n+1}))$$

and

$$t \in U(B, z) \Rightarrow (t, 1) \in U(A, (z, x_{n+1})).$$

Take $0 \le \lambda \le 1$. Then

$$\lambda(s, 0) + (1-\lambda)(t, 1) \in U(A, (z, x_{n+1}))$$

since $U(A,(z,x_{n+1}))$ is convex. Hence,

$$\begin{aligned} d(A,(z,x_{n+1})) = d(A,x) &\leq |\lambda(s,0)+(1-\lambda)(t,1)|^2 \\ &= \sum_{i=1}^{n}(\lambda s_i+(1-\lambda)t_i)^2+(1-\lambda)^2 \\ &\leq \lambda\sum s_i^2+(1-\lambda)\sum t_i^2+(1-\lambda)^2 \end{aligned}$$

So,

$$d(A,x) \leq \lambda d(A(x_{n+1}),z)+(1-\lambda)d(B,z)+(1-\lambda)^2.$$

Now we can use induction:

$$\int e^{\frac{1}{4}d(A,x)}dP^{n+1}(x) = \int_{\mathcal{X}}\int_{\mathcal{X}^n} e^{\frac{1}{4}d(A,(z,x_{n+1}))}dP^n(z)dP(x_{n+1}).$$

Then inner integral is

$$\begin{aligned} \int_{\mathcal{X}^n} e^{\frac{1}{4}d(A,(z,x_{n+1}))}dP^n(z) &\leq \int_{\mathcal{X}^n} e^{\frac{1}{4}\left(\lambda d(A(x_{n+1}),z)+(1-\lambda)d(B,z)+(1-\lambda)^2\right)}dP^n(z) \\ &= e^{\frac{1}{4}(1-\lambda)^2}\int e^{\left(\frac{1}{4}d(A(x_{n+1}),z)\right)\lambda+\left(\frac{1}{4}d(B,z)\right)(1-\lambda)}dP^n(z) \end{aligned}$$

We now use *Hölder*'s inequality:

$$\int fg dP \leq \left(\int f^p dP\right)^{1/p}\left(\int g^q dP\right)^{1/q} \text{ where } \frac{1}{p}+\frac{1}{q}=1$$

$$\begin{aligned} & e^{\frac{1}{4}(1-\lambda)^2}\int e^{\left(\frac{1}{4}d(A(x_{n+1}),z)\right)\lambda+\left(\frac{1}{4}d(B,z)\right)(1-\lambda)}dP^n(z) \\ &\leq e^{\frac{1}{4}(1-\lambda)^2}\left(\int e^{\frac{1}{4}d(A(x_{n+1}),z)}dP^n(z)\right)^{\lambda}\left(e^{\frac{1}{4}d(B,z)}dP^n(z)\right)^{1-\lambda} \\ &\leq \text{(by ind. hypoth.) } e^{\frac{1}{4}(1-\lambda)^2}\left(\frac{1}{P^n(A(x_{n+1}))}\right)^{\lambda}\left(\frac{1}{P^n(B)}\right)^{1-\lambda} \\ &= \frac{1}{P^n(B)}e^{\frac{1}{4}(1-\lambda)^2}\left(\frac{P^n(A(x_{n+1}))}{P^n(B)}\right)^{-\lambda} \end{aligned}$$

Optimizing over $\lambda\in[0,1]$, we use the Lemma proved in the beginning of the lecture with

$$0 \leq r = \frac{P^n(A(x_{n+1}))}{P^n(B)} \leq 1.$$

Thus,

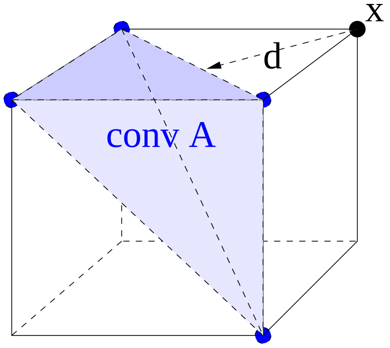
$$\frac{1}{P^n(B)}e^{\frac{1}{4}(1-\lambda)^2}\left(\frac{P^n(A(x_{n+1}))}{P^n(B)}\right)^{-\lambda} \leq \frac{1}{P^n(B)}\left(2-\frac{P^n(A(x_{n+1}))}{P^n(B)}\right).$$

Now, integrate over the last coordinate. When averaging over $x_{n+1}$, we get measure of $A$.

$$\begin{aligned}
\int e^{\frac{1}{4}d(A,x)}dP^{n+1}(x) &= \int_{\mathcal{X}}\int_{\mathcal{X}^n} e^{\frac{1}{4}d(A,(z,x_{n+1}))}dP^n(z)dP(x_{n+1}) \\
&\leq \int_{\mathcal{X}} \frac{1}{P^n(B)}\left(2-\frac{P^n(A(x_{n+1}))}{P^n(B)}\right)dP(x_{n+1}) \\
&= \frac{1}{P^n(B)}\left(2-\frac{P^{n+1}(A)}{P^n(B)}\right) \\
&= \frac{1}{P^{n+1}(A)}\frac{P^{n+1}(A)}{P^n(B)}\left(2-\frac{P^{n+1}(A)}{P^n(B)}\right) \\
&\leq \frac{1}{P^{n+1}(A)}
\end{aligned}$$

because $x(2-x) \leq 1$ for $0 \leq x \leq 1$. □

Let $\mathcal{X} = \{0,1\}$, $(x_1, \ldots, x_n) \in \{0,1\}^n$, $\mathbb{P}(x_i = 1) = p$, and $\mathbb{P}(x_i = 0) = 1 - p$. Suppose $A \subseteq \{0,1\}^n$. What is $d(A, x)$ in this case?



For a given $x$, take all $y \in A$ and compute $s$:

$$\begin{array}{rcccc} x = ( & x_1, & x_2, & \ldots, & x_n) \\ & = & \neq & \ldots & = \\ y = ( & y_1, & y_2, & \ldots, & y_n) \\ s = ( & 0, & 1, & \ldots, & 0) \end{array}$$

Build conv $V(A, x) = U(A, x)$. Finally, $d(A, x) = \min\{|x - u|^2; u \in \text{conv } A\}$

**Theorem 34.1.** *Consider a convex and Lipschitz $f : \mathbb{R}^n \mapsto \mathbb{R}$, $|f(x) - f(y)| \le L|x - y|$, $\forall x, y \in \mathbb{R}^n$. Then*

$$\mathbb{P}\left(f(x_1, \ldots, x_n) \ge M + L\sqrt{t}\right) \le 2e^{-t/4}$$

*and*

$$\mathbb{P}\left(f(x_1, \ldots, x_n) \le M - L\sqrt{t}\right) \le 2e^{-t/4}$$

*where $M$ is median of $f$: $\mathbb{P}(f \ge M) \ge 1/2$ and $\mathbb{P}(f \le M) \ge 1/2$.*

*Proof.* Fix $a \in \mathbb{R}$ and consider $A = \{(x_1, \ldots, x_n) \in \{0,1\}^n,\ f(x_1, \ldots, x_n) \le a\}$. We proved that

$$\mathbb{P}\left(\underbrace{d(A, x) \ge t}_{\text{event } E}\right) \le \frac{1}{\mathbb{P}(A)} e^{-t/4} = \frac{1}{\mathbb{P}(f \le a)} e^{-t/4}$$

$$d(A, x) = \min\{|x - u|^2; u \in \text{conv } A\} = |x - u_0|^2$$

for some $u_0 \in \text{conv } A$. Note that $|f(x) - f(u_0)| \le L|x - u_0|$.

Now, assume that $x$ is such that $d(A, x) \le t$, i.e. complement of event $E$. Then $|x - u_0| = \sqrt{d(A, x)} \le \sqrt{t}$. Hence,

$$|f(x) - f(u_0)| \le L|x - u_0| \le L\sqrt{t}.$$

So, $f(x) \leq f(u_0) + L\sqrt{t}$. What is $f(u_0)$? We know that $u_0 \in \operatorname{conv} A$, so $u_0 = \sum \lambda_i a_i$, $a_i \in A$, and $\lambda_i \geq 0$, $\sum \lambda_i = 1$. Since $f$ is convex,

$$f(u_0) = f\left(\sum \lambda_i a_i\right) \leq \sum \lambda_i f(a_i) \leq \sum \lambda_i a = a.$$

This implies $f(x) \leq a + L\sqrt{t}$. We proved

$$\{d(A,x) \leq t\} \subseteq \{f(x) \leq a + L\sqrt{t}\}.$$

Hence,

$$1 - \frac{1}{\mathbb{P}\left(f \geq a\right)} e^{-t/4} \leq \mathbb{P}\left(d(A,x) \leq t\right) \leq \mathbb{P}\left(f(x) \leq a + L\sqrt{t}\right).$$

Therefore,

$$\mathbb{P}\left(f(x) \geq a + L\sqrt{t}\right) \leq \frac{1}{\mathbb{P}\left(f \geq a\right)} e^{-t/4}.$$

To prove the first inequality take $a = M$. Since $\mathbb{P}\left(f \leq M\right) \geq 1/2$,

$$\mathbb{P}\left(f(x) \geq M + L\sqrt{t}\right) \leq 2e^{-t/4}.$$

To prove the second inequality, take $a = M - L\sqrt{t}$. Then

$$\mathbb{P}\left(f \geq M\right) \leq \frac{1}{\mathbb{P}\left(f \leq M - L\sqrt{t}\right)} e^{-t/4},$$

which means

$$\mathbb{P}\left(f(x) \leq M - L\sqrt{t}\right) \leq 2e^{-t/4}.$$

**Example 34.1.** Let $H \subseteq \mathbb{R}^n$ be a bounded set. Let

$$f(x_1, \ldots, x_n) = \sup_{h \in \mathcal{H}} \left| \sum_{i=1}^n h_i x_i \right|.$$

Let's check:

(1) convexity:

$$\begin{aligned} f(\lambda x + (1-\lambda)y) &= \sup_{h \in \mathcal{H}} \left| \sum_{i=1}^n h_i(\lambda x_i + (1-\lambda)y_i) \right| \\ &= \sup_{h \in \mathcal{H}} \left| \lambda \sum_{i=1}^n h_i x_i + (1-\lambda) \sum_{i=1}^n h_i y_i \right| \\ &\leq \lambda \sup_{h \in \mathcal{H}} \left| \sum_{i=1}^n h_i x_i \right| + (1-\lambda) \sup_{h \in \mathcal{H}} \left| \sum_{i=1}^n h_i y_i \right| \\ &= \lambda f(x) + (1-\lambda) f(y) \end{aligned}$$

(2) Lipschitz:

$$\begin{aligned}
|f(x)-f(y)| &= \left|\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n h_i x_i\right| - \sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n h_i y_i\right|\right| \\
&\leq \sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n h_i(x_i-y_i)\right| \\
&\leq \text{(by Cauchy-Schwartz)} \quad \sup_{h\in\mathcal{H}}\sqrt{\sum h_i^2}\sqrt{\sum (x_i-y_i)^2} \\
&= |x-y| \qquad \underbrace{\sup_{h\in\mathcal{H}}\sqrt{\sum h_i^2}}_{L=\text{Lipschitz constant}}
\end{aligned}$$

We proved the following

**Theorem 34.2.** *If $M$ is the median of $f(x_1,\ldots,x_n)$, and $x_1,\ldots,x_n$ are i.i.d with $\mathbb{P}(x_i=1)=p$ and $\mathbb{P}(x_i=0)=1-p$, then*

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n h_i x_i\right| \geq M + \sup_{h\in\mathcal{H}}\sqrt{\sum h_i^2}\cdot\sqrt{t}\right) \leq 2e^{-t/4}$$

*and*

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n h_i x_i\right| \leq M - \sup_{h\in\mathcal{H}}\sqrt{\sum h_i^2}\cdot\sqrt{t}\right) \leq 2e^{-t/4}$$

□

Assume we have space $\mathcal{X}$ and a class of functions $\mathcal{F} = \{f : \mathcal{X} \mapsto \mathbb{R}\}$, not necessarily bounded. Define

$$Z(x) = Z(x_1, \ldots, x_n) = \sup_{f \in \mathcal{F}} \sum f(x_i)$$

(or $\sup_{f\in\mathcal{F}} |\sum f(x_i)|$).

**Example 35.1.** $f \to \frac{1}{n}(f - \mathbb{E}f)$. $Z(x) = \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}f$.

Consider $x' = (x'_1, \ldots, x'_n)$, an independent copy of $x$. Let

$$V(x) = \mathbb{E}_{x'} \sup_{f\in\mathcal{F}} \sum_{i=1}^n (f(x_i) - f(x'_i))^2$$

be "random uniform variance" (unofficial name)

**Theorem 35.1.**

$$\mathbb{P}\left(Z(x) \geq \mathbb{E}Z(x) + 2\sqrt{V(x)t}\right) \leq 4e \cdot e^{-t/4}$$
$$\mathbb{P}\left(Z(x) \leq \mathbb{E}Z(x) - 2\sqrt{V(x)t}\right) \leq 4e \cdot e^{-t/4}$$

Recall the Symmetrization lemma:

**Lemma 35.1.** $\xi_1, \xi_2, \xi_3(x, x') : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, $\xi'_i = \mathbb{E}_{x'}\xi_i$. *If*

$$\mathbb{P}\left(\xi_1 \geq \xi_2 + \sqrt{\xi_3 t}\right) \leq \Gamma e^{-\gamma t},$$

*then*

$$\mathbb{P}\left(\xi'_1 \geq \xi'_2 + \sqrt{\xi'_3 t}\right) \leq \Gamma e \cdot e^{-\gamma t}.$$

We have

$$\mathbb{E}Z(x) = \mathbb{E}_{x'}Z(x') = \mathbb{E}_{x'} \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(x'_i)$$

and

$$V(x) = \mathbb{E}_{x'} \sup_{f\in\mathcal{F}} \sum_{i=1}^n (f(x_i) - f(x'_i))^2.$$

Use the Symmetrization Lemma with $\xi_1 = Z(x)$, $\xi_2 = Z(x')$, and

$$\xi_3 = \sup_{f\in\mathcal{F}} \sum_{i=1}^n (f(x_i) - f(x'_i))^2.$$

It is enough to prove that

$$\mathbb{P}\left(Z(x) \geq Z(x') + 2\sqrt{t \sup_{f\in\mathcal{F}} \sum_{i=1}^n (f(x_i) - f(x'_i))^2}\right) \leq 4e^{-t/4},$$

i.e.

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \sum_{i=1}^n f(x_i) \geq \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(x'_i) + 2\sqrt{t \sup_{f\in\mathcal{F}} \sum_{i=1}^n (f(x_i) - f(x'_i))^2}\right) \leq 4e^{-t/4}.$$

If we switch $x_i \leftrightarrow x_i'$, nothing changes, so we can switch randomly. Implement the permutation $x_i \leftrightarrow x_i'$:

$$I = f(x_i') + \varepsilon_i(f(x_i) - f(x_i'))$$

$$II = f(x_i) - \varepsilon_i(f(x_i) - f(x_i'))$$

where $\varepsilon_i = 0, 1$. Hence,

(1) If $\varepsilon_i = 1$, then $I = f(x_i)$ and $II = f(x_i')$.

(2) If $\varepsilon_i = 0$, then $I = f(x_i')$ and $II = f(x_i)$.

Take $\varepsilon_1 \ldots \varepsilon_n$ i.i.d. with $\mathbb{P}(\varepsilon_i = 0) = \mathbb{P}(\varepsilon_i = 1) = 1/2$.

$$\begin{aligned}
&\mathbb{P}_{x,x'}\left(\sup_{f\in\mathcal{F}}\sum_{i=1}^n f(x_i) \geq \sup_{f\in\mathcal{F}}\sum_{i=1}^n f(x_i') + 2\sqrt{t\sup_{f\in\mathcal{F}}\sum_{i=1}^n (f(x_i) - f(x_i'))^2}\right) \\
&= \mathbb{P}_{x,x',\varepsilon}\left(\sup_{f\in\mathcal{F}}\sum_{i=1}^n (f(x_i') + \varepsilon_i(f(x_i) - f(x_i'))) \geq \sup_{f\in\mathcal{F}}\sum_{i=1}^n (f(x_i) - \varepsilon_i(f(x_i) - f(x_i')))\right. \\
&\qquad\left. +2\sqrt{t\sup_{f\in\mathcal{F}}\sum_{i=1}^n (f(x_i) - f(x_i'))^2}\right) \\
&= \mathbb{E}_{x,x'}\mathbb{P}_{\varepsilon}\left(\sup_{f\in\mathcal{F}}\ldots \geq \sup_{f\in\mathcal{F}}\ldots + 2\sqrt{\ldots}\text{ for fixed } x, x'\right)
\end{aligned}$$

Define

$$\Phi_1(\varepsilon) = \sup_{f\in\mathcal{F}}\sum_{i=1}^n (f(x_i') + \varepsilon_i(f(x_i) - f(x_i')))$$

and

$$\Phi_2(\varepsilon) = \sup_{f\in\mathcal{F}}\sum_{i=1}^n (f(x_i) - \varepsilon_i(f(x_i) - f(x_i'))).$$

$\Phi_1(\varepsilon), \Phi_2(\varepsilon)$ are convex and Lipschitz with $L = \sup_{f\in\mathcal{F}}\sqrt{\sum_{i=1}^n (f(x_i) - f(x_i'))^2}$. Moreover, $Median(\Phi_1) = Median(\Phi_2)$ and $\Phi_1(\varepsilon_1, \ldots, \varepsilon_n) = \Phi_2(1 - \varepsilon_1, \ldots, 1 - \varepsilon_n)$. Hence,

$$\mathbb{P}_{\varepsilon}\left(\Phi_1 \leq M(\Phi_1) + L\sqrt{t}\right) \geq 1 - 2e^{-t/4}$$

and

$$\mathbb{P}_{\varepsilon}\left(\Phi_2 \leq M(\Phi_2) - L\sqrt{t}\right) \geq 1 - 2e^{-t/4}.$$

With probability at least $1 - 4e^{-t/4}$ both above inequalities hold:

$$\Phi_1 \leq M(\Phi_1) + L\sqrt{t} = M(\Phi_2) + L\sqrt{t} \leq \Phi_2 + 2L\sqrt{t}.$$

Thus,

$$\mathbb{P}_{\varepsilon}\left(\Phi_1 \geq \Phi_2 + 2L\sqrt{t}\right) \leq 4e^{-t/4}$$

and

$$\mathbb{P}_{x,x',\varepsilon}\left(\Phi_1 \geq \Phi_2 + 2L\sqrt{t}\right) \leq 4e^{-t/4}.$$

The "random uniform variance" is

$$V(x) = \mathbb{E}_{x'} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} (f(x_i) - f(x'_i))^2.$$

For example, if $\mathcal{F} = \{f\}$, then

$$\begin{aligned}
\frac{1}{n} V(x) &= \frac{1}{n} \mathbb{E}_{x'} \sum_{i=1}^{n} (f(x_i) - f(x'_i))^2 \\
& \frac{1}{n} \sum_{i=1}^{n} \left( f(x_i)^2 - 2f(x_i)\mathbb{E}f + \mathbb{E}f^2 \right) \\
&= \bar{f^2} - 2\bar{f}\mathbb{E}f + \mathbb{E}f^2 \\
&= \underbrace{\bar{f^2} - (\bar{f})^2}_{\text{sample variance}} + \underbrace{(\bar{f})^2 - 2\bar{f}\mathbb{E}f + (\mathbb{E}f)^2}_{(\bar{f} - \mathbb{E}f)^2} + \underbrace{\mathbb{E}f^2 - (\mathbb{E}f)^2}_{\text{variance}}
\end{aligned}$$

Let $x \in \mathcal{X}^n$. Suppose $A_1, A_2 \subseteq \mathcal{X}^n$. We want to define $d(A_1, A_2, x)$.

**Definition 36.1.**

$$d(A_1, A_2, x) = \inf\{card\ \{i \leq n : x_i \neq y_i^1\ \ and\ x_i \neq y_i^2\}, y^1 \in A_1, y^2 \in A_2\}$$

**Theorem 36.1.**

$$\mathbb{E}2^{d(A_1,A_2,x)} = \int 2^{d(A_1,A_2,x)} dP^n(x) \leq \frac{1}{P^n(A_1)P^n(A_2)}$$

*and*

$$\mathbb{P}\left(d(A_1, A_2, x) \geq t\right) \leq \frac{1}{P^n(A_1)P^n(A_2)} \cdot 2^{-t}$$

We first prove the following lemma:

**Lemma 36.1.** *Let* $0 \leq g_1, g_2 \leq 1$, $g_i : \mathcal{X} \mapsto [0, 1]$. *Then*

$$\int \min\left(2, \frac{1}{g_1(x)}, \frac{1}{g_2(x)}\right) dP(x) \cdot \int g_1(x) dP(x) \cdot \int g_2(x) dP(x) \leq 1$$

*Proof.* Notice that $\log x \leq x - 1$.

So enough to show

$$\int \min\left(2, \frac{1}{g_1}, \frac{1}{g_2}\right) dP + \int g_1 dP + \int g_2 dP \leq 3$$

which is the same as

$$\int \left[\min\left(2, \frac{1}{g_1}, \frac{1}{g_2}\right) + g_1 + g_2\right] dP \leq 3$$

It's enough to show

$$\min\left(2, \frac{1}{g_1}, \frac{1}{g_2}\right) + g_1 + g_2 \leq 3.$$

If min is equal to 2, then $g_1, g_2 \leq \frac{1}{2}$ and the sum is less than 3.

If min is equal to $\frac{1}{g_1}$, then $g_1 \geq \frac{1}{2}$ and $g_1 \geq g_2$, so $\min + g_1 + g_2 \leq \frac{1}{g_1} + 2g_1 \leq 3$. □

We now prove the Theorem:

*Proof.* Proof by induction on $n$.

$\mathbf{n = 1}$ :

$$d(A_1, A_2, x) = 0 \text{ if } x \in A_1 \cup A_2 \ \text{ and } \ d(A_1, A_2, x) = 1 \text{ otherwise}$$

$$\begin{aligned}\int 2^{d(A_1,A_2,x)} dP(x) &= \int \min\left(2, \frac{1}{I(x \in A_1)}, \frac{1}{I(x \in A_2)}\right) dP(x) \\ &\leq \frac{1}{\int I(x \in A_1) dP(x) \cdot \int I(x \in A_2) dP(x)} \\ &= \frac{1}{P(A_1)P(A_2)}\end{aligned}$$

$\mathbf{n \to n+1}$ :

Let $x \in \mathcal{X}^{n+1}$, $A_1, A_2 \subseteq \mathcal{X}^{n+1}$. Denote $x = (x_1, \ldots, x_n, x_{n+1}) = (z, x_{n+1})$.

Define

$$A_1(x_{n+1}) = \{z \in \mathcal{X}^n : (z, x_{n+1}) \in A_1\}$$

$$A_2(x_{n+1}) = \{z \in \mathcal{X}^n : (z, x_{n+1}) \in A_2\}$$

and

$$B_1 = \bigcup_{x_{n+1}} A_1(x_{n+1}), \quad B_2 = \bigcup_{x_{n+1}} A_2(x_{n+1})$$

Then

$$d(A_1, A_2, x) = d(A_1, A_2, (z, x_{n+1})) \leq 1 + d(B_1, B_2, z),$$

$$d(A_1, A_2, (z, x_{n+1})) \leq d(A_1(x_{n+1}), B_2, z),$$

and

$$d(A_1, A_2, (z, x_{n+1})) \leq d(B_1, A_2(x_{n+1}), z).$$

Now,

$$\int 2^{d(A_1, A_2, x)} dP^{n+1}(z, x_{n+1}) = \int \underbrace{\int 2^{d(A_1, A_2, (z, x_{n+1}))} dP^n(z)}_{I(x_{n+1})} dP(x_{n+1})$$

The inner integral ca ne bounded by induction as follows

$$\begin{aligned} I(x_{n+1}) &\leq \int 2^{1+d(B_1, B_2, z)} dP^n(z) \\ &= 2 \int 2^{d(B_1, B_2, z)} dP^n(z) \\ &\leq 2 \cdot \frac{1}{P^n(B_1) P^n(B_2)} \end{aligned}$$

Moreover, by induction,

$$I(x_{n+1}) \leq \int 2^{d(A_1(x_{n+1}), B_2, z)} dP^n(z) \leq \frac{1}{P^n(A_1(x_{n+1})) P^n(B_2)}$$

and

$$I(x_{n+1}) \leq \int 2^{d(B_1, A_2(x_{n+1}), z)} dP^n(z) \leq \frac{1}{P^n(B_1) P^n(A_2(x_{n+1}))}$$

Hence,

$$\begin{aligned} I(x_{n+1}) &\leq \min\left(\frac{2}{P^n(B_1)P^n(B_2)}, \frac{1}{P^n(A_1(x_{n+1}))P^n(B_2)}, \frac{1}{P^n(B_1)P^n(A_2(x_{n+1}))}\right) \\ &= \frac{1}{P^n(B_1)P^n(B_2)} \min\left(2, \underbrace{\frac{1}{P^n(A_1(x_{n+1})/P^n(B_1)}}_{1/g_1(x_{n+1})}, \underbrace{\frac{1}{P^n(A_2(x_{n+1})/P^n(B_2)}}_{1/g_2(x_{n+1})}\right) \end{aligned}$$

So,

$$\begin{aligned}\int I(x_{n+1})dP(x_{n+1}) &\le \frac{1}{P^n(B_1)P^n(B_2)}\int \min\left(2, \frac{1}{g_1}, \frac{1}{g_2}\right)dP\\ &\le \frac{1}{P^n(B_1)P^n(B_2)}\cdot\frac{1}{\int g_1 dP\cdot\int g_2 dP}\\ &= \frac{1}{P^n(B_1)P^n(B_2)}\cdot\frac{1}{P^{n+1}(A_1)/P^n(B_1)\cdot P^{n+1}(A_2)/P^n(B_2)}\\ &= \frac{1}{P^{n+1}(A_1)P^{n+1}(A_2)}\end{aligned}$$

because $\int P^n(A_1(x_{n+1}))dP(x_{n+1}) = P^{n+1}(A_1)$. □

**Lemma 37.1.** *Let*

$$V(x) = \mathbb{E}_{x'} \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} (f(x_i) - f(x'_i))^2$$

*and $a \le f \le b$ for all $f \in \mathcal{F}$. Then*

$$\mathbb{P}\left(V \le 4\mathbb{E}V + (b-a)^2 t\right) \ge 1 - 4\cdot 2^{-t}.$$

*Proof.* Consider $M$-median of $V$, i.e. $\mathbb{P}(V \ge M) \ge 1/2$, $\mathbb{P}(V \le M) \ge 1/2$. Let $A = \{y \in \mathcal{X}^n, V(y) \le M\} \subseteq \mathcal{X}^n$. Hence, $A$ consists of points with typical behavior. We will use control by 2 points to show that any other point is close to these two points.

By control by 2 points,

$$\mathbb{P}(d(A,A,x) \ge t) \le \frac{1}{\mathbb{P}(A)\,\mathbb{P}(A)} \cdot 2^{-t} \le 4\cdot 2^{-t}$$

Take any $x \in \mathcal{X}^n$. With probability at least $1 - 4\cdot 2^{-t}$, $d(A,A,x) \le t$. Hence, we can find $y^1 \in A, y^2 \in A$ such that $\text{card}\,\{i \le n, x_i \ne y_i^1, x_i \ne y_i^2\} \le t$.

Let

$$I_1 = \{i \le n : x_i = y_i^1\},\ \ I_2 = \{i \le n : x_i \ne y_i^1, x_i = y_i^2\},$$

and

$$I_3 = \{i \le n : x_i \ne y_i^1, x_i \ne y_i^2\}$$

Then we can decompose $V$ as follows

$$\begin{aligned}
V(x) &= \mathbb{E}_{x'} \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} (f(x_i) - f(x'_i))^2 \\
&= \mathbb{E}_{x'} \sup_{f\in\mathcal{F}} \left[ \sum_{i\in I_1} (f(x_i) - f(x'_i))^2 + \sum_{i\in I_2} (f(x_i) - f(x'_i))^2 + \sum_{i\in I_3} (f(x_i) - f(x'_i))^2 \right] \\
&\le \mathbb{E}_{x'} \sup_{f\in\mathcal{F}} \sum_{i\in I_1} (f(x_i) - f(x'_i))^2 + \mathbb{E}_{x'} \sup_{f\in\mathcal{F}} \sum_{i\in I_2} (f(x_i) - f(x'_i))^2 + \mathbb{E}_{x'} \sup_{f\in\mathcal{F}} \sum_{i\in I_3} (f(x_i) - f(x'_i))^2 \\
&\le \mathbb{E}_{x'} \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} (f(y_i^1) - f(x'_i))^2 + \mathbb{E}_{x'} \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} (f(y_i^2) - f(x'_i))^2 + (b-a)^2 t \\
&= V(y^1) + V(y^2) + (b-a)^2 t \\
&\le M + M + (b-a)^2 t
\end{aligned}$$

because $y^1, y^2 \in A$. Hence,

$$\mathbb{P}\left(V(x) \le 2M + (b-a)^2 t\right) \ge 1 - 4\cdot 2^{-t}.$$

Finally, $M \le 2\mathbb{E}V$ because

$$\mathbb{P}(V \ge 2\mathbb{E}V) \le \frac{\mathbb{E}V}{2\mathbb{E}V} = \frac{1}{2} \qquad \text{while} \qquad \mathbb{P}(V \ge M) \ge \frac{1}{2}.$$

□

Now, let $Z(x) = \sup_{f\in\mathcal{F}} |\sum_{i=1}^n f(x_i)|$. Then

$$Z(x) \underbrace{\leq}_{\text{with prob. } \geq 1-(4e)e^{-t/4}} \mathbb{E}Z + 2\sqrt{V(x)t} \underbrace{\leq}_{\text{with prob. } \geq 1-4\cdot 2^{-t}} \mathbb{E}Z + 2\sqrt{(4\mathbb{E}V + (b-a)^2t)t}.$$

Using inequality $\sqrt{c+d} \leq \sqrt{c} + \sqrt{d}$,

$$Z(x) \leq \mathbb{E}Z + 4\sqrt{\mathbb{E}Vt} + 2(b-a)t$$

with high probability.

We proved Talagrand's concentration inequality for empirical processes:

**Theorem 37.1.** *Assume $a \leq f \leq b$ for all $f \in \mathcal{F}$. Let $Z = \sup_{f\in\mathcal{F}} |\sum_{i=1}^n f(x_i)|$ and $V = \sup_{f\in\mathcal{F}} \sum_{i=1}^n (f(x_i) - f(x_i'))^2$. Then*

$$\mathbb{P}\left(Z \leq \mathbb{E}Z + 4\sqrt{\mathbb{E}Vt} + 2(b-a)t\right) \geq 1 - (4e)e^{-t/4} - 4\cdot 2^{-t}.$$

This is an analog of Bernstein's inequality:

$$4\sqrt{\mathbb{E}Vt} \longrightarrow \text{ Gaussian behavior}$$

$$2(b-a)t \longrightarrow \text{ Poisson behavior}$$

Now, consider the following lower bound on $V$.

$$\begin{aligned}
V &= \mathbb{E}\sup_{f\in\mathcal{F}} \sum_{i=1}^n (f(x_i) - f(x_i'))^2 \\
&> \sup_{f\in\mathcal{F}} \mathbb{E}\sum_{i=1}^n (f(x_i) - f(x_i'))^2 \\
&= \sup_{f\in\mathcal{F}} n\mathbb{E}(f(x_1) - f(x_1'))^2 \\
&= \sup_{f\in\mathcal{F}} 2n\mathrm{Var}(f) = 2n \sup_{f\in\mathcal{F}} \mathrm{Var}(f) = 2n\sigma^2
\end{aligned}$$

As for the upper bound,

$$\begin{aligned}
\mathbb{E}\sup_{f\in\mathcal{F}} \sum_{i=1}^n (f(x_i) - f(x_i'))^2 &= \mathbb{E}\sup_{f\in\mathcal{F}} \left(\sum_{i=1}^n (f(x_i) - f(x_i'))^2 - 2n\mathrm{Var}(f) + 2n\mathrm{Var}(f)\right) \\
&\leq \mathbb{E}\sup_{f\in\mathcal{F}} \sum_{i=1}^n \left[(f(x_i) - f(x_i'))^2 - \mathbb{E}(f(x_i) - f(x_i'))^2\right] + 2n\sup_{f\in\mathcal{F}} \mathrm{Var}(f) \\
&\text{(by symmetrization)} \\
&\leq 2\mathbb{E}\sup_{f\in\mathcal{F}} \sum_{i=1}^n \varepsilon_i (f(x_i) - f(x_i'))^2 + 2n\sigma^2 \\
&\leq 2\mathbb{E}\left(\sup_{f\in\mathcal{F}} \sum_{i=1}^n \varepsilon_i (f(x_i) - f(x_i'))^2\right)_+ + 2n\sigma^2
\end{aligned}$$

Note that the square function $[-(b-a),(b-a)] \mapsto \mathbb{R}$ is a contraction. Its largest derivative on $[-(b-a),(b-a)]$ is at most $2(b-a)$. Note that $|f(x_i) - f(x_i')| \le b-a$. Hence,

$$\begin{aligned}
2\mathbb{E}\left(\sup_{f\in\mathcal{F}}\sum_{i=1}^n \varepsilon_i (f(x_i)-f(x_i'))^2\right)_+ + 2n\sigma^2 &\le 2\cdot 2(b-a)\mathbb{E}\left(\sup_{f\in\mathcal{F}}\sum_{i=1}^n \varepsilon_i (f(x_i)-f(x_i'))\right)_+ + 2n\sigma^2 \\
&\le 4(b-a)\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^n \varepsilon_i |f(x_i)-f(x_i')| + 2n\sigma^2 \\
&\le 4(b-a)\cdot 2\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^n \varepsilon_i |f(x_i)| + 2n\sigma^2 \\
&= 8(b-a)\mathbb{E}Z + 2n\sigma^2
\end{aligned}$$

We have proved the following

**Lemma 37.2.**

$$\mathbb{E}V \le 8(b-a)\mathbb{E}Z + 2n\sigma^2,$$

*where* $\sigma^2 = \sup_{f\in\mathcal{F}} \mathrm{Var}(f)$.

**Corollary 37.1.** *Assume* $a \le f \le b$ *for all* $f \in \mathcal{F}$. *Let* $Z = \sup_{f\in\mathcal{F}} |\sum_{i=1}^n f(x_i)|$ *and* $\sigma^2 = \sup_{f\in\mathcal{F}} \mathrm{Var}(f)$. *Then*

$$\mathbb{P}\left(Z \le \mathbb{E}Z + 4\sqrt{(8(b-a)\mathbb{E}Z + 2n\sigma^2)t} + 2(b-a)t\right) \ge 1 - (4e)e^{-t/4} - 4\cdot 2^{-t}.$$

Using other approaches, one can get better constants:

$$\mathbb{P}\left(Z \le \mathbb{E}Z + \sqrt{(4(b-a)\mathbb{E}Z + 2n\sigma^2)t} + (b-a)\frac{t}{3}\right) \ge 1 - e^{-t}.$$

If we substitute $f - \mathbb{E}f$ instead of $f$, the result of Lecture 37 becomes:

$$\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n (f(x_i)-\mathbb{E}f)\right| \le \mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n (f(x_i)-\mathbb{E}f)\right| + \sqrt{\left(4(b-a)\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n (f(x_i)-\mathbb{E}f)\right| + 2n\sigma^2\right)t} + (b-a)\frac{t}{3}$$

with probability at least $\ge 1-e^{-t}$. Here, $a \le f \le b$ for all $f \in \mathcal{F}$ and $\sigma^2 = \sup_{f\in\mathcal{F}} \mathrm{Var}(f)$.

Now divide by $n$ to get

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(x_i)-\mathbb{E}f\right| \le \mathbb{E}\sup_{f\in\mathcal{F}}|\ldots| + \sqrt{\left(4(b-a)\mathbb{E}\sup_{f\in\mathcal{F}}|\ldots| + 2\sigma^2\right)\frac{t}{n}} + (b-a)\frac{t}{3n}$$

Compare this result to the Martingale-difference method (McDiarmid):

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(x_i)-\mathbb{E}f\right| \le \mathbb{E}\sup_{f\in\mathcal{F}}|\ldots| + \sqrt{\frac{2(b-a)^2 t}{n}}$$

The term $2(b-a)^2$ is worse than $4(b-a)\mathbb{E}\sup_{f\in\mathcal{F}}|\ldots| + 2\sigma^2$.

An algorithm outputs $f_0 \in \mathcal{F}$, $f_0$ depends on data $x_1, \ldots, x_n$. What is $\mathbb{E}f_0$? Assume $0 \le f \le 1$ (loss function). Then

$$\left|\mathbb{E}f_0 - \frac{1}{n}\sum_{i=1}^n f_0(x_i)\right| \le \sup_{f\in\mathcal{F}}\left|\mathbb{E}f - \frac{1}{n}\sum_{i=1}^n f(x_i)\right| \le \text{ use Talagrand's inequality }.$$

What if we knew that $\mathbb{E}f_0 \le \varepsilon$ and the family $\mathcal{F}_\varepsilon = \{f \in \mathcal{F}, \mathbb{E}f \le \varepsilon\}$ is much smaller than $\mathcal{F}$. Then looking at $\sup_{f\in\mathcal{F}}\left|\mathbb{E}f - \frac{1}{n}\sum_{i=1}^n f(x_i)\right|$ is too conservative.

Pin down location of $f_0$. Pretend we know $\mathbb{E}f_0 \le \varepsilon$, $f_0 \in \mathcal{F}_\varepsilon$. Then with probability at least $1-e^{-t}$,

$$\begin{aligned}\left|\mathbb{E}f_0 - \frac{1}{n}\sum_{i=1}^n f_0(x_i)\right| &\le \sup_{f\in\mathcal{F}_\varepsilon}\left|\mathbb{E}f - \frac{1}{n}\sum_{i=1}^n f(x_i)\right| \\ &\le \mathbb{E}\sup_{f\in\mathcal{F}_\varepsilon}\left|\mathbb{E}f - \frac{1}{n}\sum_{i=1}^n f(x_i)\right| + \sqrt{\left(4\mathbb{E}\sup_{f\in\mathcal{F}_\varepsilon}|\ldots| + 2\sigma_\varepsilon^2\right)\frac{t}{n}} + \frac{t}{3n}\end{aligned}$$

where $\sigma_\varepsilon^2 = \sup_{f\in\mathcal{F}_\varepsilon}\mathrm{Var}(f)$. Note that for $f \in \mathcal{F}_\varepsilon$

$$\mathrm{Var}(f) = \mathbb{E}f^2 - (\mathbb{E}f)^2 \le \mathbb{E}f^2 \le \mathbb{E}f \le \varepsilon$$

since $0 \le f \le 1$.

Denote $\varphi(\varepsilon) = \mathbb{E}\sup_{f\in\mathcal{F}_\varepsilon}\left|\mathbb{E}f - \frac{1}{n}\sum_{i=1}^n f(x_i)\right|$. Then

$$\left|\mathbb{E}f_0 - \frac{1}{n}\sum_{i=1}^n f_0(x_i)\right| \le \varphi(\varepsilon) + \sqrt{(4\varphi(\varepsilon) + 2\varepsilon)\frac{t}{n}} + \frac{t}{3n}$$

with probability at least $1-e^{-t}$.

Take $\varepsilon = 2^{-k}$, $k = 0, 1, 2, \ldots$. Change $t \to t + 2\log(k+2)$. Then, for a fixed $k$, with probability at least $1 - e^{-t}\frac{1}{(k+2)^2}$,

$$\left|\mathbb{E}f_0 - \frac{1}{n}\sum_{i=1}^n f_0(x_i)\right| \le \varphi(\varepsilon) + \sqrt{(4\varphi(\varepsilon) + 2\varepsilon)\frac{t + 2\log(k+2)}{n}} + \frac{t + 2\log(k+2)}{3n}$$

For all $k \ge 0$, the statement holds with probability at least

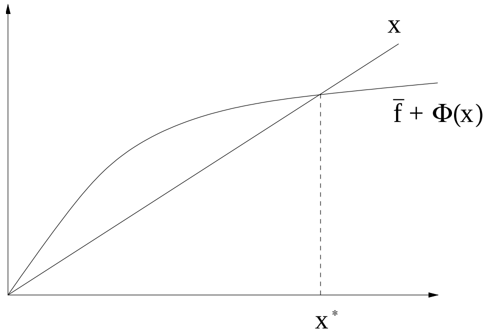
$$1 - \underbrace{\sum_{k=1}^{\infty}\frac{1}{(k+2)^2}}_{\frac{\pi^2}{6}-1} e^{-t} \ge 1 - e^{-t}$$

For $f_0$, find $k$ such that $2^{-k-1} \le \mathbb{E}f_0 < 2^{-k}$ (hence, $2^{-k} \le 2\mathbb{E}f_0$). Use the statement for $\varepsilon_k = 2^{-k}$, $k \le \log_2 \frac{1}{\mathbb{E}f_0}$.

$$\left|\mathbb{E}f_0 - \frac{1}{n}\sum_{i=1}^n f_0(x_i)\right| \le \varphi(\varepsilon_k) + \sqrt{(4\varphi(\varepsilon_k) + 2\varepsilon_k)\frac{t + 2\log(k+2)}{n}} + \frac{t + 2\log(k+2)}{3n}$$

$$\le \varphi(2\mathbb{E}f_0) + \sqrt{(4\varphi(2\mathbb{E}f_0) + 4\mathbb{E}f_0)\frac{t + 2\log(\log_2\frac{1}{\mathbb{E}f_0} + 2)}{n}} + \frac{t + 2\log(\log_2\frac{1}{\mathbb{E}f_0} + 2)}{3n} = \Phi(\mathbb{E}f_0)$$

Hence, $\mathbb{E}f_0 \le \frac{1}{n}\sum_{i=1}^n f_0(x_i) + \Phi(\mathbb{E}f_0)$. Denote $x = \mathbb{E}f_0$. Then $x \le \bar{f} + \Phi(x)$.



**Theorem 38.1.** *Let $0 \le f \le 1$ for all $f \in \mathcal{F}$. Define $\mathcal{F}_\varepsilon = \{f \in \mathcal{F}, \mathbb{E}f \le \varepsilon\}$ and $\varphi(\varepsilon) = \mathbb{E}\sup_{f\in\mathcal{F}_\varepsilon}\left|\mathbb{E}f - \frac{1}{n}\sum_{i=1}^n f(x_i)\right|$. Then, with probability at least $1 - e^{-t}$, for any $f_0 \in \mathcal{F}$, $\mathbb{E}f_0 \le x^*$, where $x^*$ is the largest solution of*

$$x^* = \frac{1}{n}\sum_{i=1}^n f_0(x_i) + \Phi(x^*).$$

Main work is to find $\varphi(\varepsilon)$. Consider the following example.

**Example 38.1.** If

$$\sup_{x_1,\ldots,x_n}\log\mathcal{D}(\mathcal{F}, u, d_x) \le \mathcal{D}(\mathcal{F}, u),$$

then

$$\mathbb{E}\sup_{f\in\mathcal{F}_\varepsilon}\left|\mathbb{E}f - \frac{1}{n}\sum_{i=1}^n f(x_i)\right| \le \frac{k}{\sqrt{n}}\int_0^{\sqrt{\varepsilon}}\log^{1/2}\mathcal{D}(\mathcal{F}, \varepsilon)d\varepsilon.$$

Let $x \in \mathcal{X}^n$, $x = (x_1, \ldots, x_n)$. Suppose $A \subseteq \mathcal{X}^n$. Define

$$V(A, x) = \{(I(x_1 \neq y_1), \ldots, I(x_n \neq y_n)) : y = (y_1, \ldots, y_n) \in A\},$$

$$U(A, x) = \text{conv } V(A, x)$$

and

$$d(A, x) = \min\{|s|^2 = \sum_{i=1}^n s_i^2,\ s \in U(A, x)\}$$

In the previous lectures, we proved

**Theorem 39.1.**

$$\mathbb{P}\left(d(A, x) \geq t\right) \leq \frac{1}{\mathbb{P}\left(A\right)} e^{-t/4}.$$

Today, we prove

**Theorem 39.2.** *The following are equivalent:*

(1) $d(A, x) \leq t$

(2) $\forall \alpha = (\alpha_1, \ldots, \alpha_n), \exists y \in A,$ *s.t.* $\sum_{i=1}^n \alpha_i I(x_i \neq y_i) \leq \sqrt{\sum_{i=1}^n \alpha_i^2 \cdot t}$

*Proof.* **(1)⇒(2):**

Choose any $\alpha = (\alpha_1, \ldots, \alpha_n)$.

$$\min_{y \in A} \sum_{i=1}^n \alpha_i I(x_i \neq y_i) = \min_{s \in U(A,x)} \sum_{i=1}^n \alpha_i s_i \leq \sum_{i=1}^n \alpha_i s_i^0 \tag{39.1}$$

$$\leq \sqrt{\sum_{i=1}^n \alpha_i^2} \sqrt{\sum_{i=1}^n (s_i^0)^2} \leq \sqrt{\sum_{i=1}^n \alpha_i^2 \cdot t} \tag{39.2}$$

where in the last inequality we used assumption (1). In the above, min is achieved at $s^0$.

**(2)⇒(1):**

Let $\alpha = (s_1^0, \ldots, s_n^0)$. There exists $y \in A$ such that

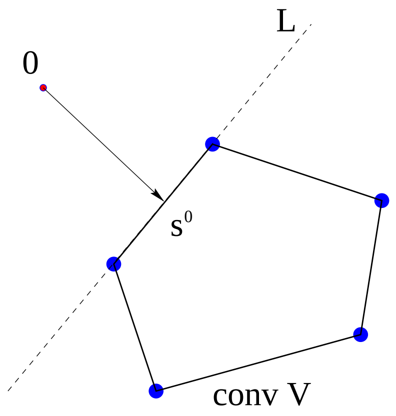
$$\sum_{i=1}^n \alpha_i I(x_i \neq y_i) \leq \sqrt{\sum_{i=1}^n \alpha_i^2 \cdot t}$$

Note that $\sum \alpha_i s_i^0$ is constant on $L$ because $s^0$ is perpendicular to the face.

$$\sum \alpha_i s_i^0 \leq \sum \alpha_i I(x_i \neq y_i) \leq \sqrt{\sum \alpha_i^2 t}$$

Hence, $\sum (s_i^0)^2 \leq \sqrt{\sum (s_i^0)^2 t}$ and $\sqrt{\sum (s_i^0)^2} \leq \sqrt{t}$. Therefore, $d(A, x) \leq \sum (s_i^0)^2 \leq t$. □

We now turn to an application of the above results: Bin Packing.



**Example 39.1.** Assume we have $x_1, \ldots, x_n$, $0 \le x_i \le 1$, and let $B(x_1, \ldots, x_n)$ be the smallest number of bins of size 1 needed to pack all $(x_1, \ldots, x_n)$. Let $S_1, \ldots, S_B \subseteq \{1, \ldots, n\}$ such that all $x_i$ with $i \in S_k$ are packed into one bin, $\bigcup S_k = \{1, \ldots, n\}$, $\sum_{i \in S_k} x_i \le 1$.

**Lemma 39.1.** $B(x_1, \ldots, x_n) \le 2\sum x_i + 1$.

*Proof.* For all but one $k$, $\frac{1}{2} \le \sum_{i \in S_k} x_i$. Otherwise we can combine two bins into one. Hence, $B - 1 \le 2\sum_k \sum_{i \in S_k} x_i = 2\sum x_i$ □

**Theorem 39.3.**

$$\mathbb{P}\left(B(x_1, \ldots, x_n) \le M + 2\sqrt{\sum x_i^2 \cdot t} + 1\right) \ge 1 - 2e^{-t/4}.$$

*Proof.* Let $A = \{y : B(y_1, \ldots, y_n) \le M\}$, where $\mathbb{P}(B \ge M) \ge 1/2$, $\mathbb{P}(B \le M) \ge 1/2$. We proved that

$$\mathbb{P}(d(A, x) \ge t) \le \frac{1}{\mathbb{P}(A)} e^{-t/4}.$$

Take $x$ such that $d(A, x) \le t$. Take $\alpha = (x_1, \ldots, x_n)$. Since $d(A, x) \le t$, there exists $y \in A$ such that $\sum x_i I(x_i \ne y_i) \le \sqrt{\sum x_i^2 \cdot t}$.

To pack the set $\{i : x_i = y_i\}$ we need $\le B(y_1, \ldots, y_n) \le M$ bins.

To pack $\{i : x_i \ne y_i\}$:

$$\begin{aligned} B(x_1 I(x_1 \ne y_1), \ldots, x_n I(x_n \ne y_n)) &\le 2\sum x_i I(x_i \ne y_i) + 1 \\ &\le 2\sqrt{\sum x_i^2 \cdot t} + 1 \end{aligned}$$

by Lemma.

Hence,

$$B(x_1, \ldots, x_n) \le M + 2\sqrt{\sum x_i^2 \cdot t} + 1$$

with probability at least $1 - 2e^{-t/4}$.

By Bernstein's inequality we get

$$\mathbb{P}\left(\sum x_i^2 \le n\mathbb{E}x_1^2 + \sqrt{n\mathbb{E}x_1^2 \cdot t} + \frac{2}{3}t\right) \ge 1 - e^{-t}.$$

Hence,

$$B(x_1, \ldots, x_n) \lesssim M + 2\sqrt{n\mathbb{E}x_1^2 \cdot t}$$

□

In this lecture, we expose the technique of deriving concentration inequalities with the entropy tensorization inequality. The entropy tensorization inequality enables us to bound the entropy of a function of $n$ variables by the sum of the $n$ entropies of this function in terms of the individual variables. The second step of this technique uses the variational formulation of the $n$ entropies to form a differential inequality that gives an upper bound of the log-Laplace transform of the function. We can subsequently use Markov inequality to get a deviation inequality involving this function.

Let $(\mathcal{X}, \mathcal{F}, \mathbb{P})$ be a measurable space, and $u : \mathcal{X} \to \mathbb{R}^+$ a measurable function. The **entropy** of $u$ with regard to $\mathbb{P}$ is defined as $\mathrm{Ent}_{\mathbb{P}}(u) \stackrel{\text{def.}}{=} \int u \log u d\mathbb{P} - \int u \cdot \left(\log\left(\int u d\mathbb{P}\right)\right) d\mathbb{P}$. If $\mathbb{Q}$ is another probability measure and $u = \frac{d\mathbb{Q}}{d\mathbb{P}}$, then $\mathrm{Ent}_{\mathbb{P}}(u) = \int \left(\log \frac{d\mathbb{Q}}{d\mathbb{P}}\right) d\mathbb{Q}$ is the **KL-divergence** between two probability measures $\mathbb{Q}$ and $\mathbb{P}$. The following lemma gives variational formulations for the entropy.

**Lemma 40.1.**

$$\begin{aligned} Ent_{\mathbb{P}}(u) &= \inf\left\{\int \left(u \cdot (\log u - \log x) - (u - x)\right) d\mathbb{P} : x \in \mathbb{R}^+\right\} \\ &= \sup\left\{\int (u \cdot g)\, d\mathbb{P} : \int \exp(g) d\mathbb{P} \le 1\right\}. \end{aligned}$$

*Proof.* For the first formulation, we define $x$ pointsizely by $\frac{\partial}{\partial x} \int \left(u \cdot (\log u - \log x) - (u - x)\right) d\mathbb{P} = 0$, and get $x = \int u d\mathbb{P} > 0$.

For the second formulation, the Laplacian corresponding to $\sup\left\{\int (u \cdot g)\, d\mathbb{P} : \int \exp(g) d\mathbb{P} \le 1\right\}$ is $\mathcal{L}(g, \lambda) = \int (ug)\, d\mathbb{P} - \lambda\left(\int \exp(g) d\mathbb{P} - 1\right)$. It is linear in $\lambda$ and concave in $g$, thus $\sup_g \inf_{\lambda \ge 0} \mathcal{L} = \inf_{\lambda \ge 0} \sup_g \mathcal{L}$. Define $g$ pointwisely by $\frac{\partial}{\partial g}\mathcal{L} = u - \lambda \exp(g) = 0$. Thus $g = \log \frac{u}{\lambda}$, and $\sup_g \mathcal{L} = \int \left(u \log \frac{u}{\lambda}\right) d\mathbb{P} - \int u d\mathbb{P} + \lambda$. We set $\frac{\partial}{\partial \lambda} \sup_g \mathcal{L} = -\frac{\int u d\mathbb{P}}{\lambda} + 1 = 0$, and get $\lambda = \int u d\mathbb{P}$. As a result, $\inf_\lambda \sup_g \mathcal{L} = \mathrm{Ent}_{\mathbb{P}}(u)$. □

Entropy $\mathrm{Ent}_{\mathbb{P}}(u)$ is a convex function of $u$ for any probability measure $\mathbb{P}$, since

$$\begin{aligned} \mathrm{Ent}_{\mathbb{P}}(\sum \lambda_i u_i) &= \sup\left\{\int \left(\sum \lambda_i u_i \cdot g\right) d\mathbb{P} : \int \exp(g) d\mathbb{P} \le 1\right\} \\ &\le \sum \lambda_i \sup\left\{\int (u_i \cdot g_i)\, d\mathbb{P} : \int \exp(g_i) d\mathbb{P} \le 1\right\} \\ &= \sum \lambda_i \mathrm{Ent}_{\mathbb{P}}(u_i). \end{aligned}$$

**Lemma 40.2.** *[Tensorization of entropy]* $\mathcal{X} = (\mathcal{X}_1, \cdots, \mathcal{X}_n)$, $\mathbb{P}^n = \mathbb{P}_1 \times \cdots \times \mathbb{P}_n$, $u = u(x_1, \cdots, x_n)$, $Ent_{\mathbb{P}^n}(u) \le \int \left(\sum_{i=1}^n Ent_{\mathbb{P}_i}(u)\right) d\mathbb{P}^n$.

*Proof.* Proof by induction. When $n = 1$, the above inequality is trivially true. Suppose

$$\int u \log u d\mathbb{P}^n \le \int u d\mathbb{P}^n \log \int u d\mathbb{P}^n + \int \sum_{i=1}^n \mathrm{Ent}_{\mathbb{P}_i}(u) d\mathbb{P}^n.$$

Integrate with regard to $\mathbb{P}_{n+1}$,

$$\begin{aligned}
& \int u \log u d\mathbb{P}^{n+1} \\
\leq \quad & \int \left( \overbrace{\int u d\mathbb{P}^n}^{v} \log \overbrace{\int u d\mathbb{P}^n}^{v} \right) d\mathbb{P}_{n+1} + \int \sum_{i=1}^{n} \mathrm{Ent}_{\mathbb{P}_i}(u) d\mathbb{P}^{n+1} \\
\underbrace{=}_{\text{definition of entropy}} \quad & \int \overbrace{\int u d\mathbb{P}^n}^{v} d\mathbb{P}_{n+1} \cdot \left( \log \int \overbrace{\int u d\mathbb{P}^n}^{v} d\mathbb{P}_{n+1} \right) + \mathrm{Ent}_{\mathbb{P}_{n+1}} \left( \overbrace{\int u d\mathbb{P}^n}^{v} \right) + \int \sum_{i=1}^{n} \mathrm{Ent}_{\mathbb{P}_i}(u) d\mathbb{P}^{n+1} \\
\underbrace{=}_{\text{Foubini's theorem}} \quad & \int u d\mathbb{P}^{n+1} \cdot \left( \log \int u d\mathbb{P}^{n+1} \right) + \mathrm{Ent}_{\mathbb{P}_{n+1}} \left( \int u d\mathbb{P}^n \right) + \int \sum_{i=1}^{n} \mathrm{Ent}_{\mathbb{P}_i}(u) d\mathbb{P}^{n+1} \\
\underbrace{\leq}_{\text{convexity of entropy}} \quad & \int u d\mathbb{P}^{n+1} \cdot \left( \log \int u d\mathbb{P}^{n+1} \right) + \int \mathrm{Ent}_{\mathbb{P}_{n+1}}(u) d\mathbb{P}^n + \int \sum_{i=1}^{n} \mathrm{Ent}_{\mathbb{P}_i}(u) d\mathbb{P}^{n+1} \\
= \quad & \int u d\mathbb{P}^{n+1} \cdot \left( \log \int u d\mathbb{P}^{n+1} \right) + \int \mathrm{Ent}_{\mathbb{P}_{n+1}}(u) d\mathbb{P}^{n+1} + \int \sum_{i=1}^{n} \mathrm{Ent}_{\mathbb{P}_i}(u) d\mathbb{P}^{n+1} \\
\leq \quad & \int u d\mathbb{P}^{n+1} \cdot \left( \log \int u d\mathbb{P}^{n+1} \right) + \int \sum_{i=1}^{n+1} \mathrm{Ent}_{\mathbb{P}_i}(u) d\mathbb{P}^{n+1}.
\end{aligned}$$

By definition of entropy, $\mathrm{Ent}_{\mathbb{P}^{n+1}}(u) \leq \int \sum_{i=1}^{n+1} \mathrm{Ent}_{\mathbb{P}_i}(u) d\mathbb{P}^{n+1}$. □

The tensorization of entropy lemma can be trivially applied to get the following tensorization of Laplace transform.

**Theorem 40.3.** *[Tensorization of Laplace transform] Let $x_1, \cdots, x_n$ be independent random variables and $x'_1, \cdots, x'_n$ their indepent copies, $Z = Z(x_1, \cdots, x_n)$, $Z^i = Z(x_1, \cdots, x_{i-1}, x'_i, x_{i+1}, \cdots, x_n)$, $\phi(x) = e^x - x - 1$, and $\psi(x) = \phi(x) + e^x \phi(-x) = x \cdot (e^x - 1)$, and $I$ be the indicator function. Then*

$$\begin{aligned}
\mathbb{E}\left(e^{\lambda Z} \cdot \lambda Z\right) - \mathbb{E}e^{\lambda Z} \cdot \log \mathbb{E}e^{\lambda Z} \quad & \leq \quad \mathbb{E}_{x_1, \cdots, x_n, x'_1, \cdots, x'_n} e^{\lambda Z} \sum_{i=1}^{n} \phi\left(-\lambda(Z - Z^i)\right) \\
\mathbb{E}\left(e^{\lambda Z} \cdot \lambda Z\right) - \mathbb{E}e^{\lambda Z} \cdot \log \mathbb{E}e^{\lambda Z} \quad & \leq \quad \mathbb{E}_{x_1, \cdots, x_n, x'_1, \cdots, x'_n} e^{\lambda Z} \sum_{i=1}^{n} \psi\left(-\lambda(Z - Z^i)\right) \cdot I(Z \geq Z^i).
\end{aligned}$$

*Proof.* Let $u = \exp(\lambda Z)$ where $\lambda \in \mathbb{R}$, and apply the tensorization of entropy lemma,

$$
\begin{aligned}
& \underbrace{\mathbb{E}\left(e^{\lambda Z}\cdot\lambda Z\right) - \mathbb{E}e^{\lambda Z}\cdot\log\mathbb{E}e^{\lambda Z}}_{\mathrm{Ent}_{\mathbb{P}^n}\log u} \\
\leq \quad & \mathbb{E}\sum_{i=1}^n \mathrm{Ent}_{\mathbb{P}_i} e^{\lambda Z} \\
\underbrace{=}_{\text{variational formulation}} \quad & \mathbb{E}\sum_{i=1}^n \inf\left\{\int\left(e^{\lambda Z}(\lambda Z - \lambda x) - (e^{\lambda Z} - e^{\lambda x})\right)d\mathbb{P}_i : x \in \mathbb{R}^+\right\} \\
\leq \quad & \mathbb{E}\sum_{i=1}^n \mathbb{E}_{x_i x_i'}\left(e^{\lambda Z}(\lambda Z - \lambda Z^i) - (e^{\lambda Z} - e^{\lambda Z^i})\right) \\
= \quad & \mathbb{E}\sum_{i=1}^n \mathbb{E}_{x_i x_i'} e^{\lambda Z}\left(e^{-\lambda(Z-Z^i)} - \left(-\lambda\cdot\left(Z-Z^i\right)\right) - 1\right) \\
= \quad & \mathbb{E}_{x_1,\cdots,x_n,x_1',\cdots,x_n'} e^{\lambda Z}\sum_{i=1}^n \phi\left(-\lambda\cdot\left(Z-Z^i\right)\right).
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
& \mathbb{E}e^{\lambda Z}\sum_{i=1}^n \phi\left(-\lambda\cdot\left(Z-Z^i\right)\right) \\
= \quad & \mathbb{E}\sum_{i=1}^n e^{\lambda Z}\phi\left(-\lambda\cdot\left(Z-Z^i\right)\right)\cdot\left(\underbrace{I\left(Z\geq Z^i\right)}_{\mathbf{I}} + \underbrace{I\left(Z^i\geq Z\right)}_{\mathbf{II}}\right) \\
= \quad & \mathbb{E}\sum_{i=1}^n\left(\underbrace{e^{\lambda Z^i}\phi\left(-\lambda\cdot\left(Z^i-Z\right)\right)\cdot I\left(Z\geq Z^i\right)}_{\text{switch } Z \text{ and } Z^i \text{ in } \mathbf{II}} + \underbrace{e^{\lambda Z}\phi\left(-\lambda\cdot\left(Z-Z^i\right)\right)\cdot I\left(Z\geq Z^i\right)}_{\mathbf{I}}\right) \\
= \quad & \mathbb{E}\sum_{i=1}^n e^{\lambda Z}\cdot I\left(Z\geq Z^i\right)\cdot\left(\underbrace{e^{\lambda(Z^i-Z)}\cdot\phi\left(-\lambda\cdot\left(Z^i-Z\right)\right)}_{\mathbf{II}} + \underbrace{\phi\left(-\lambda\cdot\left(Z-Z^i\right)\right)}_{\mathbf{I}}\right) \\
= \quad & \mathbb{E}\sum_{i=1}^n e^{\lambda Z}\cdot I\left(Z\geq Z^i\right)\cdot\psi\left(-\lambda\cdot\left(Z^i-Z\right)\right).
\end{aligned}
$$

□

Recall the tensorization of entropy lemma we proved previously. Let $x_1, \cdots, x_n$ be independent random variables, $x_1', \cdots, x_n'$ be their independent copies, $Z = Z(x_1, \cdots, x_n)$, $Z^i = (x_1, \cdots, x_{i-1}, x_i', x_{i+1}, \cdots, x_n)$, and $\phi(x) = e^x - x - 1$. We have $\mathbb{E}e^{\lambda Z} - \mathbb{E}e^{\lambda Z} \log \mathbb{E}e^{\lambda Z} \le \mathbb{E}e^{\lambda Z} \sum_{i=1}^n \phi(-\lambda(Z - Z^i))$. We will use the tensorization of entropy technique to prove the following Hoeffding-type inequality. This theorem is Theorem 9 of `Pascal Massart. About the constants in Talagrand's concentration inequalities for empirical processes. The Annals of Probability, 2000, Vol 28, No. 2, 863-884.`

**Theorem 41.1.** *Let $\mathcal{F}$ be a finite set of functions $|\mathcal{F}| < \infty$. For any $f = (f_1, \cdots, f_n) \in \mathcal{F}$, $a_i \le f_i \le b_i$, $L = \sup_f \sum_{i=1}^n (b_i - a_i)^2$, and $Z = \sup_f \sum_{i=1}^n f_i$. Then $\mathbb{P}(Z \ge \mathbb{E}Z + \sqrt{2Lt}) \le e^{-t}$.*

*Proof.* Let

$$\begin{aligned} Z^i &= \sup_{f \in \mathcal{F}} \left( a_i + \sum_{j \ne i} f_j \right) \\ Z &= \sup_{f \in \mathcal{F}} \sum_{i=1}^n f_i \overset{\text{def.}}{=} \sum_{i=1}^n f_i^\circ. \end{aligned}$$

It follows that

$$0 \le Z - Z^i \le \sum_i f_i^\circ - \sum_{j \ne i} f_j^\circ - a_i = f_i^\circ - a_i \le b_i(f^\circ) - a_i(f^\circ).$$

Since $\frac{\phi(x)}{x^2} = \frac{e^x - x - 1}{x^2}$ is increasing in $\mathbb{R}$ and $\lim_{x \to 0} \frac{\phi(x)}{x^2} \to \frac{1}{2}$, it follows that $\forall x < 0$, $\phi(x) \le \frac{1}{2}x^2$, and

$$\begin{aligned} \mathbb{E}e^{\lambda Z}\lambda Z - \mathbb{E}e^{\lambda Z} \log \mathbb{E}e^{\lambda Z} &\le \mathbb{E}e^{\lambda Z} \sum_i \phi\left(-\lambda(Z - Z^i)\right) \\ &\le \frac{1}{2}\mathbb{E}e^{\lambda Z} \sum_i \lambda^2 (Z - Z^i)^2 \\ &\le \frac{1}{2} L\lambda^2 \mathbb{E}e^{\lambda Z}. \end{aligned}$$

Center $Z$, and we get

$$\mathbb{E}e^{\lambda(Z - \mathbb{E}Z)}\lambda(Z - \mathbb{E}Z) - \log \mathbb{E}e^{\lambda(Z - \mathbb{E}Z)} \le \frac{1}{2} L\lambda^2 \mathbb{E}e^{\lambda(Z - \mathbb{E}Z)}.$$

Let $F(\lambda) = \mathbb{E}e^{\lambda(Z-\mathbb{E}Z)}$. It follows that $F'_\lambda(\lambda) = \mathbb{E}e^{\lambda(Z-\mathbb{E}Z)}(Z - \mathbb{E}Z)$ and

$$\begin{aligned}
\lambda F'_\lambda(\lambda) - F(\lambda)\log F(\lambda) &\leq \frac{1}{2}L\lambda^2 F(\lambda)\\
\frac{1}{\lambda}\frac{F'_\lambda(\lambda)}{F(\lambda)} - \frac{1}{\lambda^2}\log F(\lambda) &\leq \frac{1}{2}L\\
\left(\frac{1}{\lambda}\log F(\lambda)\right)'_\lambda &\leq \frac{1}{2}L\\
\frac{1}{\lambda}\log F(\lambda) &= \frac{1}{t}\log F(t)\big|_{t\to 0} + \int_0^\lambda \left(\frac{1}{t}\log F(t)\right)'_t dt\\
&\leq \frac{1}{2}L\lambda\\
F(\lambda) &\leq \exp(\frac{1}{2}L\lambda^2).
\end{aligned}$$

By Chebychev inequality, and minimize over $\lambda$, we get

$$\begin{aligned}
\mathbb{P}(Z \geq \mathbb{E}Z + t) &\leq e^{-\lambda t}\mathbb{E}e^{\lambda(Z-\mathbb{E}Z)}\\
&\leq e^{-\lambda t}e^{\frac{1}{2}L\lambda^2}\\
\overset{\text{minimize over }\lambda}{\mathbb{P}(Z \geq \mathbb{E}Z + t)} &\leq e^{-t^2/(2L)}
\end{aligned}$$

□

Let $f_i$ above be Rademacher random variables and apply Hoeffding's inequality, we get $\mathbb{P}(Z \geq \mathbb{E}Z + \sqrt{Lt/2}) \leq e^{-t}$. As a result, The above inequality improves the constant of Hoeffding's inequality.

The following Bennett type concentration inequality is Theorem 10 of

`Pascal Massart. About the constants in Talagrand's concentration inequalities for empirical processes. The Annals of Probability, 2000, Vol 28, No. 2, 863-884.`

**Theorem 41.2.** *Let $\mathcal{F}$ be a finite set of functions $|\mathcal{F}| < \infty$. $\forall f = (f_1, \cdots, f_n) \in \mathcal{F}$, $0 \leq f_i \leq 1$, $Z = \sup_f \sum_{i=1}^n f_i$, and define $h$ as $h(u) = (1+u)\log(1+u) - u$ where $u \geq 0$. Then $\mathbb{P}(Z \geq \mathbb{E}Z + x) \leq e^{-\mathbb{E}Z\cdot h(x/\mathbb{E}Z)}$.*

*Proof.* Let

$$\begin{aligned}
Z &= \sup_{f\in\mathcal{F}}\sum_{i=1}^n f_i \overset{\text{def.}}{=} \sum_{i=1}^n f_i^\circ\\
Z^i &= \sup_{f\in\mathcal{F}}\sum_{j\neq i} f_j.
\end{aligned}$$

It follows that $0 \leq Z - Z^i \leq f_i^\circ \leq 1$. Since $\phi = e^x - x - 1$ is a convex function of $x$,

$$\phi(-\lambda(Z - Z^i)) = \phi(-\lambda \cdot (Z - Z^i) + 0 \cdot (1 - (Z - Z^i))) \leq (Z - Z^i)\phi(-\lambda)$$

and

$$
\begin{aligned}
\mathbb{E}\left(\lambda Z e^{\lambda Z}\right) - \mathbb{E}e^{\lambda Z}\log \mathbb{E}e^{\lambda Z} &\leq \mathbb{E}\left(e^{\lambda Z}\sum_{i=1}^{n}\phi\left(-\lambda(Z-Z^i)\right)\right) \\
&\leq \mathbb{E}\left(e^{\lambda Z}\phi(-\lambda)\sum_i\left(Z-Z^i\right)\right) \\
&\leq \phi(-\lambda)\mathbb{E}\left(e^{\lambda Z}\cdot\sum_i f_i^\circ\right) \\
&= \phi(-\lambda)\mathbb{E}\left(Z\cdot e^{\lambda Z}\right).
\end{aligned}
$$

Set $\tilde{Z} = Z - \mathbb{E}Z$ (i.e., center $Z$), and we get

$$
\begin{aligned}
\mathbb{E}\left(\lambda\tilde{Z}e^{\lambda\tilde{Z}}\right) - \mathbb{E}e^{\lambda\tilde{Z}}\log\mathbb{E}e^{\lambda\tilde{Z}} &\leq \phi(-\lambda)\mathbb{E}\left(\tilde{Z}\cdot e^{\lambda\tilde{Z}}\right) \leq \phi(-\lambda)\mathbb{E}\left(\left(\tilde{Z}+\mathbb{E}Z\right)\cdot e^{\lambda\tilde{Z}}\right) \\
\left(\lambda-\phi(-\lambda)\right)\mathbb{E}\left(\tilde{Z}e^{\lambda\tilde{Z}}\right) - \mathbb{E}e^{\lambda\tilde{Z}}\log\mathbb{E}e^{\lambda\tilde{Z}} &\leq \phi(-\lambda)\cdot\mathbb{E}Z\cdot\mathbb{E}e^{\lambda\tilde{Z}}.
\end{aligned}
$$

Let $v = \mathbb{E}Z$, $F(\lambda) = \mathbb{E}e^{\lambda\tilde{Z}}$, $\Psi = \log F$, and we get

$$
\begin{aligned}
\left(\lambda-\phi(-\lambda)\right)\frac{F'_\lambda(\lambda)}{F(\lambda)} - \log F(\lambda) &\leq v\phi(-\lambda) \\
\left(\lambda-\phi(-\lambda)\right)\left(\log F(\lambda)\right)'_\lambda - \log F(\lambda) &\leq v\phi(-\lambda). \qquad (41.1)
\end{aligned}
$$

Solving the differential equation

$$
\left(\lambda-\phi(-\lambda)\right)\left(\underbrace{\log F(\lambda)}_{\Psi_0}\right)'_\lambda - \underbrace{\log F(\lambda)}_{\Psi_0} = v\phi(-\lambda), \qquad (41.2)
$$

yields $\Psi_0 = v\cdot\phi(\lambda)$. We will proceed to show that $\Psi$ satisfying 41.1 has the property $\Psi \leq \Psi_0$:

$$
\begin{aligned}
&\scriptstyle \text{Substract 41.2 from 41.1, and let } \Psi_1=\Psi-\Psi_0 \\
(1-e^{-\lambda})\Psi'_1 - \Psi_1 &\leq 0 \\
&\scriptstyle \left(e^\lambda-1\right)\left(1-e^{-\lambda}\right)\frac{1}{e^\lambda-1}=1-e^{-\lambda}\text{ ,and }\left(e^\lambda-1\right)\left(1-e^{-\lambda}\right)\frac{e^\lambda}{\left(e^\lambda-1\right)^2}=1 \\
\left(e^\lambda-1\right)\left(1-e^{-\lambda}\right)\underbrace{\left(\frac{1}{e^\lambda-1}\Psi'_1 - \frac{e^\lambda}{\left(e^\lambda-1\right)^2}\Psi_1\right)}_{\left(\frac{\Psi_1}{e^\lambda-1}\right)'_\lambda} &\leq 0 \\
\frac{\Psi_1(\lambda)}{e^\lambda-1} &\leq \lim_{\lambda\to 0}\frac{\Psi_1(\lambda)}{e^\lambda-1} = 0.
\end{aligned}
$$

It follows that $\Psi \leq v\phi(\lambda)$, and $F = \mathbb{E}e^{\lambda Z} \leq e^{v\phi(\lambda)}$. By Chebychev's inequality, $\mathbb{P}\left(Z \geq \mathbb{E}Z + t\right) \leq e^{-\lambda t + v\phi(\lambda)}$. Minimizing over all $\lambda > 0$, we get $\mathbb{P}\left(Z \geq \mathbb{E}Z + t\right) \leq e^{-v\cdot h(t/v)}$ where $h(x) = (1+x)\cdot\log(1+x) - x$. □

The following sub-additive increments bound can be found as Theorem 2.5 in

`Olivier Bousquet. Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms. PhD thesis, Ecole Polytechnique, 2002.`

**Theorem 41.3.** *Let* $Z = \sup_{f\in\mathcal{F}} \sum f_i$, $\mathbb{E}f_i = 0$, $\sup_{f\in\mathcal{F}} var(f) = \sup_{f\in\mathcal{F}} \sum_{i=1}^n f_i^2 \overset{def.}{=} \sigma^2$, $\forall i \in \{1,\cdots,n\}, f_i \leq u \leq 1$. *Then* $\mathbb{P}\left(Z \geq \mathbb{E}Z + \sqrt{(1+u)\mathbb{E}Z + n\sigma^2 x} + \frac{x}{3}\right) \leq e^{-x}$.

*Proof.* Let

$$\begin{aligned}
Z &= \sup_{f\in\mathcal{F}} \sum_{i=1}^n f_i \overset{\text{def.}}{=} \sum_{i=1}^n f_i^\circ \\
Z_k &= \sup_{f\in\mathcal{F}} \sum_{i\neq k} f_i \\
Z_k' &= f_k \text{ such that } Z_k = \sup_{f\in\mathcal{F}} \sum_{i\neq k} f_i.
\end{aligned}$$

It follows that $Z_k' \leq Z - Z_k \leq u$. Let $\psi(x) = e^{-x} + x - 1$. Then

$$\begin{aligned}
e^{\lambda Z}\psi(\lambda(Z - Z_k)) &= e^{\lambda Z_k} - e^{\lambda Z} + \lambda(Z - Z_k)e^{\lambda Z} \\
&= f(\lambda)\left(Z - Z_k\right)e^{\lambda Z} + \left(\lambda - f(\lambda)\right)\left(Z - Z_k\right)e^{\lambda Z} + e^{\lambda Z_k} - e^{\lambda Z} \\
&= f(\lambda)\left(Z - Z_k\right)e^{\lambda Z} + g(Z - Z_k)e^{\lambda Z_k}.
\end{aligned}$$

In the above, $g(x) = 1 - e^{\lambda x} + \left(\lambda - f(\lambda)\right)xe^{\lambda x}$, and we define $f(\lambda) = \left(1 - e^{\lambda} + \lambda e^{\lambda}\right)/\left(e^{\lambda} + \alpha - 1\right)$ where $\alpha = 1/\left(1+u\right)$. We will need the following lemma to make use of the bound on the variance.

**Lemma 41.4.** *For all* $x \leq 1$, $\lambda \geq 0$ *and* $\alpha \geq \frac{1}{2}$, $g(x) \leq f(x)\left(\alpha x^2 - x\right)$.

Continuing the proof, we have

$$\begin{aligned}
e^{\lambda Z}\psi(\lambda(Z - Z_k)) &= f(\lambda)\left(Z - Z_k\right)e^{\lambda Z} + g(Z - Z_k)e^{\lambda Z_k} \\
&\leq f(\lambda)\left(Z - Z_k\right)e^{\lambda Z} + e^{\lambda Z_k}f(\lambda)\left(\alpha\left(Z - Z_k\right)^2 - \left(Z - Z_k\right)\right) \\
&\leq f(\lambda)\left(Z - Z_k\right)e^{\lambda Z} + e^{\lambda Z_k}f(\lambda)\left(\alpha\left(Z_k'\right)^2 - Z_k'\right).
\end{aligned}$$

Sum over all $k = 1,\cdots,n$, and take expectation, we get

$$\begin{aligned}
e^{\lambda Z}\sum_k \psi(\lambda(Z - Z_k)) &\leq f(\lambda)Ze^{\lambda Z} + f(\lambda)\sum_k e^{\lambda Z_k}\left(\alpha\left(Z_k'\right)^2 - Z_k'\right) \\
\mathbb{E}e^{\lambda Z}\sum_k \psi(\lambda(Z - Z_k)) &\leq f(\lambda)\mathbb{E}Ze^{\lambda Z} + f(\lambda)\sum_k \mathbb{E}e^{\lambda Z_k}\left(\alpha\left(Z_k'\right)^2 - Z_k'\right).
\end{aligned}$$

Since $\mathbb{E}Z'_k = 0$, $\mathbb{E}_{X_k}\left(Z'_k\right)^2 = \mathbb{E}f_k^2 = \operatorname{var}(f_k) \le \sup_{f\in\mathcal{F}}\operatorname{var}(f) \le \sigma^2$, it follows that

$$\begin{aligned}
\mathbb{E}e^{\lambda Z_k}\left(\alpha\left(Z'_k\right)^2 - Z'_k\right) &= \mathbb{E}e^{\lambda Z_k}\left(\alpha\mathbb{E}_{f_k}\left(Z'_k\right)^2 - \mathbb{E}_{f_k}Z'_k\right) \\
&\le \alpha\sigma^2\mathbb{E}e^{\lambda Z_k} \\
&\le \alpha\sigma^2\mathbb{E}e^{\lambda Z_k+\lambda\mathbb{E}Z'_k} \\
&\overset{\text{Jensen's inequality}}{\le} \alpha\sigma^2\mathbb{E}e^{\lambda Z_k+\lambda Z'_k} \\
&\le \alpha\sigma^2\mathbb{E}e^{\lambda Z}.
\end{aligned}$$

Thus

$$\begin{aligned}
\mathbb{E}\left(\lambda Ze^{\lambda Z}\right) - \mathbb{E}e^{\lambda Z}\log\mathbb{E}e^{\lambda Z} &\le \mathbb{E}e^{\lambda Z}\sum_k \psi(\lambda(Z-Z_k)) \\
&\le f(\lambda)\mathbb{E}Ze^{\lambda Z} + f(\lambda)\alpha n\sigma^2\mathbb{E}e^{\lambda Z}.
\end{aligned}$$

Let $Z_0 = Z - \mathbb{E}$, and center $Z$, we get

$$\mathbb{E}\left(\lambda Z_0e^{\lambda Z_0}\right) - \mathbb{E}e^{\lambda Z_0}\log\mathbb{E}e^{\lambda Z_0} \le f(\lambda)\mathbb{E}Z_0e^{\lambda Z_0} + f(\lambda)\left(\alpha n\sigma^2 + \mathbb{E}Z\right)\mathbb{E}e^{\lambda Z_0}.$$

Let $F(\lambda) = \mathbb{E}e^{\lambda Z_0}$, and $\Psi(\lambda) = \log F(\lambda)$, we get

$$\begin{aligned}
(\lambda - f(\lambda))\,F'(\lambda) - F(\lambda)\log F(\lambda) &\le f(\lambda)\left(\alpha n\sigma^2 + \mathbb{E}Z\right)F(\lambda) \\
(\lambda - f(\lambda))\underbrace{\frac{F'(\lambda)}{F(\lambda)}}_{\Psi'(\lambda)} - \underbrace{\log F(\lambda)}_{\Psi(\lambda)} &\le f(\lambda)\left(\alpha n\sigma^2 + \mathbb{E}Z\right).
\end{aligned}$$

Solve this inequality, we get $F(\lambda) \le e^{v\psi(-\lambda)}$ where $v = n\sigma^2 + (1+u)\mathbb{E}Z$. □

This lecture reviews the method for proving concentration inequalities developed by Sourav Chatterjee based on Stein's method of exchangeable pairs. The lecture is based on

`Sourav Chatterjee. Stein's method for concentration inequalities. Probab. Theory Relat. Fields (2007) 138:305-321. DOI 10.1007/s00440-006-0029-y.`

**Theorem 42.1.** *Let $(X, X')$ be an exchangeable pair on $\mathcal{X}$ (i.e., $d\mathbb{P}(X, X') = d\mathbb{P}(X', X)$. Let $F(x, x') = -F(x, x')$ be antisymmetric, $f(x) = \mathbb{E}(F(x, x')|x)$. Then $\mathbb{E}f(x) = 0$. If further*

$$\Delta(x) = \frac{1}{2}\mathbb{E}\left(\left|(f(x) - f(x'))\, F(x, x')\right| \middle| x\right) \leq Bf(x) + C,$$

*then $\mathbb{P}(f(x) \geq t) \leq \exp\left(-\frac{t^2}{2(C+Bt)}\right)$.*

*Proof.*

$$\begin{aligned}
\mathbb{E}\left(h(X)f(X)\right) &\overset{\text{by definition of } f(X)}{=} \mathbb{E}\left(h(X)\cdot\mathbb{E}\left(F(X, X')|X\right)\right) = \mathbb{E}\left(h(X)\cdot F(X, X')\right) \\
&\overset{X, X' \text{ are exchangeable}}{=} \mathbb{E}\left(h(X')\cdot F(X', X)\right) \\
&\overset{F(X, X') \text{ is anti-symmetric}}{=} -\mathbb{E}\left(h(X')\cdot F(X, X')\right) \\
&= \frac{1}{2}\mathbb{E}\left((h(X) - h(X'))\cdot F(X, X')\right).
\end{aligned}$$

Take $h(X)$=1, we get $\mathbb{E}f(x) = 0$. Take $h(X) = f(X)$, we get $\mathbb{E}f^2 = \frac{1}{2}\mathbb{E}\left((f(X) - f(X'))\cdot F(X, X')\right)$.

We proceed to bound the moment generating function $m(\lambda) = \mathbb{E}\exp(\lambda f(X))$, and use Chebychev's inequality to complete the proof. The derivative of $m(\lambda)$ satisfies,

$$\begin{aligned}
|m'_\lambda(\lambda)| &= \left|\mathbb{E}\left(\underbrace{e^{\lambda f(X)}}_{h(X)}\cdot f(X)\right)\right| \\
&= \left|\frac{1}{2}\mathbb{E}\left(\left(e^{\lambda f(X)} - e^{\lambda f(X')}\right)\cdot F(X, X')\right)\right| \\
&\leq \mathbb{E}\left|\frac{1}{2}\left(\left(e^{\lambda f(X)} - e^{\lambda f(X')}\right)\cdot F(X, X')\right)\right| \\
&\overset{\frac{e^a - e^b}{a-b} = \int_0^1 e^{b+t(a-b)}dt \leq \int_0^1 (te^a + (1-t)e^b)dt = \frac{1}{2}(e^a + e^b)}{\leq} |\lambda|\cdot\mathbb{E}\left(\frac{1}{2}\left(e^{\lambda f(X)} + e^{\lambda f(X')}\right)\cdot\left|\frac{1}{2}\left(f(X) - f(X')\right)\cdot F(X, X')\right|\right) \\
&= |\lambda|\cdot\mathbb{E}\left(e^{\lambda f(X)}\cdot\left|\frac{1}{2}\left(f(X) - f(X')\right)\cdot F(X, X')\right|\right) \\
&= |\lambda|\cdot\mathbb{E}\left(e^{\lambda f(X)}\cdot\underbrace{\mathbb{E}\left(\left|\frac{1}{2}\left(f(X) - f(X')\right)\cdot F(X, X')\right|\middle| X\right)}_{\Delta(X)}\right) \\
&\leq |\lambda|\cdot\mathbb{E}\left(e^{\lambda f(X)}\cdot(B\cdot f(X) + C)\right) = |\lambda|\cdot(B\cdot m'_\lambda(\lambda) + C\cdot m(\lambda))
\end{aligned}$$

Since $m(\lambda)$ is a convex function in $\lambda$, and $m'(0)=0$, $m'(\lambda)$ always has the same sign as $\lambda$. In the interval $0 \leq \lambda < 1/B$, the above inequality can be expressed as

$$\begin{aligned}
m'(\lambda) &\leq \lambda \cdot (B \cdot m'(\lambda) + C \cdot m(\lambda)) \\
(\log m(\lambda))'_\lambda &\leq \frac{\lambda \cdot C}{(1-\lambda B)} \\
\log m(\lambda) &\leq \int_0^\lambda \frac{s \cdot C}{1 - s \cdot B} ds \;\leq\; \frac{C}{1-\lambda \cdot B} \int_0^\lambda s ds \;=\; \frac{1}{2} \cdot \frac{C \cdot \lambda^2}{1 - \lambda \cdot B}.
\end{aligned}$$

By Chebyshev's inequality $\mathbb{P}(f(x) \geq t) \leq \exp\left(-\lambda t + \frac{1}{2} \cdot \frac{C \cdot \lambda^2}{1-\lambda \cdot B}\right)$. Minimize the inequality over $0 \leq \lambda < \frac{1}{B}$, we get $\lambda = \frac{t}{C+Bt}$, and $\mathbb{P}(f(x) \geq t) \leq \exp\left(-\frac{t^2}{2 \cdot (C+Bt)}\right)$. □

We will use the following three examples to illustrate how to use the above theorem to prove concentration inequalities.

**Example 42.2.** Let $X_1, \cdots, X_n$ be i.i.d random variables with $\mathbb{E}x_i = \mu_i$, $\text{var}(x_i) = \sigma_i^2$, and $|x_i - \mu_i| \leq c_i$. Let $X = \sum_{i=1}^n X_i$, our goal is to bound $|X - \mathbb{E}X|$ probabilistically. To apply the above theorem, take $X_i'$ be an independent copy of $X_i$ for $i = 1, \cdots, n$, $I \sim \text{unif}\{1, \cdots, n\}$ be a random variable uniformly distributed over $1, \cdots, n$, and $X' = \sum_{i \neq I} X_i + X_I$. Define $F(X, X')$, $f(X)$, and $\Delta(X)$ as the following,

$$\begin{aligned}
F(X, X') &\overset{\text{def.}}{=} n \cdot (X - X') = n \cdot (X_I - X_I') \\
f(X) &\overset{\text{def.}}{=} \mathbb{E}\left(F(X, X')|X\right) = \mathbb{E}\left(n \cdot (X_I - X_I')|X\right) = \frac{1}{n} \sum_{I=1}^n \mathbb{E}\left(n \cdot (X_I - X_I')|X\right) = X - \mathbb{E}X \\
\Delta(X) &\overset{\text{def.}}{=} \frac{1}{2} \mathbb{E}\left(\left|(f(X) - f(X')) \cdot F(X, X')\right| \big| X\right) \\
&= \frac{n}{2} \cdot \frac{1}{n} \sum_{I=1}^n \mathbb{E}\left((X_I - X_I')^2 \big| X\right) \\
&= \frac{1}{2} \sum_{I=1}^n \left( \mathbb{E}\left( \underbrace{(X_I - \mathbb{E}X_I)^2}_{\leq c_i^2} \big| X \right) + \underbrace{\mathbb{E}\left((X_I' - \mathbb{E}X_I')^2\right)}_{=\sigma_i^2} \right) \\
&\leq \frac{1}{2} \sum_{I=1}^n \left(c_i^2 + \sigma_i^2\right).
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbb{P}\left(\sum X_i - \mathbb{E}\sum X_i \geq t\right) &\leq \exp\left(-\frac{t^2}{\sum_i c_i^2 + \sigma_i^2}\right) \\
\mathbb{P}\left(\left(-\sum X_i\right) - \mathbb{E}\left(-\sum X_i\right) \geq t\right) &\leq \exp\left(-\frac{t^2}{\sum_i c_i^2 + \sigma_i^2}\right) \\
\overset{\text{union bound}}{\mathbb{P}\left(\left|\sum X_i - \mathbb{E}\sum X_i\right| \geq t\right)} &\leq 2\exp\left(-\frac{t^2}{\sum_i c_i^2 + \sigma_i^2}\right).
\end{aligned}$$

**Example 42.3.** Let $(a_{ij})_{i,j=1,\cdots,n}$ be a real matrix where $a_{ij} \in [0,1]$ for $1 \leq i,j \leq n$, $\pi$ be a random variable uniformly distributed over the permutations of $1, \cdots, n$. Let $X = \sum_{i=1}^n a_{i,\pi(i)}$, then $\mathbb{E}X = \frac{1}{n}\sum_{i,j} a_{i,j}$, and our goal is to bound $|X - \mathbb{E}X|$ probabilistically. To apply the above theorem, we define exchangeable pairs of permutations in the following way. Given permutation $\pi$, pick $I, J$ uniformly and independently from $\{1, \cdots, n\}$, and construct $\pi' = \pi \circ (I,J)$ where $(I,J)$ is a transposition of $I$ and $J$. The two random variables $\pi$ and $\pi'$ are exchangeable. We can define $F(X,X')$, $f(X)$, and $\Delta(X)$ as the following,

$$\begin{aligned}
F(X,X') &\overset{\text{def.}}{=} \frac{n}{2}(X - X') = \frac{n}{2}\left(\sum_{i=1}^n a_{i,\pi(i)} - \sum_{i=1}^n a_{i,\pi'(i)}\right) \\
f(X) &\overset{\text{def.}}{=} \mathbb{E}\left(F(X,X')|X\right) \\
&= \frac{n}{2}\left(a_{I,\pi(I)} + a_{J,\pi(J)} - a_{I,\pi(J)} - a_{J,\pi(I)}|\pi\right) \\
&= \frac{n}{2}\cdot\frac{1}{n}\sum_I a_{I,\pi(I)} + \frac{n}{2}\cdot\frac{1}{n}\sum_J a_{J,\pi(J)} - \frac{n}{2}\cdot\frac{1}{n^2}\sum_{I,J} a_{I,J} - \frac{n}{2}\cdot\frac{1}{n^2}\sum_{I,J} a_{I,J} \\
&= \sum_i a_{i,\pi(i)} - \frac{1}{n}\sum_{i,j} a_{i,j} \\
&= X - \mathbb{E}X \\
\Delta(X) &\overset{\text{def.}}{=} \frac{1}{2}\cdot\frac{n}{2}\mathbb{E}\left((X-X')^2|\pi\right) \\
&= \frac{n}{4}\mathbb{E}\left(\left(a_{I,\pi(I)} + a_{J,\pi(J)} - a_{I,\pi(J)} - a_{J,\pi(I)}\right)^2|\pi\right) \\
&\leq \frac{n}{2}\mathbb{E}\left(a_{I,\pi(I)} + a_{J,\pi(J)} - a_{I,\pi(J)} - a_{J,\pi(I)}|\pi\right) \\
&= X + \mathbb{E}X \\
&= f(X) + 2\mathbb{E}X.
\end{aligned}$$

Apply the theorem above, and take union bound, we get $\mathbb{P}\left(|X - \mathbb{E}X| \geq t\right) \leq 2\exp\left(-\frac{t^2}{4\mathbb{E}X + t}\right)$.

**Example 42.4.** In this example, we consider a concentration behavior of the Curie-Weiss model. Let $\sigma = (\sigma_1, \cdots, \sigma_n) \in \{-1,1\}^n$ be random variables observing the probability distribution

$$G\left((\sigma_1,\cdots,\sigma_n)\right) = \frac{1}{Z}\exp\left(\frac{\beta}{n}\sum_{i<j}\sigma_i\sigma_j + \beta\cdot h\sum_{i=1}^n \sigma_i\right).$$

We are interested in the concentration of $m(\sigma) = \frac{1}{n}\sum_i \sigma_i$ around $\tanh\left(\beta\cdot m(\sigma) + \beta\cdot h\right)$ where $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. Given any $\sigma$, we can pick $I$ uniformly and independently from $\{1,\cdots,n\}$, and generate $\sigma_I'$ according to the conditional distribution of $\sigma_i$ on $\{\sigma_j : j \neq i\}$ (Gibbs sampling):

$$\begin{aligned}
\mathbb{P}(\sigma'_i = +1 | \{\sigma_j : j \neq i\}) &= \frac{\exp(\frac{\beta}{n}\sum_{j\neq i}\sigma_j + \beta \cdot h)}{2 \cdot (\exp(\frac{\beta}{n}\sum_{j\neq i}\sigma_j + \beta \cdot h) + \exp(-\frac{\beta}{n}\sum_{j\neq i}\sigma_j - \beta \cdot h))} \\
\mathbb{P}(\sigma'_i = -1 | \{\sigma_j : j \neq i\}) &= \frac{\exp(-\frac{\beta}{n}\sum_{j\neq i}\sigma_j - \beta \cdot h)}{2 \cdot (\exp(\frac{\beta}{n}\sum_{j\neq i}\sigma_j + \beta \cdot h) + \exp(-\frac{\beta}{n}\sum_{j\neq i}\sigma_j - \beta \cdot h))}.
\end{aligned}$$

Let $\sigma'_j = \sigma_j$ for $j \neq I$. The two random variables $\sigma$ and $\sigma'$ are exchangeable pairs. To apply the above theorem, we define $F(X, X')$, $f(X)$, and $\Delta(X)$ as the following,

$$\begin{aligned}
F(\sigma, \sigma') &\overset{\text{def.}}{=} \sum \sigma_i - \sum \sigma'_i = \sigma_I - \sigma'_I \\
f(\sigma) &\overset{\text{def.}}{=} \mathbb{E}\left(F(\sigma, \sigma')|\sigma\right) \\
&= \mathbb{E}\left(\sigma_I - \sigma'_I|\sigma\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(\sigma_i - \sigma'_i|\sigma\right) \text{ , where } \sigma'_1, \cdots, \sigma'_n \text{ are all by Gibbs sampling.} \\
&= \frac{1}{n}\sum_{i=1}^{n}\sigma_i - \frac{1}{n}\sum_{i=1}^{n}\tanh\left(\frac{\beta}{n}\sum_{j\neq i}\sigma_j + \beta h\right) \\
\Delta(X) &\overset{\text{def.}}{=} \frac{1}{2}\mathbb{E}\left(|(f(X) - f(X')) \cdot F(X, X')|\,\big| X\right) \\
&\overset{|F(\sigma,\sigma')|\leq 2, |f(\sigma)-f(\sigma')|\leq 2(1+\beta)/n}{\leq} \frac{1}{2} \cdot 2 \cdot \frac{2(1+\beta)}{n}.
\end{aligned}$$

Thus $\mathbb{P}\left(\left|\frac{1}{n}\sum_i \sigma_i - \frac{1}{n}\sum_i \tanh\left(\frac{\beta}{n}\sum_{j\neq i}\sigma_i + \beta h\right)\right| \geq t\right) \leq 2\exp\left(-\frac{t^2 n}{4(1+\beta)}\right)$. Since $|\tanh(\beta m_i(\sigma) + \beta h) - \tanh(\beta m(\sigma) + \beta h)| \leq \frac{\beta}{n}$ where $m_i(\sigma) = \frac{1}{n}\sum_{j\neq i}\sigma_i$, we have $\mathbb{P}\left(\left|\frac{1}{n}\sum_i \sigma_i - \tanh\left(\beta \cdot m(\sigma) + \beta h\right)\right| \geq \frac{\beta}{n} + \frac{t}{\sqrt{n}}\right) \leq 2\exp\left(-\frac{t^2 n}{4(1+\beta)}\right)$.