Notes on ML theory and high-dimensional statistics
Young Ahn (@chiyahn)

---

1 Basic concentration bounds

1.1 Sub-Gaussian variables

Proposition 1.1 (Chernoff's bound)**.**

$$\mathbf{P}(Z - \mathbf{E}Z > t) \leq \frac{\mathbf{E}e^{\lambda(Z-\mathbf{E}Z)}}{e^{\lambda t}} \tag{1}$$

The following lemma is useful for providing a sharp bound in moment generating functions:

Lemma 1.1 (Hoeffding's lemma)**.** Let X be a random variable whose mean is zero whose support is on $[a, b]$. Then, for every $\lambda \in \mathbb{R}$, we have

$$\mathbf{E}e^{\lambda X} \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right) \tag{2}$$

Definition 1.1 (Sub-Gaussian random variable)**.** A random variable X is sub-Gaussian if there exists $\sigma^2 > 0$ such that

$$\mathbf{E}e^{\lambda(X-\mu)} \leq e^{\sigma^2\lambda/2} \tag{3}$$

By the Chernoff's bound, and optimizing over λ, we have the following concentration inequality:

Lemma 1.2 (Concentration inequality for sub-Gaussian random variables)**.** X satisfies the following ineqaulity if X is sub-Gaussian:

$$\mathbf{P}(|X - \mu| > t) \leq 2e^{-\frac{t^2}{2\sigma^2}} \tag{4}$$

which results in the Hoeffding bound:

Proposition 1.2 (Hoeffding bound)**.** Let X_i be independent random variables such that each X_i has mean M_i and sub-Gaussian parameter σ_i. Then we have

$$\mathbf{P}\left[\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right] \leq \exp\left[-\frac{t^2}{2\sum_{i=1}^{N}\sigma_i^2}\right] \tag{5}$$

Proposition 1.3 (Maximal inequality)**.** Let $Y_1, ..., Y_N$ be sub-Gaussian with parameter σ. Then we ahve

$$\mathbf{E} \max_{i=1,...,N} Y_i \leq \sigma\sqrt{2\log N} \tag{6}$$

Examples 1.1 (Maximal bound on point estimator convergence)**.** Let
$A_1, ..., A_M \in \mathcal{X}$ and let $X_1, ..., X_N$ be random variables that takes value on one of $\mathcal{X}$. Let $\mathbf{P}(A_j) = \mathbf{P}(X_1 \in A_j)$ and $\mathbf{P}_N(A_j) = \frac{1}{N}\sum_{i=1}^N 1(X_i \in A_j)$. First, by the Hoeffding's lemma, $\mathbf{P}_N(A_j)$ is sub-Gaussian for all $j = 1, ..., M$

$$\mathbf{E}e^{\lambda(\mathbf{P}(A_j)-\mathbf{P}_N(A_j))} = \prod_{i=1}^n \mathbf{E}e^{(\lambda/N)(\mathbf{P}(A_j)-1(X_i\in A_j))} \tag{7}$$

$$\leq \exp(\lambda^2/(8N)) \tag{8}$$

so by maximal inequality, we have

$$\mathbf{E} \max_{j=1,...,M} (\mathbf{P}(A_j) - \mathbf{P}_n(A_j)) \leq \sqrt{\frac{\log M}{2N}} \tag{9}$$

Proposition 1.4 (Rademacher variables are sub-Gaussian with $\sigma = 1$)**.** Let ϵ be a Rademacher random variable. Then we have

$$\mathbf{E}e^{\lambda\epsilon} = \frac{1}{2}\left(e^{-\lambda} + e^{\lambda}\right) \tag{10}$$

$$= \frac{1}{2}\left(\sum_{k=0}^{\infty}\frac{(-\lambda)^k}{k!} + \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\right) \tag{11}$$

$$= \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{(2k)!} \tag{12}$$

$$\leq \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{2^k k!} = e^{\lambda^2/2} \tag{13}$$

Examples 1.2 (Bounded random variable by a symmetrization argument)**.** We show that if X is bounded then X is sub-Gaussian (which is also shown by the Hoeffding lemma above). WLOG, let X be mean-zero supported on interval $[a, b]$. Let X' be an independent copy of X and let σ be a Rademacher random variable independent of X and X'. Then we have

$$\mathbf{E}_X e^{\lambda X} = \mathbf{E}_X e^{\lambda[X - \mathbf{E}_{X'}X']} \tag{14}$$

$$= \mathbf{E}_{X,X'} e^{\lambda[X-X']} \tag{15}$$

$$= \mathbf{E}_{X,X'}\mathbf{E}_\sigma e^{\lambda\sigma[X-X']} \tag{16}$$

$$\leq \mathbf{E}_{X,X'} e^{\lambda^2[X-X']/2} \tag{17}$$

$$\leq e^{\lambda^2(b-a)^2/2} \tag{18}$$

Some random variables are not sub-Gaussian.

Definition 1.2 (Sub-exponential random variable)**.** A random variable X is sub-Gaussian if there exists $\sigma^2 > 0$ such that

$$\mathbf{E}e^{\lambda(X-\mu)} \leq e^{\sigma^2\lambda/2} \tag{19}$$

for all $|\lambda| < \alpha^{-1}$ for some $\alpha > 0$.

There are some random variables that are sub-exponential but not sub-Gaussian.

Examples 1.3 (Sub-exponential but not sub-Gaussian)**.** Consider $X \sim \chi^2(1)$.

$$\mathbf{E}e^{\lambda(X-1)} = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \tag{20}$$

which is not defined when $\lambda > 1/2$.

The following Bernstein bound is useful as it uses an extra infromation about the variance of random variables to control the bound, unlike the Hoeffding that only uses boundedness:

Proposition 1.5 (Bernstein bound)**.** Let $X_1, ..., X_n$ be independent with $X_i \leq 1$. Let $v = \sum_{i=1}^N \mathbf{E}X_i^2$. Then we have

$$\mathbf{P}\left(\sum_{i=1}^N (X_i - \mathbf{E}X_i) \geq t\right) \leq \exp\left(-\frac{t^2}{2(v+t/3)}\right) \tag{21}$$

1.2 Martingale methods

Definition 1.3 (Doob martingale)**.** For $\{X_k\}_{k=1}^n$ random variables, consider $Z = f(X_1, ..., X_n)$. Then writing $\Delta_k = \mathbf{E}_k Z - \mathbf{E}_{k-1} Z$, we can represent the difference between Z and its mean as

$$Z - \mathbf{E}Z = \sum_{k=1}^n \Delta_k \tag{22}$$

which is called the Doob martingale representation of Z.

Proposition 1.6 (Efron-Stein inequality)**.** Let $X_1', ..., X_n'$ be independent copies of $X_1, ..., X_n$ and define $Z^{\setminus k} = f(X_1, ..., X_{k-1}, X_k', X_{k+1}, ..., X_n)$. Then the variance of Z is bounded above by

$$\mathrm{Var}(Z) \leq \frac{1}{2}\mathbf{E}\sum_{k=1}^n (Z - Z_i')^2 \tag{23}$$

Proposition 1.7 (Bounded difference inequality)**.** Suppose that f satisfies the bounded difference, i.e., for each $k = 1, ..., N$, we have

$$|f(x) - f(x^{\setminus k})| \leq L_k \tag{24}$$

for some L_k and elements of the random vector $X = (X_1, ..., X_N)$ are all independent. Then we have

$$\mathbf{P}[|f(X) - \mathbf{E}f(X)| \geq t] \leq 2\exp\left(-\frac{2t^2}{\sum_{k=1}^n L_k^2}\right) \tag{25}$$

i.e., $f(X)$ is sub-Gaussian.

Proof. First, we use the following lemma:

Lemma 1.3 (Azuma-Hoeffding)**.** Suppose that the martingale difference sequence $\{\Delta_k, \mathcal{F}_k\}_{k=1}^n$ is bounded such that $\Delta_k \in [a_k, b_k]$ almost surely. Then, for all $t \geq 0$, we have

$$\mathbf{P}\left[\left|\sum_{k=1}^n \Delta_k\right| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{\sum_{k=1}^n (b_k - a_k)^2}\right) \tag{26}$$

Applying iterated expectation and using sub-Gaussian property yields the conclusion easily. It remains to show that each summand of the Doob martingale,

$$\begin{aligned}
\Delta_k &= \mathbf{E}_k f(X) - \mathbf{E}_{k-1} f(X) && (27)\\
&= \mathbf{E}f(X)|X_1, ..., X_{k-1}, X_k - \mathbf{E}f(X)|X_1, ..., X_{k-1} && (28)
\end{aligned}$$

is bounded. First define

$$\begin{aligned}
A_k &= \inf_x \quad \mathbf{E}[f(X)|X_1, ..., X_{k-1}, x] - \mathbf{E}f(X)|X_1, ..., X_{k-1} && (29)\\
B_k &= \sup_x \quad \mathbf{E}[f(X)|X_1, ..., X_{k-1}, x] - \mathbf{E}f(X)|X_1, ..., X_{k-1} && (30)
\end{aligned}$$

Note that $A_k \leq \Delta_k \leq B_k$. It suffices to show that $B_k - A_k$ is bounded. Note that

$$\begin{aligned}
B_k - A_k &= \sup_x \mathbf{E}[f(X)|X_1, ..., X_{k-1}, x] - \inf_x \mathbf{E}[f(X)|X_1, ..., X_{k-1}, x] && (31)\\
&\leq \sup_x \mathbf{E}[f(X)|X_1, ..., X_{k-1}, x - f(X)|X_1, ..., X_{k-1}, y] && (32)\\
&= \sup_x \mathbf{E}_{k+1}[f(X_1, ..., X_{k-1}, x, X_{k+1}^n) - f(X_1, ..., X_{k-1}, y, X_{k+1}^n)] && (33)\\
&\leq L_k && (34)
\end{aligned}$$

where X_{k+1}^n is a vector $(X_{k+1}, ..., X_n)$.

□

Examples 1.4 (Bounding a pairwise U-statistics)**.** Let $X_k \in \mathbb{R}^d$ be a sequence of random variable and g be a function that takes values in two values of X. The quantity

$$U = \frac{1}{\binom{n}{2}} \sum_{j<k} g(X_j, X_k) \tag{35}$$

is a pairwise U-statistics. Suppose that g is uniformly bounded by b. Consider U as a function of $f(x) = f(x_1, ..., x_n)$. Then we have a bounded difference inequality since

$$|f(x) - f(x^{\backslash k})| \leq \frac{1}{\binom{n}{2}} \sum_{j \neq k} |g(x_j, x_k) - g(x_j, x_k')| \tag{36}$$

$$\leq \frac{(n-1)2b}{\binom{n}{2}} = \frac{4b}{n} \tag{37}$$

so we have, by the bounded difference inequality,

$$\mathbf{P}\left(|U - \mathbf{E}U| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{8b^2}\right) \tag{38}$$

Examples 1.5 (Kernel density estimation, Wainwright Exercise 2.15)**.** Let $\{X_i\}_{i=1}^n$ be iid sequence of random variables with density f. Consider the following kernel density estimate:

$$\widehat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{39}$$

where $\int_{-\infty}^{\infty} K(t)dt = 1$ and $h > 0$. Show that

$$\mathbf{P}\left(\|\widehat{f}_n - f\|_1 - \mathbf{E}\|\widehat{f}_n - f\|_1 \geq \delta\right) \geq \exp\left(-\frac{n\delta^2}{8}\right) \tag{40}$$

2 Concentration of measure

2.1 Entropy to sub-Gaussian tail bounds

Definition 2.1 ($\phi-$Entropy)**.** Given a convex function $\phi : \mathbb{R} \to \mathbb{R}$, $\varphi-$entropy of a random variable X is defined as

$$H_\phi(X) = \mathbf{E}\phi(X) - \phi(\mathbf{E}X) \tag{41}$$

One particular common choice of ϕ is $\phi(u) = u \log u$ (reminicent of Shannon entropy). Note that for this particular choice of the entropy, a random variable $Z = e^{\lambda X}$ yields the following representation:

$$H(e^{\lambda X}) = \lambda \varphi_x'(\lambda) - \varphi_x(\lambda) \log \varphi_x(\lambda) \tag{42}$$

where $\varphi_x(\lambda) = \mathbf{E}e^{\lambda X}$ is the mgf of X. Note that for a mean-zero Gaussian random variable Z with variance σ^2, we have

$$H(e^{\lambda Z}) = \frac{1}{2}\lambda^2\sigma^2\varphi_Z(\lambda) \tag{43}$$

which suggests that there should be a link to the entropy of random variables to sub-Gaussian tail behaviours as follows:

Proposition 2.1 (Herbst argument)**.** Suppose that the entropy $H(e^{\lambda X})$ satisfies the inequality

$$H(e^{\lambda X}) \leq \frac{1}{2}\lambda^2\sigma^2\varphi_X(\lambda) \tag{44}$$

Then we have

$$\log \mathbf{E}e^{\lambda(X-\mathbf{E}X)} \leq \frac{1}{2}\lambda^2\sigma^2 \tag{45}$$

Remarks 2.1. This implies that when the entropy of a random variable behaves in a sub-Gaussian way, it is essentially sub-Gaussian, which implies that its sub-Gaussian tail bound can be obtained using Chernoff arguments.

Now we develop some methods to bound entropy.

Lemma 2.1 (Entropy bound for univariate functions)**.** Suppose X is supported on $[a, b]$ and $g : \mathbb{R} \to \mathbb{R}$ is convex and Lipschitz. Then we have

$$H(e^{\lambda g(X)}) \leq \lambda^2(b-a)^2\mathbf{E}[(g'(X))^2e^{\lambda g(X)}] \tag{46}$$

Lemma 2.2 (Entropy tensorization)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ and let $\{X_k\}_{k=1}^n$ be independent random variables and let X be its vectorized form. Then we have

$$H(e^{\lambda f(X)}) \leq \mathbf{E}\left[\sum_{k=1}^{n} H(e^{\lambda f_k(X_k)})|X^{\setminus k})\right] \tag{47}$$

The first lemma can be proven easily with the Rademacher lemma. The second lemma is somewhat difficult to derive on its own (see Lemma 3.8 in Wainwright – TODO). Combining these results together, we have the following powerful result:

Proposition 2.2. Let $X = (X_1, ..., X_n)$ be independent random variables that are each supported on $[a, b]$. Let f be separately convex and $L-$Lipschitz with Euclidean norm. Then, for all $\delta > 0$ we have

$$\mathbf{P}\left(f(X) \geq \mathbf{E}f(X) + \delta\right) \leq \exp\left(-\frac{\delta^2}{4L^2(b-a)^2}\right) \tag{48}$$

2.2 Tail bounds for empirical process

One very practical use of the entropy methods involves Hoeffding bounds for empirical processes. To begin with, we provide an alternative way to bound the entropy:

Lemma 2.3 (Entropy bound for univariate functions, second)**.** Let $X, Y \sim \mathbf{P}$ be a pair of iid variates. Then, for any $g : \mathbb{R} \to \mathbb{R}$, we have

$$H(e^{\lambda\gamma(X)}) \leq \lambda^2 \mathbf{E}\left[(g(X) - g(Y))^2 e^{\lambda g(X)} 1\left[g(X) \geq g(Y)\right]\right] \tag{49}$$

Proof. Proof technique is similar to that of symmetrization arguments. First, by the definition of entropy, we have

$$\begin{aligned}
H(e^{\lambda g(X)}) &= \mathbf{E}_X \lambda g(X) e^{\lambda g(X)} - \mathbf{E}_X e^{\lambda g(X)} \log \mathbf{E}_X e^{\lambda g(X)} && (50)\\
&\leq \mathbf{E}_X \lambda g(X) e^{\lambda g(X)} - \mathbf{E}_{X,Y} e^{\lambda g(X)} \lambda g(Y) && (51)\\
&= \frac{1}{2} \mathbf{E}_{X,Y} \lambda (g(X) - g(Y)) \left(e^{\lambda g(X)} - e^{\lambda g(Y)}\right) && (52)\\
&= \mathbf{E}_{X,Y} \lambda (g(X) - g(Y)) \left(e^{\lambda g(X)} - e^{\lambda g(Y)}\right) 1\left[g(X) \geq g(Y)\right] && (53)\\
&\leq \mathbf{E}_{X,Y} \lambda^2 (g(X) - g(Y))^2 e^{\lambda g(X)} 1\left[g(X) \geq g(Y)\right] && (54)
\end{aligned}$$

where the first inequality follows from Jensen's inequality and the last inequality follows by convexity of the exponential ($e^x - e^y \leq e^x(x - y)$ for all x, y) □

Proposition 2.3 (Functional Hoeffding theorem)**.** Let $Z = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(X_i)$. Suppose that for each $f \in \mathcal{F}$ $f(x)$, $f(x) \in [a_{i,f}, b_{i,f}]$ for all $x \in \mathcal{X}_i$. Then for all $\delta \geq 0$, we have

$$\mathbf{P}\left(Z \geq \mathbf{E}Z + \delta\right) \leq \exp\left(-\frac{n\delta^2}{4 \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (b_{i,f} - a_{i,f})^2}\right) \tag{55}$$

Proof. It suffices to prove the result for a finite class of functions $\mathcal{F}$ as we can take the general result by taking limits over an increasing sequence of such finite classes. First, consider Z as a function that takes random variables $(X_1, ..., X_n)$. For each index $j = 1, ..., n$, define the random function $x_j \to Z_j(x_j) := Z(X_1, ..., X_{j-1}, x_j, X_{j+1}, ..., X_n)$. First, by the lemma above and tensorization lemma, we have

$$H(e^{\lambda g(X)}) \leq \lambda^2 \mathbf{E}\left[\sum_{j=1}^n \mathbf{E}_{X_j}(Z_j(X_j) - Z_j(Y_j))^2 e^{\lambda g(X)} 1\left[g(X) \geq g(Y)\right] | X^{\setminus j}\right] \tag{56}$$

so it suffices to bind the quantity $(Z_j(X_j) - Z_j(Y_j))^2 1\left[Z_j(X_j) \geq Z_j(Y_j)\right]$. Define the disjoint set $\mathcal{A}(f) := \{(x_1, ..., x_n) | Z = \sum_{i=1}^n f(x_i)\}$. For any $x \in \mathcal{A}$, we have that $Z_j(x_j) - Z_j(y_j) \leq$

$f(x_j) - f(y_j)$. Thus we have that

$$(Z_j(x_j) - Z_j(y_j))^2 1\left[Z_j(x_j) \geq Z_j(y_j)\right] \leq \sum_{f \in \mathcal{F}} 1[x \in \mathcal{A}(f)](Z_j(x_j) - Z_j(y_j))^2 \tag{57}$$

$$\leq \sum_{f \in \mathcal{F}} 1[x \in \mathcal{A}(f)](b_{j,f} - a_{j,f})^2 \tag{58}$$

$$\leq \sup_{f \in \mathcal{F}} (b_{j,f} - a_{j,f})^2 \tag{59}$$

Combining this with the sub-Gaussian entropy bound with $t = n\delta$ gives the result. □

2.3 Transport inequality

Here is the main philosophy:

Remarks 2.2 (Wasserstein distance bounds f of the form $\int f(d\mathbf{Q} - d\mathbf{P})$)**.** Transport inequality is useful since any Wasserstein distance gives an upper bound for any Lipschitz f function of the form

$$\int f(d\mathbf{Q} - d\mathbf{P}) \tag{60}$$

by the dual representation of optimal transport problems.

Definition 2.2 ($\rho-$transportation cost inequality)**.** Define the KL divergence as

$$D(\mathbf{Q}|\mathbf{P}) = \mathbf{E}_{\mathbf{Q}} \log \frac{d\mathbf{Q}}{d\mathbf{P}} \tag{61}$$

$$= \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x)} \tag{62}$$

We say that the probability measure $\mathbf{P}$ satisfies $\rho-$transportation cost inequality with $\gamma > 0$ if for all probability measure $\mathbf{Q}$ we have

$$W_\rho(\mathbf{Q}, \mathbf{P}) \leq \sqrt{2\gamma D(\mathbf{Q}|\mathbf{P})} \tag{63}$$

Examples 2.1 (Pinsker-Csiszar-Kullback inequality for bounded variation)**.** Let $\mathbf{P}, \mathbf{Q}$ be probability distributions. With $\rho(x, x') = 1(x \neq x')$, the Wasserstein distance with respect to ρ between two distributions is the total variation distance:

$$\|\mathbf{Q} - \mathbf{P}\|_{TV} = \sup_{A \subset \mathbf{P}(X)} |\mathbf{Q}(A) - \mathbf{P}(A)| \tag{64}$$

Now we show that the transportation cost can be used to offer a sharp bound.

Proposition 2.4. [From transporation cost to concentration for Lipschitz functions] Let f be a Lipschitz function and $(\mathbf{P}, \mathcal{X}, \rho)$ satisfy the $\rho-$transportation cost inequality. Then we have

$$\mathbf{P}\left(|f(X) - \mathbf{E}f(X) \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\gamma L^2}\right) \tag{65}$$

Proof. Note that for any positive numbers (u, v, λ), $\sqrt{2uv} \leq u\lambda/2 + v/\lambda$ is satisfied. Thus we have

$$\int f(d\mathbf{Q} - d\mathbf{P}) \leq LW_\rho(\mathbf{Q}, \mathbf{P}) \leq \sqrt{2L^2\gamma D(\mathbf{Q}|\mathbf{P})} \tag{66}$$

$$\leq \frac{\lambda\gamma L^2}{2} + \frac{1}{\lambda}D(\mathbf{Q}|\mathbf{P}) \tag{67}$$

where the first inequality follows from the dual form of the optimal transport cost function.

Now define a distribution $\mathbf{Q}$ with Radon-Nikodym derivative

$$\frac{d\mathbf{Q}}{d\mathbf{P}}(x) = \frac{e^{g(x)}}{\mathbf{E}_\mathbf{P} e^{g(x)}} \tag{68}$$

with $g(x) = \lambda(f(x) - \mathbf{E}_\mathbf{P} f) - L^2\gamma\lambda^2/2$. Then we have

$$D(\mathbf{Q}|\mathbf{P}) = \mathbf{E}_\mathbf{Q} \log\left(\frac{e^{g(X)}}{\mathbf{E}_\mathbf{P} e^{g(X)}}\right) \tag{69}$$

$$= \lambda\left[\mathbf{E}_\mathbf{Q} f(X) - \mathbf{E}_\mathbf{P} f(X)\right] - \frac{\gamma L^2\lambda^2}{2} - \log \mathbf{E}_\mathbf{P} e^{g(X)} \tag{70}$$

Combining this with the inequality above, we have that $\log \mathbf{E}_\mathbf{P} e^{g(X)} \leq 0$, or equivalently

$$\mathbf{E}_\mathbf{P} e^{\lambda(f(X) - \mathbf{E}_{X'}\mathbf{P} f(X'))} \leq \exp\left(\frac{\lambda^2\gamma L^2}{2}\right) \tag{71}$$

The result follows from the Chernoff bound. The same argument can be applied to $-f$ to yield the lower tail bound as well. □

Proposition 2.5. [Transport cost tensorization] Suppose that a univariate distribution $\mathbf{P}_k$ satisfies ρ_k-transporation cost inequality with parameter γ_k. Then the product distribution $\mathbf{P} = \otimes_{k=1}^N \mathbf{P}_k$ satisfies the tranportation cost inequality with

$$W_\rho(\mathbf{Q}, \mathbf{P}) \leq \sqrt{2\left(\sum_{k=1}^n \gamma_k\right) D(\mathbf{Q}|\mathbf{P})} \tag{72}$$

Examples 2.2 (Bounded difference inequality, using transportation argument)**.** Suppose that f satisfies the bounded difference inequality with parameter L_k for each coordinate k. Then it is not difficult to show that f is a 1-Lipschitz function with respect to the rescaled

Hamming metric $\rho(x,y) = \sum_{k=1}^{n} \rho_k(x_k, y_k)$. Note that by the Pinsker-Csiszar-Kullback inequality, each univariate distribution $\mathbf{P}_k$ then satisfies a ρ_k-transportation cost inequality with $\gamma_k = L_k^2/4$. Hence, by Proposition 2.4 and Proposition 2.5, we have

$$\mathbf{P}\left(|f(X) - \mathbf{E}f(X)| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{k=1}^{n} L_k^2}\right) \tag{73}$$

Now we show that Wasserstein distance bound, with l_2-norm, yields **dimension-independent** concentration inequalities. To do so, let us first introduce a coupling-based distance between probability distributions that is asymmetric in arguments.

Definition 2.3 (Coupling distance between two distributions)**.** A coupling distance $\mathbf{Q}$ on $\mathbf{P}$ is defined as

$$C(\mathbf{Q}|\mathbf{P}) = \inf_{\gamma \in \prod(\mathbf{P},\mathbf{Q})} \sqrt{\int \sum_{i=1}^{n} \gamma(Y_i \neq x_i | X_i = x_i)^2 d\mathbf{P}(x)} \tag{74}$$

Note that this can be alternatively defined as

$$C(\mathbf{Q}|\mathbf{P}) = \sqrt{\int \left|1 - \frac{d\mathbf{Q}}{d\mathbf{P}}(x)\right|_+^2 d\mathbf{P}(x)} \tag{75}$$

Remarks 2.3 (Why coupling distance is useful)**.** Although the coupling distance is not Wasserstein, as Wasserstein is useful to yield an upper bound when f is Lipschitz (Remarks 2.2), it can be used to upper bound usch differences when f is Lipschitz and convex and bounded.

For proof, we need the following result:

Lemma 2.4 (Samson, 2000)**.** For any product distribution $\mathbf{P}$ in n variables, we have

$$\max[C(\mathbf{Q}|\mathbf{P}), C(\mathbf{P}|\mathbf{Q})] \leq \sqrt{2D(Q|P)} \tag{76}$$

Here's an actual theorem and result:

Proposition 2.6 (Dimension-free bounds for Lipschitz functions)**.** Consider a vector of independent $X_1, ..., X_n$, bounded on $[0,1]$, and let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and $L-$Lipschitz with respect to the Euclidean norm. Then for all $t \geq 0$, we have

$$\mathbf{P}\left(|f(X) - \mathbf{E}f(X)| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2L^2}\right) \tag{77}$$

Proof. By convexity of f, we have

$$f(y) - f(x) \leq \sum_{j=1}^{n} \left| \frac{\partial f}{\partial y_j}(y) \right| |x_j - y_j| \tag{78}$$

$$\leq \sum_{j=1}^{n} \left| \frac{\partial f}{\partial y_j}(y) \right| 1(x_j \neq y_j) \tag{79}$$

where the second inequality follows by the assumption that $X_1, ..., X_n$ is bounded on $[0, 1]$. Hence we have that, for any coupling γ on $(\mathbf{P}, \mathbf{Q})$,

$$\int f(y)d\mathbf{Q} - \int f(x)d\mathbf{P} \leq \sum_{j=1}^{n} \left| \frac{\partial f}{\partial y_j}(y) \right| 1(x_j \neq y_j) d\gamma(x, y) \tag{80}$$

$$= \int \sum_{j=1}^{n} \left| \frac{\partial f}{\partial y_j}(y) \right| \gamma(X_j \neq y_j | Y_j = y_k) d\mathbf{Q}(y) \tag{81}$$

$$\leq \int \|\nabla f(y)\|_2 \sqrt{\sum_{j=1}^{n} \gamma^2(X_j \neq y_j | Y_j = y_j)} d\mathbf{Q}(y) \quad \text{(by CS)} \tag{82}$$

$$\leq \int L \sqrt{\sum_{j=1}^{n} \gamma^2(X_j \neq y_j | Y_j = y_j)} d\mathbf{Q}(y) \quad \text{(by Lipschitz)} \tag{83}$$

$$\leq L \sqrt{\int \sum_{j=1}^{n} \gamma^2(X_j \neq y_j | Y_j = y_j) d\mathbf{Q}(y)} \quad \text{(by Jensen)} \tag{84}$$

$$\leq LC(\mathbf{P}|\mathbf{Q}) \tag{85}$$

The result follows from Samson. The lower bound can be analogously derived by considering a concave Lipschitz functions. □

3 ULLN and metric entropy

3.1 Uniform laws of large numbers

Proposition 3.1 (Glivenko-Cantelli theorem)**.** For any distribution, $\|\widehat{F}_n - F\|_\infty \to_{a.s.} 0$, where $\widehat{F}_n$ is the empirical CDF for F.

Definition 3.1 (Glivenko-Cantelli class)**.** We say that $\mathcal{F}$ is a Glivenko-Cantelli class for $\mathbf{P}$ if

$$\|\mathbf{P}_n - \mathbf{P}\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbf{E}f(X) \right| \tag{86}$$

converges to zero in probability.

In statistics, it is common to study the excess risk by bounding

$$E(\widehat{\theta}, \theta^*) = R(\widehat{\theta}, \theta^*) - \inf_{\theta \in \Omega_0} R(\theta, \theta^*) \tag{87}$$

where $\widehat{\theta}_n$ is the estimate from the principle of empirical risk minimization

$$\widehat{\theta} = \arg\max_{\theta} \widehat{R}_n(\theta, \theta^*) \tag{88}$$

$$= \arg\max_{\theta} \frac{1}{n} \sum_{i=1}^{n} L_\theta(X_i) \tag{89}$$

and the population risk is defined as

$$R(\theta, \theta^*) = \mathbf{E}_{\theta^*} L_\theta(X) \tag{90}$$

for some loss function $L(X)$.

Examples 3.1 (Maximum likelihood estimate)**.** Consider $\mathcal{F}$ parametrized by $\theta \in \Theta$. The loss function corresponding to maximum likelihood estimator is then

$$L_\theta(x) = \log \frac{p_{\theta^*}(x)}{p_\theta(x)} \tag{91}$$

whose population risk is in fact KL-divergence:

$$R(\theta, \theta^*) = \mathbf{E}_{\theta^*} \log \frac{p_{\theta^*(X)}}{p_\theta(X)} \tag{92}$$

Examples 3.2 (Binary classification)**.** Consider a binary estimation such that $f : \mathbb{R} \to \{-1, +1\}$ that minimizes $\mathbf{P}[f(X) \neq Y]$. The corresponding loss function is

$$L_f(X, Y) = 1(f(X) \neq Y) \tag{93}$$

Remarks 3.1 (Decomposing excess risk)**.** Suppose that the population risk supremum is achieved by some $\theta_0 \in \Omega_0$. Note that we can then decompose the excess risk as follows:

$$E(\widehat{\theta}, \theta^*) = R(\widehat{\theta}, \theta^*) - R(\theta_0, \theta^*) \tag{94}$$

$$= R(\widehat{\theta}, \theta^*) - \widehat{R}_n(\widehat{\theta}, \theta^*) + \widehat{R}_n(\widehat{\theta}, \theta^*) - \widehat{R}_n(\theta_0, \theta^*) + \widehat{R}_n(\theta_0, \theta^*) - R(\theta_0, \theta^*) \tag{95}$$

The second difference $\widehat{R}_n(\widehat{\theta}, \theta^*) - \widehat{R}_n(\theta_0, \theta^*)$ is non-positive since it minimizes $\widehat{R}_n$ over Ω_0. The third difference, $\widehat{R}_n(\theta_0, \theta^*) - R(\theta_0, \theta^*)$, can be easily bounded since θ_0 is a fixed quantity.

The first difference is quite tricky though, since it depends on $\widehat{\theta}$ which depends on the samples. Note that we have

$$R(\widehat{\theta}, \theta^*) - \widehat{R}_n(\widehat{\theta}, \theta^*) = \mathbf{E}_X L_{\widehat{\theta}}(X) - \frac{1}{n} \sum_{i=1}^{n} L_{\widehat{\theta}}(X_i) \tag{96}$$

$$= \sup_{\theta \in \Omega_0} \left| \mathbf{E}_X L_{\widehat{\theta}}(X) - \frac{1}{n} \sum_{i=1}^{n} L_{\widehat{\theta}}(X_i) \right| \tag{97}$$

$$= \|\mathbf{P}_n - \mathbf{P}\|_{\mathcal{L}(\Omega_0)} \tag{98}$$

where $\mathcal{L}(\Omega_0) := \{x \to L_\theta(x); \theta \in \Omega_0\}$.

3.2 A uniform law via Rademacher complexity

Now we introduce the notion of Rademacher complexity, which measures the possible largest correlation with a randomly drawn noise vector.

Definition 3.2 (Rademacher complexity)**.** For a fixed collection x_1^n of n points, the **empirical Rademacher complexity** is defined as

$$\mathcal{R}(\mathcal{F}(x_1^n)/n) := \mathbf{E}_\varepsilon \sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right| \tag{99}$$

The **Rademacher complexity** of the function class $\mathcal{F}$ is defined as the expectation of the empirical Rademacher complexity:

$$\mathcal{R}_n(\mathcal{F}) := \mathbf{E}_{X,\varepsilon} \sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \tag{100}$$

We show that $\mathcal{R}_n(\mathcal{F}) = o(1)$ implies the Glivenko-Cantelli property.

Proposition 3.2 (Rademacher complexity to Glivenko-Cantelli)**.** Let $\mathcal{F}$ be a $b-$uniformly bounded class of functions. Then, for any $n \geq 1$ and $\delta > 0$, we have

$$\|\mathbf{P}_n - \mathbf{P}\|_\mathcal{F} \leq 2\mathcal{R}_n(\mathcal{F}) + \delta \tag{101}$$

with $\mathbf{P}-$probability at least $1-\exp(-n\delta^2/2b^2)$. Hence, $\mathcal{R}_n(\mathcal{F}) = o(1)$ implies $\|\mathbf{P}_n - \mathbf{P}\|_\mathcal{F} \to 0$ almost surely.

Proof. We first prove that (1) $\|\mathbf{P}_n - \mathbf{P}\|_\mathcal{F}$ is concentrated around mean with the rate of $\exp(-n\delta^2/2b^2)$, and then (2) show that its mean is bounded uniformly by $2\mathcal{R}_n(\mathcal{F})$.

(1) Concentration around mean. We use bounded difference inequality in Proposition 1.7. Define $\overline{f}(x) := f(x) - \mathbf{E}f(X)$ and rewrite the random variable $\|\mathbf{P}_n - \mathbf{P}\|_\mathcal{F}$ as $\sup_{f\in\mathcal{F}} |n^{-1}\sum_{i=1}^n \overline{f}(X_i)|$. For each X, define the function $G(x) := \sup_{f\in\mathcal{F}} |n^{-1}\sum_{i=1}^n \overline{f}(x_i)|$. It is an easy exercise to show that G satisfies bounded difference inequality, thus following Proposition 1.7, we have

$$\|\mathbf{P}_n - \mathbf{P}\|_\mathcal{F} - \mathbf{E}\|\mathbf{P}_n - \mathbf{P}\|_\mathcal{F} \leq t \tag{102}$$

holds with the probability of $1 - \exp(-nt^2/2b^2)$ for all $t \geq 0$.

(2) Upper bound on mean by Rademacher complexity. We use symmetrization

arguments; let Y be an independent copy of X:

$$\mathbf{E}_X \|\mathbf{P}_n - \mathbf{P}\|_{\mathcal{F}} = \mathbf{E}_X \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbf{E}f(X) \right| \tag{103}$$

$$= \mathbf{E}_X \sup_{f\in\mathcal{F}} \mathbf{E}_Y \left| \frac{1}{n}\sum_{i=1}^{n} [f(X_i) - f(Y_i)] \right| \tag{104}$$

$$= \mathbf{E}_{X,Y} \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} [f(X_i) - f(Y_i)] \right| \tag{105}$$

$$= \mathbf{E}_{X,Y,\varepsilon} \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i [f(X_i) - f(Y_i)] \right| \tag{106}$$

$$\leq 2\mathcal{R}_n(\mathcal{F}) \tag{107}$$

□

It turns out the empirical process is bounded below by the Rademacher complexity as well; thus, the Rademacher complexity provides both necessary and sufficient conditions for the empirical process to be Glivenko-Cantelli.

Proposition 3.3 (Glivenko-Cantelli to Rademacher complexity)**.** For any $b-$bounded function class $\mathcal{F}$, any integer $n \geq 1$ and $\delta \geq 0$, we have

$$\|\mathbf{P}_n - \mathbf{P}\|_{\mathcal{F}} \geq \frac{1}{2}\mathcal{R}_n(\mathcal{F}) - \frac{\sup_{f\in\mathcal{F}} \mathbf{E}f}{2\sqrt{n}} - \delta \tag{108}$$

with $\mathbf{P}-$probability at least $1 - \exp(-nb^2/2b^2)$.

See Proposition 4.12 in Wainwright for proof.

3.3 VC dimension

First, we show that the empirical Rademacher complexity of a function collection can be bounded above by polynomial discrimination:

Definition 3.3 (Polynomial discrimination)**.** A class of functions $\mathcal{F}$ has polynomial discrimination of order $\nu \geq 1$ if for each positive integer n and a fixed collection x_1^n, the cardinality of the set $\mathcal{F}(x_1^n)$ is upper bounded by $(n+1)^\nu$.

Lemma 3.1 (Bounding empirical Rademacher complexity by polynomial discrimination order)**.** See Lemma 4.14 in Wainwright.

$$\mathcal{R}(\mathcal{F}(x_1^n/n)) \leq 4 \sup_{f\in\mathcal{F}} \sqrt{\frac{\sum_{i=1}^n f^2(x_i)}{n}} \sqrt{\frac{\nu \log(n+1)}{n}} \tag{109}$$

Now we show that a polynomial discrimination order can be bounded by the VC dimension.

Definition 3.4. Let $\mathcal{F}$ be a class of binary-valued functions. x_1^n is said to be shattered by $\mathcal{F}$ if the cardinality of $\mathcal{F}(x_1^n)$ is 2^n. The VC dimension $\nu(\mathcal{F})$ is the largest integer n for which there is some collection $x_1^n = (x_1, ..., x_n)$ of n points that is shattered by $\mathcal{F}$.

Proposition 3.4 (Sauer–Shelah lemma)**.** Suppose $\nu(S) < \infty$. Then for any collection of points $P = (x_1, ..., x_n)$ with $n \geq \nu(S)$, we have

$$\text{card}(S(P)) \leq \sum_{i=0}^{\nu(S)} \binom{n}{i} \leq (n+1)^{\nu(S)} \tag{110}$$

4 Sub-Gaussian process and metric entropy

In the last section we have studied how controlling Rademacher complexity can be used to show Glivenko-Cantelli properties of a function class. In this section, we generalize this idea to a general sub-Gaussian process and how it can be bounded by metric entropy.

4.1 Sub-Gaussian process

4.1.1 Gaussian complexity

Recall that the Rademacher complexity is defined as

$$\mathcal{R}(\mathcal{F}) := \mathbf{E}_\varepsilon \sup_{\theta \in \mathcal{F}} \langle \theta, \varepsilon \rangle \tag{111}$$

Similarly, one can define the Gaussian complexity:

Definition 4.1 (Gaussian complexity)**.** A **canonical Gaussian process** associated with $\theta \in \mathcal{F}$ is defined as, for $\mathcal{F} \subset \mathbb{R}^d$,

$$G_\theta := \langle w, \theta \rangle \tag{112}$$

where w is a $d-$vector of iid standard normal random variables. The corresponding **Gaussian complexity** is defined as

$$\mathcal{G}(\mathcal{F}) := \mathbf{E}_w \sup_{\theta \in \mathcal{F}} \langle \theta, w \rangle \tag{113}$$

Here Rademahcer complexity is in fact bounded above by Gaussian compleixty:

Lemma 4.1 (Rademacher complexity is bounded by Gaussian complexity)**.** For any set $\mathcal{F}$, the following inequality holds:

$$\mathcal{R}(\mathcal{F}) \leq \sqrt{\frac{\pi}{2}}\mathcal{G}(\mathcal{F}) \tag{114}$$

The following result shows that $b-$uniformly bounded functions have Gaussian complexity bounded by b:

Proposition 4.1 (Trivial bound for uniformly bounded functions)**.** Suppose $\|f\|_\infty \leq b$ for all $f \in \mathcal{F}$. Then, we have

$$\mathcal{G}\left(\frac{\mathcal{F}(X_1^n)}{n}\right) = \mathbf{E}_w \sup_{f\in\mathcal{F}} \sum_{i=1}^n \frac{w_i}{\sqrt{n}} \frac{f(x_i)}{\sqrt{n}} \tag{115}$$

$$\leq b\frac{\mathbf{E}\|w\|_2}{\sqrt{n}} \tag{116}$$

$$\leq b \tag{117}$$

4.1.2 Sub-Gaussian process

We can generalize the notion of Gaussian and Rademacher process as follows:

Definition 4.2 (Sub-Gaussian process)**.** A collection of zero-mean random variables X_θ is a sub-Gaussian process with respect to ρ_X if

$$\mathbf{E}e^{\lambda(X_\theta - X_{\tilde{\theta}})} \leq e^{\frac{\lambda\rho_X^2(\theta,\tilde{\theta})}{2}} \tag{118}$$

which implies $\mathbf{P}(|X_\theta - X_{\tilde{\theta}}| \geq t) \leq 2e^{-\frac{t^2}{2\rho_X^2(\theta,\tilde{\theta})}}$

Note that Gaussian and Rademacher processes are both sub-Gaussian with respect to the Euclidean metric.

4.2 Metric entropy

Definition 4.3 ($\delta-$covering number $N(\delta, \mathcal{F}, \rho)$)**.** A $\delta-$cover of a set $\mathcal{F}$ with respect to ρ is a set $\{f_1, ..., f_N\} \in \mathcal{F}$ such that for each $f \in \mathcal{F}$, there exists i such that $\rho(\theta, \theta_i) \leq \delta$. The $\delta-$covering number $N(\delta, \mathcal{F}, \rho)$ is the cardinality of the smallest $\delta-$cover. We use $N_\infty(\delta, \mathcal{F})$ for the $\delta-$covering number with respect to sup-norm.

Definition 4.4 ($\delta-$packing number $M(\delta, \mathcal{F}, \rho)$)**.** A $\delta-$packing of a set $\mathcal{F}$ with respect to ρ is a set $\{f_1, ..., f_N\} \in \mathcal{F}$ such that for each i, j, $\rho(\theta_i, \theta_j) > \delta$. The $\delta-$packing number $M(\delta, \mathcal{F}, \rho)$ is the cardinality of the largest $\delta-$packing.

The following lemma is useful in describing the scale of metric entropy:

Lemma 4.2. For $\delta > 0$, the packing and covering numbers are related as follows:

$$M(2\delta; \mathcal{F}, \rho) \leq N(\delta; \mathcal{F}, \rho) \leq M(\delta; \mathcal{F}, \rho) \tag{119}$$

Examples 4.1 (A parametric class of functions)**.** Let $f_\theta(x) := 1 - e^{-\theta x}$ and consider the function class $\mathcal{P} = \{f_\theta : [0,1] \to \mathbb{R}; \theta \in [0,1]\}$ equipped with the sup-norm. We claim that the covering number is bounded above and below as

$$1 + \frac{1 - 1/e}{2\delta} \leq N_\infty(\delta, \mathcal{P}) \leq \frac{1}{2\delta} + 2 \tag{120}$$

The upper bound can be established by taking the $\delta-$cover for $\mathcal{P}$ by considering $\{\theta_i\}_{i=1}^T$ that has been evenly placed on $[0,1]$ by $T = 1/2\delta$. The lower bound can be established by bounding the packing number, by defining $\theta_i = -\log(1 - \delta i)$ for $i = 1, ..., T$ with $T = (1 - 1/e)/\delta$ and showing that $\|f_{\theta_i} - f_{\theta_j}\| \geq \delta$.

Examples 4.2 (Lipschitz functions on the unit interval)**.** Proof is constructive. Let $\mathcal{F}_L$ be a class of L-Lipschitz functions on the unit interval $[0,1]$. We claim that the metric entropy of the class $\mathcal{F}_L$ scales as

$$\log N_\infty(\delta; \mathcal{F}_L) \asymp \frac{L}{\delta} \tag{121}$$

For a given $\varepsilon > 0$ to be determined later, slice $[0,1]$ evenly by $\varepsilon = \delta/L$ to get $\{x_i\}_{i=1}^M$ where $M = 1/\varepsilon$, and construct the foolowing functions:

$$f_\beta(y) = \sum_{i=1}^M \beta_i L \varepsilon \phi\left(\frac{y - x_i}{\varepsilon}\right) \tag{122}$$

where $\beta \in \{-1, +1\}^M$ and $\phi(u)$ is a kernel

$$\phi(u) = \begin{cases} 0 & \forall u < 0 \\ u & \forall u \in [0,1] \\ 1 & \forall u > 1 \end{cases} \tag{123}$$

Verify that functions in $\{f_\beta\}$ are L-Lipschitz and construct a covering and packing.

4.3 Controlling sub-Gaussian processes by metric entropy

Proposition 4.2 (One-step discretization bound)**.** Let X_θ be a zero mean sub-Gaussian process with respect to the metric ρ_X with $D := \sup_{\theta, \widetilde{\theta} \in \mathcal{F}} \rho_X(\theta, \widetilde{\theta})$. Then for any $\delta \in [0, D]$ such that $N_X(\delta; \mathcal{F}) \geq 10$, we have

$$\mathbf{E} \sup_{\theta, \tilde{\theta} \in \mathcal{F}} (X_\theta - X_{\tilde{\theta}}) \leq 2\mathbf{E} \sup_{\gamma, \tilde{\gamma} \in \mathcal{F}; \rho_X(\gamma, \tilde{\gamma}) \leq \delta} (X_\gamma - X_{\tilde{\gamma}}) + 4\sqrt{D^2 \log N_X(\delta; \mathcal{F})} \tag{124}$$

Remarks 4.1. Even though the one-step discretization bound at first glampse looks like a bound for just difference, the zero-mean assumption makes it applicable to $\mathbf{E}\sup_{\theta\in\mathcal{F}} X_\theta$ since

$$\mathbf{E}\sup_{\theta\in\mathcal{F}} X_\theta = \mathbf{E}\sup_{\theta\in\mathcal{F}} (X_\theta - X_{\theta_0}) \tag{125}$$

$$\leq \mathbf{E}\sup_{\theta,\widetilde{\theta}\in\mathcal{F}} \left(X_\theta - X_{\widetilde{\theta}}\right) \tag{126}$$

A similar argument can be applied to the Gaussian compleixty as well:

Proposition 4.3 (Bound by localized Gaussian compleixty)**.** Let $\mathcal{G}(\widetilde{\mathcal{F}}(\delta))$ be a localized Gaussian complexity where

$$\widetilde{\mathcal{F}}(\delta) := \{\gamma - \gamma' | \gamma, \gamma' \in \mathcal{F}, \|\gamma - \gamma'\|_2 \leq \delta\} \tag{127}$$

Then we have

$$\mathcal{G}(\mathcal{F}) \leq \min_{\delta\in[0,D]} \left(\mathcal{G}(\widetilde{\mathcal{F}}(\delta)) + 2\sqrt{D^2 \log N_2(\delta;\mathcal{F})}\right) \tag{128}$$

By the Cauchy-Schwarz, we have $\mathcal{G}(\widetilde{T}(\delta)) \leq \delta\mathbf{E}\|w\|_2 \leq \delta\sqrt{d}$. This results in the naive discretization bound:

Corollary 4.1 (Naive discretization bound for Gaussian complexity)**.**

$$\mathcal{G}(\mathcal{F}) \leq \min_{\delta\in[0,D]} \left(\delta\sqrt{d} + 2\sqrt{D^2 \log N_2(\delta;\mathcal{F})}\right) \tag{129}$$

Examples 4.3 (Gaussian compleixty for smoothness classes)**.** Recall that a class of $L-$Lipschitz functions on the unit interval, $\mathcal{F}_L$, has log-covering number scaled by L/δ. Thus, for a sufficiently small $\delta_0 > 0$, we have $\log N_\infty(\delta;\mathcal{F}_L) \leq cL/\delta$ for some constant c for all $\delta \in (0,\delta_0)$. Using the naive bound, we have

$$\mathcal{G}(\mathcal{F}_L(x_1^n)/n) \leq \frac{1}{\sqrt{n}} \inf_{\delta\in(0,\delta_0)} \left(\delta\sqrt{n} + 3\sqrt{\frac{cL}{\delta}}\right) \tag{130}$$

Optimizing the upper bound by $\delta = n^{-1/3}$, we have

$$\mathcal{G}(\mathcal{F}_L(x_1^n)/n) \lesssim n^{-1/3} \tag{131}$$

Proposition 4.4 (Dudley's entropy integral bound)**.** Let X_θ be a zero mean sub-Gaussian process with respect to the metric ρ_X with $D := \sup_{\theta,\widetilde{\theta}\in\mathcal{F}} \rho_X(\theta,\widetilde{\theta})$. Then for any $\delta \in [0,D]$, we have

$$\mathbf{E}\sup_{\theta,\tilde{\theta}\in\mathcal{F}} (X_\theta - X_{\tilde{\theta}}) \leq 2\mathbf{E}\sup_{\gamma,\tilde{\gamma}\in\mathcal{F};\rho_X(\gamma,\tilde{\gamma})\leq\delta} (X_\gamma - X_{\tilde{\gamma}}) + 32\int_\delta^D \sqrt{\log N_X(u;\mathcal{F})}du \tag{132}$$

In particular, with $\delta = 0$, we have

$$\mathbf{E}\sup_{\theta,\tilde{\theta}\in\mathcal{F}} (X_\theta - X_{\tilde{\theta}}) \leq 32\int_0^D \sqrt{\log N_X(u;\mathcal{F})}du \tag{133}$$

This results in the foolowing well-known bound for VC classes:

Proposition 4.5 (Bounds for VC classes)**.** It is well known that for finite VC-class $\mathcal{F}$ with constant ν, the metric entropy can be bounded as follows:

$$N(\epsilon; \mathcal{F}, \|\cdot\|_{\mathbf{P}_n}) \leq C\nu(16e)^{\nu}\frac{b}{\epsilon}^{2\nu} \tag{134}$$

Define the zero-mean random variable $Z_f := n^{-1/2}\sum_{i=1}^{n}\varepsilon_i f(x_i)$ and consider the stochastic process $\{Z_f | f \in \mathcal{F}\}$ where $\mathcal{F}$ is a class of functions that are uniformly bounded by b. Note that $Z_f - Z_g$ is sub-Gaussian with parameter $\|f-g\|^2_{\mathbf{P}_n} = n^{-1}\sum_{i=1}^{n}(f(x_i)-g(x_i))^2$, which is bounded above by $2b$. Thus, by Dudley's entropy integral, we have

$$\mathbf{E}_{\varepsilon}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x_i)\right|\right) \leq \frac{24}{\sqrt{n}}\int_0^{2b} N(t; \mathcal{F}, \|\cdot\|_{\mathbf{P}_n})dt \tag{135}$$

$$\leq c_0\sqrt{\frac{\nu}{n}} \tag{136}$$