ORIE 7790 High Dimensional Probability and Statistics Lecture 5–6 - 02/04-06/2020

Lectures 5–6: Concentration for Lipschitz Functions

Lecturer: Yudong Chen *Scribe: Sean Sinclair, Connor Lawless*

Reading:

- M. J. Wainwright, "High-dimensional statistics: A non-asymptotic viewpoint", Section 3.1.
- J. Duchi, "Lecture notes for Statistics 311/Electrical Engineering 377: Information Theory and Statiscs", Section 3.3.
- R. Vershynin, "High dimensional Probability", Section 5.

1 Brief Review

Last class we discussed concentration inequalities for the case when $X = (X_1, \dots, X_n)$ is a random vector with independent coordinates. We showed concentration results for the two cases when:

$$f(X_1, \dots, X_n) = \frac{1}{n}\sum_{i=1}^{n} \big(X_i - \mathbb{E}[X_i]\big),$$
$$f(X_1, \dots, X_n) = \|X_1, \dots, X_n\|_2 .$$

2 This Week's Goal

The main goal of today's class will be to extend the result to functions which are L-Lipschitz and separately convex.[1] In particular, we will be interested in showing the following:

Theorem 1. *Let $X_1, \dots, X_n$ be independent random variables each supported on $[a, b]$. Further let $f : \mathbb{R}^n \to \mathbb{R}$ be separately convex and L-Lipschitz. Then for all $t \geq 0$*

$$Pr\left[f(X_1, \dots, X_n) - \mathbb{E}\left[f(X_1, \dots, X_n)\right] \geq t\right] \leq \exp\left(-\frac{t^2}{4L^2(b-a)^2}\right).$$

We start by considering a few remarks of the theorem.

Remark

- A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be separately convex when the function $x_k \to f(x_1, \dots, x_k, \dots, x_n)$ is convex for fixed $(x_j : j \neq k)$.
- If f is convex then f is also separately convex.
- f is L-Lipschitz when $|f(x) - f(y)| \leq L \|x - y\|_2$ for any $x, y \in \mathbb{R}^n$.
- Theorem 1 is considered *dimension-free*: the concentration holds for a quantity independent of n.

For the proof we will use the *Entropy Method.* We need to show that for $X = (X_1, \dots, X_n)$ that the (upper tail of) random variable $f(X)$ is sub-Gaussian with parameter $\sigma^2 = 2L^2(b-a)^2$. Afterwards we can use the standard sub-Gaussian tail bound to obtain the result. To show that $f(X)$ is sub-Gaussian we will bound the moment generating function $\mathbb{E}\left[e^{\lambda f(X)}\right]$. The proof will follow in two main steps:

[1] *Reading:* Sec 3.1 of Wainwright book. Also relevant: Duchi notes Sec 3.3 and Vershynin HDP book Sec 5.

Step 1: Show the result for $n = 1$. This will be done with a Herbst argument, relating the MGF of a random variable to its entropy.

Step 2: Tensorize the result for general n.

3 Entropy

Before starting the proof we begin with some notation and preliminary lemmas. We will start by showing the case when $n = 1$.

Definition 1. *For a convex function $\phi : \mathbb{R} \to \mathbb{R}$, the **$\phi$-entropy** of Z is*

$$H_\phi(Z) = \mathbb{E}\left[\phi(Z)\right] - \phi(\mathbb{E}\left[Z\right]).$$

Remark

- By Jensen's inequality and the convexity of ϕ, we always have that $H_\phi(Z) \geq 0$.
- Specializing $\phi(u) = u^2$ you obtain that $H_\phi(Z) = \mathrm{Var}\left[Z\right]$.
- Taking $\phi(u) = -\log(u)$ then a straightforward calculation shows that $H_\phi(e^{\lambda X}) = \log(M_X(\lambda))$ where $M_X(\lambda) := \mathbb{E}\left[e^{\lambda X}\right]$ is the moment generating function. Thus, $H_\phi(e^{\lambda X})$ recovers the log moment generating function.

We will fix $\phi(u) = u \log(u)$ for the rest of the lecture and omit the subscript ϕ in the bottom of H_ϕ. After applying it to $e^{\lambda X}$ we have that

$$\begin{aligned}
H(e^{\lambda X}) &= \mathbb{E}\left[e^{\lambda X}\log(e^{\lambda X})\right] - \mathbb{E}\left[e^{\lambda X}\right]\log\left(\mathbb{E}\left[e^{\lambda} X\right]\right) \\
&= \mathbb{E}\left[\lambda X e^{\lambda X}\right] - \mathbb{E}\left[e^{\lambda X}\right]\log\left(\mathbb{E}\left[e^{\lambda} X\right]\right) \\
&= \lambda M_X'(\lambda) - M_X(\lambda)\log(M_X(\lambda)).
\end{aligned}$$

Notice that if we take $X \sim N(0, \sigma^2)$ then after plugging in the moment generating function we get that

$$H(e^{\lambda X}) = \frac{1}{2}\lambda^2\sigma^2 M_X(\lambda).$$

Lemma 1 (Herbst Argument)**.** *If $H(e^{\lambda X}) \leq \frac{1}{2}\lambda^2\sigma^2 M_X(\lambda)$ for all $\lambda \geq 0$ then $\mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right] \leq e^{\frac{1}{2}\lambda^2\sigma^2}$ for all $\lambda \geq 0$. In particular, X satisfies the sub-Gaussian MGF bound for $\lambda \geq 0$.*

Proof From before we have by assumption that

$$\begin{aligned}
H(e^{\lambda X}) &= \lambda M_X'(\lambda) - M_X(\lambda)\log(M_X(\lambda)) \\
&\leq \frac{1}{2}\lambda^2\sigma^2 M_X(\lambda).
\end{aligned}$$

Now define $G(\lambda) = \frac{1}{\lambda}\log(M_X(\lambda))$. Moreover, $\lim_{\lambda\to 0} G(\lambda) = \mathbb{E}\left[X\right]$ by the derivatives of the MGF yielding the moments. We thus extension to $G(0) = \mathbb{E}\left[X\right]$. Then by an application of the chain rule

$$G'(\lambda) = \frac{1}{\lambda}\frac{M_X'(\lambda)}{M_X(\lambda)} - \frac{\log(M_X(\lambda))}{\lambda^2}.$$

Rewriting the original inequality in terms of G gives that $G'(\lambda) \leq \frac{1}{2}\sigma^2$. This differential equation has a known solution that

$$G(\lambda) - G(0) \leq \frac{1}{2}\sigma^2\lambda.$$

Thus we find that

$$\begin{aligned} \frac{1}{\lambda}\log(M_X(\lambda)) - \mathbb{E}[X] &\le \frac{1}{2}\sigma^2\lambda \\ \Rightarrow \log(M_X(\lambda)) - \lambda\mathbb{E}[X] &\le \frac{1}{2}\sigma^2\lambda^2 \\ \Rightarrow \mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] &\le \frac{1}{2}\sigma^2\lambda^2 \end{aligned}$$

as needed. □

We next show the following result, which relates the entropy of $g(X)$ to its derivative and MGF.

Lemma 2. *If X is supported on $[a, b]$ and g is a convex function then*

$$H(e^{\lambda g(X)}) \le \frac{1}{2}\lambda^2(b-a)^2\mathbb{E}\left[g'(X)^2 e^{\lambda g(X)}\right].$$

Proof We use a symmetrization argument. Let Y be an independent copy of X. Then

$$\begin{aligned} H(e^{\lambda g(X)}) &= \mathbb{E}\left[\lambda g(X)e^{\lambda g(X)}\right] - \mathbb{E}\left[e^{\lambda g(X)}\right]\log\left(\mathbb{E}\left[e^{\lambda g(X)}\right]\right) \\ &= \mathbb{E}\left[\lambda g(X)e^{\lambda g(X)}\right] - \mathbb{E}\left[e^{\lambda g(X)}\right]\log\left(\mathbb{E}\left[e^{\lambda g(Y)}\right]\right) \\ &\le \mathbb{E}\left[\lambda g(X)e^{\lambda g(X)}\right] - \mathbb{E}\left[e^{\lambda g(X)}\lambda g(Y)\right] \text{ by Jensen's inequality and independence of } X \text{ and } Y \\ &= \mathbb{E}\left[\lambda g(X)e^{\lambda g(X)} - e^{\lambda g(X)}\lambda g(Y)\right] \\ &= \frac{1}{2}\mathbb{E}\left[(\lambda g(X) - \lambda g(Y))(e^{\lambda g(X)} - e^{\lambda g(Y)})\right] \\ &= \mathbb{E}\left[(\lambda g(X) - \lambda g(Y))(e^{\lambda g(X)} - e^{\lambda g(Y)})\mathbb{1}_{[g(X)\ge g(Y)]}\right]. \end{aligned}$$

The second to last line comes from the fact that X and Y are independent. The last line is from noticing that the terms on the inside are non-negative and symmetric, and so we can decompose the expectation into the two equal-sized portions from when $g(X) \ge g(Y)$ and other way around.

However, a simple fact shows that $e^s - e^t \le e^s(s-t)$. Rearranging this inequality shows that

$$(s-t)(e^s - e^t)\mathbb{1}_{[s\ge t]} \le e^s(s-t)^2\mathbb{1}_{[s\ge t]}.$$

We apply the above inequality where $s = \lambda g(X)$ and $t = \lambda g(Y)$ to get that

$$\begin{aligned} H(e^{\lambda g(X)}) &\le \mathbb{E}\left[\lambda^2(g(X) - g(Y))^2 e^{\lambda g(X)}\mathbb{1}_{[g(X)\ge g(Y)]}\right] \\ &= \lambda^2\mathbb{E}\left[(g(X) - g(Y))^2 e^{\lambda g(X)}\mathbb{1}_{[g(X)\ge g(Y)]}\right] \\ &\le \lambda^2\mathbb{E}\left[g'(X)^2(X-Y)^2 e^{\lambda g(X)}\mathbb{1}_{[g(X)\ge g(Y)]}\right] \\ &\le \frac{1}{2}\lambda^2(b-a)^2\mathbb{E}\left[g'(X)^2 e^{\lambda g(X)}\right]. \end{aligned}$$

In the second to last line we used the definition of the derivative of a convex function, and the last line that X and Y are supported in $[a, b]$ and a symmetrization argument again. □

Using these facts we are now ready to show Theorem 1 for the case when $n = 1$.

Proof As stated earlier, it suffices to show that $f(X) = f(X_1)$ is sub-Gaussian with parameter $\sigma^2 = 2L^2(b-a)^2$. However, by Lemma 2 since f is convex we know that

$$H(e^{\lambda f(X)}) \leq \frac{1}{2}\lambda^2(b-a)^2\mathbb{E}\left[f'(X)^2 e^{\lambda f(X)}\right].$$

As f is L-Lipschitz we know $\max_x |f'(x)| \leq L$ and so this can be bounded by $\frac{1}{2}\lambda^2 L^2(b-a)^2\mathbb{E}\left[e^{\lambda f(X)}\right]$. Thus we find that $H(e^{\lambda f(X)}) \leq \frac{1}{2}\lambda^2 L^2(b-a)^2 M_X(\lambda)$. By Lemma 1 this shows that $f(x)$ is sub-Gaussian with parameter $\sigma^2 = 2L^2(b-a)^2$. □

4 Tensorization

We now start to show the more general case by a tensorization argument. We start with some notation. For a vector $x \in \mathbb{R}^n$ set $x_{-k} = (x_i \mid i \neq k) \in \mathbb{R}^{n-1}$. For fixed x_{-k} define $f_k : \mathbb{R} \to \mathbb{R}$ by

$$f_k(x_k) = f(x_k, x_{-k}).$$

We define the conditional entropy for a random variable X_k as

$$H(e^{\lambda f_k(X_k)} \mid x_{-k}) = H(e^{\lambda f(X_k, x_{-k})}).$$

Notice here that the only randomness is X_k as x_{-k} is fixed.

Lemma 3 (Tensorization of Entropy). *If $X = (X_1, \ldots, X_n)$ has independent coordinates then*

$$H(e^{\lambda f(X)}) \leq \sum_{k=1}^{n} \mathbb{E}\left[H(e^{\lambda f_k(X_k)} \mid X_{-k})\right].$$

Before proving the Lemma, we will need the following claim,

Claim 1 (Variational Representation of Entropy).

$$H(e^{\lambda f(X)}) = \sup_g \left\{\mathbb{E}\left[g(X)e^{\lambda f(X)}\right] \mid \mathbb{E}\left[e^{g(X)}\right] \leq 1\right\}.$$

Proof We first show that the left hand side is upper bounded by the right hand side. Consider the function $g(x) = \lambda f(x) - \log\left(\mathbb{E}\left[e^{\lambda f(X)}\right]\right)$. Then

$$\begin{aligned} H(e^{\lambda f(X)}) &= \mathbb{E}\left[\lambda f(X)e^{\lambda f(X)}\right] - \mathbb{E}\left[e^{\lambda f(X)}\right]\log\left(\mathbb{E}\left[e^{\lambda f(X)}\right]\right) \\ &= \mathbb{E}\left[g(X)e^{\lambda f(X)}\right]. \end{aligned}$$

Noticing that $\mathbb{E}\left[e^{g(X)}\right] = 1$ the first inequality follows.

For the other direction consider the function $\Theta(u) = u\log(u) - u$. Then using the fact that e^y is the Fenchel-conjugate of Θ we have that

$$\Theta(u) = \sup_y\{uy - e^y\}.$$

However,

$$
\begin{aligned}
H(e^{\lambda f(X)}) &= \mathbb{E}\left[\Theta(e^{\lambda f(X)})\right] - \Theta\left(\mathbb{E}\left[e^{\lambda f(X)}\right]\right) \\
&= \mathbb{E}\left[\sup_y y e^{\lambda f(X)} - e^y\right] - \Theta\left(\mathbb{E}\left[e^{\lambda f(X)}\right]\right) \\
&= \sup_{\tilde{g}} \mathbb{E}\left[\tilde{g}(X) e^{\lambda f(X)} - e^{\tilde{g}(X)}\right] - \mathbb{E}\left[e^{\lambda f(X)}\right] \log\left(\mathbb{E}\left[e^{\lambda f(X)}\right]\right) + \mathbb{E}\left[e^{\lambda f(X)}\right] \\
&= \sup_{\tilde{g}} \mathbb{E}\left[\left(\tilde{g}(X) - \log\left(\mathbb{E}\left[e^{\lambda f(X)}\right]\right) e^{\lambda f(X)}\right] - \mathbb{E}\left[e^{\tilde{g}(X)}\right] + \mathbb{E}\left[e^{\lambda f(X)}\right] \\
&= \sup_g \mathbb{E}\left[g(X) e^{\lambda f(X)}\right] + \mathbb{E}\left[e^{\lambda f(X)}\right] (1 - \mathbb{E}\left[e^{\lambda g(X)}\right]) \\
&\geq \sup_g \left\{\mathbb{E}\left[g(X) e^{\lambda f(X)}\right] \mid \mathbb{E}\left[e^{g(X)}\right] \leq 1\right\},
\end{aligned}
$$

where in the second to last line we defined $g(x) = \tilde{g}(x) - \log\left(\mathbb{E}\left[e^{\lambda f(X)}\right]\right)$. □

We now complete the proof for Lemma 3.

Proof Let g be any function satisfying $\mathbb{E}\left[e^{g(X)}\right] \leq 1$. We also define X_j^n, and $g^k(X_k^n)$ as follows:

$$
\begin{aligned}
X_j^n &= (X_j, X_{j+1}, ..., X_n), \quad j = 1, \dots, n \\
g^k(X_k^n) &= \log \frac{\mathbb{E}\left[e^{g(X)} | X_k^n\right]}{\mathbb{E}\left[e^{g(X)} | X_{k+1}^n\right]}, \quad k = 1, \dots, n.
\end{aligned}
$$

Note that by construction, we get:

$$
\sum_{k=1}^n g^k(X_k^n) = g(X) - \log \mathbb{E}\left[e^{g(X)}\right] \geq g(X). \tag{1}
$$

We also have:

$$
\mathbb{E}\left[e^{g^k(X_k^n)} | X_{-k}\right] = \mathbb{E}\left[\frac{\mathbb{E}\left[e^{g(X)} | X_k^n\right]}{\mathbb{E}\left[e^{g(X)} | X_{k+1}^n\right]} | X_{-k}\right] = \frac{\mathbb{E}\left[e^{g(X)} | X_{k+1}^n\right]}{\mathbb{E}\left[e^{g(X)} | X_{k+1}^n\right]} = 1, \tag{2}
$$

where we used the fact that by independence $\mathbb{E}\left[\mathbb{E}\left[\cdot | X_k^n\right] | X_{-k}\right] = \mathbb{E}\left[\cdot | X_{k+1}^n\right] = \mathbb{E}\left[\mathbb{E}\left[\cdot | X_{k+1}^n\right] | X_{-k}\right]$ Combining this together we find that

$$
\begin{aligned}
\mathbb{E}\left[g(X) e^{\lambda f(x)}\right] &\leq \sum_{k=1}^n \mathbb{E}\left[g^k(X_k^n e^{\lambda f(X)}\right] && \text{by (1)} \\
&= \sum_{k=1}^n \mathbb{E}\left[\mathbb{E}\left[g^k(X_k^n e^{\lambda f(X)} | X_{-k}\right]\right] \\
&\leq \sum_{k=1}^n \mathbb{E}\left[H(e^{\lambda f(X)} | X_{-k})\right]. && \text{by (1) and Claim 1}
\end{aligned}
$$

Taking the supremum over g we conclude the proof:

$$
H(e^{\lambda f(X)}) \leq \sum_{k=1}^n \mathbb{E}\left[H(e^{\lambda f(X)} | X_{-k})\right].
$$

□

We can now finish our proof of Theorem 1.
Proof By Lemma 2:

$$H(e^{\lambda f(X)}|X_{-k}) \leq \lambda^2(b-a)^2\mathbb{E}\left[f_k'(X_k)^2 e^{\lambda f_k(X_k)} \mid X_{-k}\right].$$

By Lemma 3:

$$\begin{aligned} H(e^{\lambda f(X)}) &\leq \lambda^2(b-a)^2\mathbb{E}\left[\sum_{k=1}^{n} f_k'(X_k)^2 e^{\lambda f(X)}\right] \\ &= \lambda^2(b-a)^2\mathbb{E}\left[\|\nabla f(X)\|_2^2 e^{\lambda f(X)}\right] \\ &\leq \lambda^2(b-a)^2 L^2\mathbb{E}\left[e^{\lambda f(X)}\right]. \end{aligned}$$

Combining this result with Lemma 1 we conclude that $f(X)$ satisfies the $2L^2(b-a)^2$ sub-Gaussian upper tail bound as needed. □

While we used that f is separately convex to prove Theorem 1, if we impose the stronger assumption of convexity we can obtain the following two-sided inequality (note that this stronger assumption is required for a two-sided bound):

Theorem 2. *Let $X_1, \ldots, X_n$ be independent random variables each supported on $[a, b]$. Further let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and L-Lipschitz. Then $\forall t \geq 0$*

$$Pr\left[|f(X_1, \ldots, X_n) - \mathbb{E}\left[f(X_1, \ldots, X_n)\right]| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2L^2(b-a)^2}\right).$$

Note that the convexity assumption cannot be dropped in general; see Ledoux and Talagrand 1991, pp17.
Furthermore, if X_i are distributed normally, we no longer need the convexity assumption resulting in the following theorem:

Theorem 3. *Let $X_1, \ldots, X_n$ be independent random variables each distributed $\mathcal{N}(0, 1)$. Further let $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz. Then $\forall t \geq 0$*

$$Pr\left[|f(X_1, \ldots, X_n) - \mathbb{E}\left[f(X_1, \ldots, X_n)\right]| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2L^2}\right).$$

We can compare these results to the bounded difference (aka McDiarmid's) inequality.

Theorem 4 (Bounded Difference Inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables. Further let $f : \mathbb{R}^n \to \mathbb{R}$ satisfy the bounded difference property:*

$$|f(x_k, x_{-k}) - f(x_k', x_{-k})| \leq L_k \quad \textit{for all} \quad k, x_k, x_k', x_{-k}.$$

Then for all $t \geq 0$,

$$Pr\left[|f(X_1, \ldots, X_n) - \mathbb{E}\left[f(X_1, \ldots, X_n)\right]| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{\sum_{k=1}^{n} L_k^2}\right).$$

In many problems $\sum_{k=1}^{n} L_k^2 \gg L^2$, and thus Theorem 4 is much weaker than Theorems 1, 2, and 3.

5 Applications

We now turn our attention to some applications of these inequalities.

5.1 Concentration of Norm

Our first application is the concentration of norms of random vectors, which we have looked at in last lecture. Start by noting that norms are convex, and 1-Lipschitz by the triangle inequality:

$$| \|X\|_2 - \|Y\|_2 | \leq \|X - Y\|_2 .$$

Thus if the X_i's are bounded or Gaussian, by Theorem 2 or 3 we have $\|X\|_2 - \mathbb{E}[\|X\|_2]$ is $\mathcal{O}(1)$ sub-Gaussian. If the X_i's are bounded, the norm also satisfies the bounded difference property:

$$| \|x_1, \ldots, x_k, \ldots, x_n\|_2 - \|x_1, \ldots, x_k', \ldots, x_n\|_2 | \leq |x_k - x_k'| = \mathcal{O}(1).$$

Thus by using the bounded difference inequality, we get that the norm is $\mathcal{O}(n)$ sub-Gaussian — a much weaker result.

5.2 Max Singular Value

Next let's consider a random matrix $X \in \mathbb{R}^{n \times n}$, where $X_{i,j}$ is independently distributed and either bounded or Gaussian. We define the operator norm (the largest singular value) as follows:

$$\|X\|_{op} = \sigma_1(X) = \sup_{\|u\|_2 \leq 1, \|v\|_2 \leq 1} u^T X v.$$

Note that the operator norm is convex (maximum of affine function), and is 1-Lipschitz as:

$$| \|X\|_{op} - \|Y\|_{op} | \leq \|X - Y\|_{op} \leq \|X - Y\|_F.$$

Thus by Theorems 2 and 3, $\|X\|_{op} - \mathbb{E}[\|X\|_{op}]$ is $\mathcal{O}(1)$ sub-Gaussian.

5.3 Any Singular Value for a Gaussian Matrix

We now extend our approach to look at other singular values ($\sigma_k(X)$ where $k \geq 2$). Note that in this case, $\sigma_k(X)$ is no longer convex, so we restrict our analysis to the Gaussian case as it doesn't require convexity. However, $\sigma_k(X)$ is still 1-Lipschitz as we can see by using Weyl's Inequality:

$$|\sigma_k(X) - \sigma_k(Y)| \leq \|X - Y\|_{op} \leq \|X - Y\|_F.$$

Thus by Theorem 3 $\sigma_k(X) - \mathbb{E}[\sigma_k(X)]$ is $\mathcal{O}(1)$ sub-Gaussian.

5.4 Rademacher Complexity

Definition 2. *Let $A \subset \mathbb{R}^n$. The **Rademacher complexity** of A is*

$$R_n(A) = \mathbb{E}\left[\sup_{a \in A} \sum_{i=1}^{n} a_i \epsilon_i\right],$$

where $\epsilon_i \in \{-1, +1\}$ are i.i.d. Rademacher random variables. Similarly, let

$$\hat{R}_n(A) = \sup_{a \in A} \sum_{i=1}^{n} a_i \epsilon_i.$$

Note that $\hat{R}_n(A)$ is a convex function of ϵ with Lipschitz constant $W(A)$ as:

$$| \sup_{a \in A} \langle a, \epsilon \rangle - \sup_{a \in A} \langle a, \epsilon' \rangle | \leq | \sup_{a \in A} \langle a, \epsilon - \epsilon' \rangle | \leq \sup_{a \in A} \|a\|_2 \|\epsilon - \epsilon'\|_2 = W(A) \|\epsilon - \epsilon'\|_2.$$

Thus by theorem 2, we get:

$$\Pr\left[|\hat{R}_n(A) - R_n(A)| \geq t\right] \leq 2 \exp\left(\frac{-t^2}{8W(A)^2}\right).$$

6 Closing Remarks

Some final closing remarks on these concentration inequalities:
Remark

- You can apply Theorems 1, 2, and 3 to unbounded RVs by a truncation trick.
- Theorems 2 and 3 imply Hoeffding (as $\sum_i X_i$ is convex and $\sqrt{n}$-Lipschitz).
- There are "Bernstein" versions of these inequalities that account for variance.

This type of inequalities are also often used to bound the supremum of empirical processes:

$$f(x) = \sup_{g \in G} \frac{1}{n} \sum_{i=1}^{n} g(x_i).$$

In particular, we have the functional Hoeffding theorem:

Theorem 5 (Functional Hoeffding Theorem)**.** *If $X_i \in \mathcal{X}_i$ are independent, and for each $g \in G$:*

$$g(x_i) \in [a_{i,g}, b_{i,g}], \qquad \forall x_i \in \mathcal{X}_i.$$

Then:

$$Pr\left[f(x) - \mathbb{E}\left[f(x)\right] \geq t\right] \leq \exp\left(-\frac{nt^2}{4L^2}\right),$$

where $L^2 = \sup_{g \in G} \frac{1}{n} \sum_{i=1}^{n} (b_{i,g} - a_{i,g})^2$.

Note that if we used the bounded difference inequality we need $L^2 = \frac{1}{n} \sum_{i=1}^{n} \sup_{g \in G} (b_{i,g} - a_{i,g})^2$, which is often much weaker than the functional Hoeffding bound.