22.1 Sub-Gaussian Processes

We are often interested in bounding expressions of the form

$$E\left[ \sup_{\theta \in T} \theta^T \epsilon \right],$$

where $T \subset \mathbb{R}^n$, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ is a vector of independent $\text{SG}(\sigma^2)$ random variables.

**Example:** Suppose we have a class of the form $T = F(x_1^n) = \{ (f(x_1), \ldots, f(x_n)) : f \in F, x_i \in \mathbb{R}^d \}$ or $A(x_1^n) = \{ A \cap x_1^n, A \in A \}$, where $A$ is a collection of subsets of $\mathbb{R}^n$. 

- Suppose $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ is a vector of $n$ independent Rademacher random variables.
  Then $R_n(T) = E\left[ \sup_{\theta \in T} \theta^T \epsilon \right]$ is the Rademacher complexity of $T$ (or $F$).

- Suppose $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ is a vector of $n$ independent $\mathcal{N}(0,1)$ (or $\mathcal{N}(0,\sigma^2)$) random variables.
  Then $G_n(T) = E\left[ \sup_{\theta \in T} \theta^T \epsilon \right]$ is the Gaussian complexity of $T$ (or $F$).

**Remark:** Rademacher and Gaussian complexities are sometimes of a similar order and sometimes of different orders.

If $T = \{ \theta \in \mathbb{R}^d : ||\theta||_2 \leq 1 \}$, then $R_n(T) \approx G_n(T) \leq \sqrt{d}$.

If $T = \{ \theta \in \mathbb{R}^d : ||\theta||_1 \leq 1 \}$, then $R_n(T) = 1$ and $G_n(T) \leq \sqrt{\log d}$.

To show these results, we would use the facts $||x||_2 = \sup_{v : ||v||_2 \leq 1} v^T x = \sup_{v : ||v||_1 \leq 1} v^T x$ and $||x||_1 = \sup_{v : ||v||_\infty \leq 1} v^T x$.

Let $\{X_\theta, \theta \in T\}$ be a mean zero stochastic process indexed by $T$. Similar to above, we may be interested in expression of the form $E\left[ \sup_{\theta \in T} X_\theta \right]$.

**Examples:**

1) Rademacher and Gaussian complexities. In the examples above, we could represent $X_\theta = \epsilon^T \theta$, where we are interested in $E\left[ \sup_{\theta \in T} X_\theta \right]$.

2) Non-parametric least-squares regression. We observe $n$ pairs $(Y_1, x_1), \ldots, (Y_n, x_n)$ where $x_1, \ldots, x_n$ are deterministic points in $[0,1]$. We assume that $Y_i = f^*(x_i) + \epsilon_i$, where $(\epsilon_1, \ldots, \epsilon_n) \sim \text{SG}(\sigma^2)$ and...
\( f^* \in \mathcal{F} \) (a class of real-valued functions on \([0,1]\)). Let \( \hat{f} \in \text{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum (Y_i - f(x_i))^2 \) be the least squares estimator. We want to bound \( \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} (\hat{f}(x_i) - f^*(x_i))^2 \right] \).

Small \( \text{MSE} \) means that \( \hat{f} \) is a good approximation to \( f^* \). To analyze the performance of \( \hat{f} \), we start with the basic inequality:

\[
\text{MSE} \leq \frac{2}{n} \sum_{i=1}^{n} \epsilon_i \left( \hat{f}(x_i) - f^*(x_i) \right)
\]

\[
\leq \frac{2}{\sqrt{n}} \sup_{f,g \in \mathcal{F}} |X_f - X_g|
\]

where \( X_f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i f(x_i) \). So

\[
\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} \left( \hat{f}(x_i) - f^*(x_i) \right)^2 \right] \leq \frac{2}{\sqrt{n}} \mathbb{E} \left[ \sup_{f,g \in \mathcal{F}} |X_f - X_g| \right].
\]

Also, for any two functions \( f, g \in \mathcal{F} \),

\[
V(X_f - X_g) = \mathbb{E} \left[ \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i (f(x_i) - g(x_i)) \right)^2 \right]
\]

\[
= \frac{1}{n} \mathbb{E} \left[ \left( \sum_{i=1}^{n} \epsilon_i (f(x_i) - g(x_i)) \right)^2 \right]
\]

\[
\leq \frac{1}{n} \mathbb{E} \left[ \sum_{i=1}^{n} \epsilon_i^2 \left( \sum_{i=1}^{n} (f(x_i) - g(x_i))^2 \right) \right]
\]

\[
= \frac{1}{n} \mathbb{E} \left[ \sum_{i=1}^{n} \epsilon_i^2 \right] \sum_{i=1}^{n} (f(x_i) - g(x_i))^2
\]

\[
\leq \frac{1}{n} \cdot n \sigma^2 \|f - g\|_2^2
\]

\[
= \sigma^2 \|f - g\|_2^2.
\]

3) Estimation in Wasserstein distance. Essentially, the Wasserstein distance is the amount of mass one must move from one distribution to another to make them equal.

Suppose \( P \) and \( Q \) are distributions on \( \mathbb{R} \). The Wasserstein distance between \( P \) and \( Q \) is

\[
W_1(P, Q) = \sup_{f \in \mathcal{F}} |Pf - Qf|,
\]

where \( Pf = \mathbb{E}_{X \sim P}[f(X)] \) and \( \mathcal{F} = \{ f : [0,1] \to \mathbb{R}, f \text{ is 1-Lipschitz} \} \). (That is, for any \( f \in \mathcal{F} \) and \( x, y \in [0,1] \), \( |f(x) - f(y)| \leq |x - y| \).)

An equivalent characterization is given by

\[
W_1(P, Q) = \inf_{(x,y)} \mathbb{E}[|X - Y|, X \sim P, Y \sim Q].
\]
We might want to use this metric to compare a true distribution to its empirical distribution. Suppose \((X_1, \ldots, X_n) \simid P\) and \(P_n\) is the corresponding empirical measure. Then \(W_1(P_n, P) = \sup_{f \in F} |X_f|\), where \(X_f = P_n f - Pf\). So \(E[X_f] = 0\) for all \(f\). Then we see
\[
E[W_1(P_n, P)] = E\left[ \sup_{f \in F} |X_f| \right].
\]

### 22.1.1 Sub-Gaussian Processes

**Definition 22.1 (Sub-Gaussian process.)** A zero-mean stochastic process \(X_\theta : \theta \in \mathbb{T}\) is sub-Gaussian with respect to metric \(d\) on \(\mathbb{T}\) if for \(\theta, \theta' \in \mathbb{T}\) and \(\lambda \in \mathbb{R}\),
\[
\mathbb{E}\left[ e^{\lambda (X_\theta - X_{\theta'})} \right] \leq \exp \left[ \frac{\lambda^2}{2} d^2(\theta, \theta') \right].
\]

Equivalently, for \(\theta, \theta' \in \mathbb{T}\), \(X_\theta - X_{\theta'} \in \text{SG}(d^2(\theta, \theta'))\). \(d(\cdot)\) is called the canonical metric. In the case of Gaussian random variables, the canonical metric is given by \(d(\theta, \theta') = \sqrt{V(X_\theta - X_{\theta'})}\).

By Hoeffding’s inequality for sub-Gaussians,
\[
\mathbb{P}(|X_\theta - X_{\theta'}| \geq t) \leq 2 \exp \left\{ - \frac{t^2}{2d^2(\theta, \theta')} \right\}.
\]

**Examples:**

1) Rademacher and Gaussian complexities. In these cases, \(\mathbb{T} \subseteq \mathbb{R}^n\). These processes are sub-Gaussian with respect to \(d(\theta, \theta') = ||\theta - \theta'||_2\) on \(\mathbb{T}\) because
\[
V(X_\theta - X_{\theta'}) = V(\epsilon^T \theta - \epsilon^T \theta') \leq ||\theta - \theta'||_2 \sigma^2.
\]

(This proves that \(X_\theta - X_{\theta'} \in \text{SG}(||\theta - \theta'||_2 \sigma^2)\).) In the case where \(\epsilon_i\) is Rademacher, \(\sigma^2 = 1\). In the case where \(\epsilon_i \sim N(0, \sigma^2)\), \(\sigma^2\) in equation 22.1 is the same \(\sigma^2\) as the normal variance.

2) Non-parametric least squares regression. As before, we define \(X_f = \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f(x_i)\), where \(x_1, \ldots, x_n\) are deterministic. \(X_f\) is SG, and so is \(X_f - X_g\). Previously, we showed \(V(X_f - X_g) \leq \sigma^2 ||f - g||_2^2\). In this case, we can use the canonical distance \(d(f, g) = \sqrt{\frac{1}{n} \sum_{i=1}^n (f(x_i) - g(x_i))^2}\).

3) Wasserstein distance. In this problem, we had \(X_f = P_n f - Pf\). This is an SG process with respect to \(d(f, g) = \frac{||f - g||_1}{\sqrt{n}}\). This is an exercise, and the result can be obtained by using Azuma-Hoeffding.

### 22.1.2 Metric Entropy

**Definition 22.2 (Metric entropy.)** Let \(\mathbb{T} \subseteq \mathbb{R}^n\) and let \(d\) be a distance metric on \(\mathbb{T}\). For \(\delta > 0\), the metric entropy of \(\mathbb{T}\) with respect to \(d\) is given by \(\log N(\mathbb{T}, \delta)\), where \(N(\mathbb{T}, \delta)\) is the \(\delta\)-covering number of \(\mathbb{T}\).

**Definition 22.3 (Diameter of \(\mathbb{T}\).)** Let \(\mathbb{T} \subseteq \mathbb{R}^n\) and let \(d\) be a distance metric on \(\mathbb{T}\). The diameter of the set \(\mathbb{T}\) is given by \(D = \sup_{\theta, \theta' \in \mathbb{T}} d(\theta, \theta')\).
Proposition 22.4 (1-step discretization bound.) Assume \( \{X_\theta : \theta \in \mathbb{T}\} \) is a SG process with respect to \( d \). Then for all \( \delta \in (0, D] \),

\[
E \left[ \sup_{\theta, \theta' \in \mathbb{T}} |X_\theta - X_{\theta'}| \right] \leq 2E \left[ \sup_{\gamma, \gamma' \in \mathbb{T}} |X_\gamma - X_{\gamma'}| \right] + 4D \sqrt{\log N(\mathbb{T}, \delta)}.
\]

\( \delta \) is a tuning parameter. As \( \delta \) decreases, the first term decreases and the second term increases.

Remarks:

1) For arbitrary \( \theta_0 \in \mathbb{T} \),
\[
E \left[ \sup_{\theta \in \mathbb{T}} X_\theta \right] = E \left[ \sup_{\theta \in \mathbb{T}} (X_\theta - X_{\theta_0}) \right] \leq E \left[ \sup_{\theta, \theta' \in \mathbb{T}} (X_\theta - X_{\theta'}) \right].
\]

2) Constants are not optimal.

Proof: Let \( \theta_1, \ldots, \theta_N \) be a minimal \( \delta \)-cover of \( \mathbb{T} \), where \( N = N(\mathbb{T}, \delta) \). Then for all \( \theta \in \mathbb{T} \), there exists \( j \) \( (1 \leq j \leq N) \) such that \( d(\theta, \theta_j) \leq \delta \).

Fix \( \theta \in \mathbb{T} \). Choose \( j \) such that \( d(\theta, \theta_j) \leq \delta \). Then

\[
X_\theta - X_{\theta_1} = X_\theta - X_{\theta_j} + X_{\theta_j} - X_{\theta_1} \leq \sup_{\gamma, \gamma' \in \mathbb{T}} (X_\gamma - X_{\gamma'}) + \max_i |X_{\theta_j} - X_{\theta_1}|.
\]

We can obtain a similar bound for \( X_{\theta_j} - X_{\theta'} \), where \( \theta' \) is another point in \( \mathbb{T} \).

Adding up and using the fact that \( \theta \) and \( \theta' \) are arbitrary,

\[
\sup_{\theta, \theta' \in \mathbb{T}} (X_\theta - X_{\theta'}) \leq 2 \sup_{\gamma, \gamma' \in \mathbb{T}} (X_\gamma - X_{\gamma'}) + 2 \max_i |X_{\theta_j} - X_{\theta_1}|.
\]

To finish the proof, we will take the expectation of both sides. Since \( X_{\theta_j} - X_{\theta_1} \in SG(D^2) \), we know
\[
E[\max_i |X_{\theta_j} - X_{\theta_1}|] \leq 2D \sqrt{\log N(\mathbb{T}, \delta)}.
\]
(See the maximal inequality from the 9-13 lecture notes.) So

\[
E \left[ \sup_{\theta, \theta' \in \mathbb{T}} |X_\theta - X_{\theta'}| \right] \leq 2E \left[ \sup_{\gamma, \gamma' \in \mathbb{T}} |X_\gamma - X_{\gamma'}| \right] + 2E \left[ \max_i |X_{\theta_j} - X_{\theta_1}| \right] \leq 2E \left[ \sup_{\gamma, \gamma' \in \mathbb{T}} |X_\gamma - X_{\gamma'}| \right] + 4D \sqrt{\log N(\mathbb{T}, \delta)}.
\]

Applications: For \( \mathbb{T} \subseteq \mathbb{R}^n \) and \( \delta \in (0, D] \) (where \( D \) is the diameter of \( \mathbb{T} \)), define

\[
\mathbb{T}(\delta) := \{ \gamma - \gamma' : \gamma, \gamma' \in \mathbb{T}, \|\gamma - \gamma'\|_2 \leq \delta \}.
\]

Then where \( \epsilon \) is a vector of Rademacher random variables and \( d(\cdot) \) is Euclidean distance,

\[
\mathcal{R}_n(\mathbb{T}(\delta)) = E \left[ \sup_{\gamma, \gamma' \in \mathbb{T}} \epsilon^T (\gamma - \gamma') \right] \leq E[\|\epsilon\|_2] \leq \sqrt{n} \delta.
\]
The same inequality holds for $G_n(\tilde{T}(\delta))$ if $\epsilon$ is a vector of $N(0,1)$ random variables.

Applying the 1-step discretization bound, we see

$$E \left[ \sup_{\theta, \theta' \in \mathcal{T}} \epsilon^T (\theta - \theta') \right] \lesssim \min_{\delta \in (0,D]} \{ \delta \sqrt{n} + \sqrt{\log N(\mathcal{T}, \delta)} \}$$

(up to constants). Again, as $\delta \to 0$, $\delta \sqrt{n} \to 0$ and $\sqrt{\log N(\mathcal{T}, \delta)}$ increases (often to infinity). To balance, we set $\delta \sqrt{n} = \sqrt{\log N(\mathcal{T}, \delta)}$ and solve for $\delta$.

References