Theoretical Statistics. Lecture 14.

Peter Bartlett

Metric entropy.

1. Chaining: Dudley's entropy integral

Recall: Sub-Gaussian processes

Definition: A stochastic process $\theta \mapsto X_\theta$ with indexing set T is sub-Gaussian with respect to a metric d on T if, for all $\theta, \theta' \in T$ and all $\lambda \in \mathbb{R}$,

$$\mathbf{E} \exp\left(\lambda(X_\theta - X_{\theta'})\right) \leq \exp\left(\frac{\lambda^2 d(\theta, \theta')^2}{2}\right).$$

Lemma: [Finite Classes] For X_θ sub-Gaussian wrt d on T, and A a set of pairs from T,

$$\mathbf{E} \max_{(\theta,\theta') \in A} (X_\theta - X_{\theta'}) \leq \max_{(\theta,\theta') \in A} d(\theta, \theta') \sqrt{2 \log |A|}.$$

Recall: Covering number bound

Theorem: Consider a zero-mean process X_θ that is sub-Gaussian wrt the metric d on T. Suppose that the diameter of T is $D = \sup_{\theta,\theta'} d(\theta, \theta')$. Then for any ϵ,

$$\mathbf{E} \sup_\theta X_\theta \le 2\mathbf{E} \sup_{d(\theta,\theta')\le\epsilon} (X_\theta - X_{\theta'}) + 2D\sqrt{\log N(\epsilon, T, d)}.$$

Dudley's entropy integral

Theorem: Let X_θ be a zero-mean stochastic process that is sub-Gaussian wrt a pseudo-metric d on the indexing set T. Then

$$\mathbf{E} \sup_\theta X_\theta \leq 8\sqrt{2} \int_0^\infty \sqrt{\log N(\epsilon, T, d)}\, d\epsilon.$$

Note that we can always rewrite the integral as an integral from 0 to the diameter of T.

Dudley's entropy integral: Proof

As before,

$$\mathbf{E} \sup_{\theta} X_\theta = \mathbf{E} \sup_{\theta} (X_\theta - X_{\theta'}) \leq \mathbf{E} \sup_{\theta, \theta'} (X_\theta - X_{\theta'}),$$

and choosing $\hat{\theta} \in \hat{T}$ (a minimal ϵ-cover) with $d(\hat{\theta}, \theta) \leq \epsilon$ (and similarly for θ'), we have

$$\begin{aligned} X_\theta - X_{\theta'} &= X_\theta - X_{\hat{\theta}} + X_{\hat{\theta}} - X_{\hat{\theta}'} + X_{\hat{\theta}'} - X_{\theta'} \\ &\leq 2 \sup_{d(\theta, \hat{\theta}) \leq \epsilon} (X_\theta - X_{\hat{\theta}}) + \sup_{\hat{\theta}, \hat{\theta}' \in \hat{T}} X_{\hat{\theta}} - X_{\hat{\theta}'}. \end{aligned}$$

Dudley's entropy integral: Proof

Consider bounding $\mathbf{E}\sup_{\hat{\theta},\hat{\theta}'}(X_{\hat{\theta}} - X_{\hat{\theta}'})$. Previously, we bounded the supremum over the ϵ-cover $\hat{T}$ (for which the diameter is that of T). Instead, we consider a sequence of progressively better approximations to elements of $\hat{T}$ (which leads to sets with progressively smaller diameters). Suppose the diameter of $\hat{T}$ is D. We first define $\hat{T}_k = \hat{T}$, and think of it as a $(2^{-k}D)$-cover of $\hat{T}$, where $k = \lceil \log_2(D/\epsilon) \rceil$ ensures that $2^{-k}D \leq \epsilon$. Then we define $\hat{T}_{i-1} =$ a minimal $(2^{-(i-1)}D)$-cover of $\hat{T}_i$, for i going from $k-1$ down to 0. Notice that $\hat{T}_0$ is a minimal D-cover of $\hat{T}_1$, so $|\hat{T}_0| = 1$.

[PICTURE].

Dudley's entropy integral: Proof

Pick $\hat{\theta}_k = \hat{\theta}$, and then pick $\hat{\theta}_{i-1} \in \hat{T}_{i-1}$ as the best approximation of $\hat{\theta}_i$. We can write $\hat{\theta}_{i-1} = f_{i-1}(\hat{\theta}_i)$, where $f_{i-1} : \hat{T}_i \to \hat{T}_{i-1}$ is the best approximation operator.

Then we can write

$$X_{\hat{\theta}} = X_{\hat{\theta}_k} = X_{\hat{\theta}_0} + \sum_{i=1}^{k} \left(X_{\hat{\theta}_i} - X_{\hat{\theta}_{i-1}} \right)$$

and, using the same notation for $\hat{\theta}'$, we have

$$\begin{aligned} X_{\hat{\theta}} - X_{\hat{\theta}'} &= X_{\hat{\theta}_k} - X_{\hat{\theta}'_k} \\ &= \sum_{i=1}^{k} \left(X_{\hat{\theta}_i} - X_{\hat{\theta}_{i-1}} \right) - \sum_{i=1}^{k} \left(X_{\hat{\theta}'_i} - X_{\hat{\theta}'_{i-1}} \right). \end{aligned}$$

Dudley's entropy integral: Proof

Thus,

$$\mathbf{E} \sup_{\hat{\theta}, \hat{\theta}' \in \hat{T}} X_{\hat{\theta}} - X_{\hat{\theta}'} \leq 2 \sum_{i=1}^{k} \mathbf{E} \sup_{\hat{\theta}_i \in \hat{T}_i} \left(X_{\hat{\theta}_i} - X_{f_{i-1}(\hat{\theta}_i)} \right).$$

Since $d(\hat{\theta}_i, \hat{\theta}_{i-1}) \leq 2^{-(i-1)} D$, the Finite Lemma shows that

$$\begin{aligned} \mathbf{E} \sup_{\hat{\theta}_i \in \hat{T}_i} \left(X_{\hat{\theta}_i} - X_{f_{i-1}(\hat{\theta}_i)} \right) &\leq 2^{-(i-1)} D \sqrt{2 \log |\hat{T}_i|} \\ &\leq 2^{-(i-1)} D \sqrt{2 \log N(2^{-i} D, T)}. \end{aligned}$$

Dudley's entropy integral: Proof

Finally, since $\log N(2^{-i}D) \leq \log N(u)$ for $u \leq 2^{-i}D$, we can approximate the area of the rectangle from $(2^{-(i+1)}D, 0)$ to $(2^{-i}D, \sqrt{2\log N(2^{-i}D)})$ by the integral under $\sqrt{2\log N(u)}$ for u in that interval (which has length $2^{-(i+1)}D$):

$$\begin{aligned} 2^{-(i-1)}D\sqrt{2\log N(2^{-i}D)} &= 4 \times 2^{-(i+1)}D\sqrt{2\log N(2^{-i}D)} \\ &\leq 4\int_{2^{-(i+1)}D}^{2^{-i}D} \sqrt{2\log N(u,T)}\,du. \end{aligned}$$

Dudley's entropy integral: Proof

Combining, we have

$$
\begin{aligned}
\mathbf{E}\sup_{\theta} X_\theta &\le 2\mathbf{E}\sup_{d(\theta,\hat\theta)\le\epsilon}(X_\theta - X_{\hat\theta}) + 2\sum_{i=1}^{k}\mathbf{E}\sup_{\hat\theta_i\in\hat T_i}\left(X_{\hat\theta_i} - X_{f_{i-1}(\hat\theta_i)}\right)\\
&\le 2\mathbf{E}\sup_{d(\theta,\hat\theta)\le\epsilon}(X_\theta - X_{\hat\theta}) + 2\sum_{i=1}^{k}2^{-(i-1)}D\sqrt{2\log N(2^{-i}D,T)}\\
&\le 2\mathbf{E}\sup_{d(\theta,\hat\theta)\le\epsilon}(X_\theta - X_{\hat\theta}) + 8\sqrt{2}\int_{2^{-(k+1)}D}^{D/2}\sqrt{\log N(u,T)}du.
\end{aligned}
$$

When $\epsilon \to 0$, the first term goes to zero and (since $k = \lceil \log_2(D/\epsilon)\rceil$), the second term approaches the integral from 0 to $D/2$, which gives the result.

Dudley's entropy integral

We actually proved the following result:

Theorem: Let X_θ be a zero-mean stochastic process that is sub-Gaussian wrt a pseudo-metric d on the indexing set T. Then

$$\mathbf{E} \sup_{\theta} X_\theta \leq 2\mathbf{E} \sup_{d(\theta,\theta')\leq\delta} (X_\theta - X_{\theta'}) + 8\sqrt{2} \int_{\delta/2}^{D/2} \sqrt{\log N(\epsilon, T, d)}\, d\epsilon.$$

When the entropy integral does not exist (because $N(\epsilon, T, d)$ grows too quickly as $\epsilon \to 0$), this can still give a useful bound.

Dudley's entropy integral

When does the entropy integral exist? Suppose T has diameter D and $\log N(\epsilon, T, d) = O(\epsilon^{-d})$. Then

$$\int_0^D \sqrt{\log N(\epsilon, T, d)}\, d\epsilon \leq C \int_0^D \epsilon^{-d/2}\, d\epsilon$$
$$= \frac{C}{1 - d/2} D^{1-d/2}$$

provided that $d < 2$. The integral does not exist otherwise.

Entropy Integral: Lipschitz parameterized class

Suppose that F is a parameterized class, $F = \{f(\theta, \cdot) : \theta \in \Theta\}$, where $\Theta = B_2 \subset \mathbb{R}^p$. The parameterization is L-Lipschitz wrt Euclidean distance on Θ, so that for all x,

$$|f(\theta, x) - f(\theta', x)| \leq L\|\theta - \theta'\|_2.$$

Suppose also that $F = -F$ (that is, F is closed under negations).

Theorem:

$$\mathbf{E}\|R_n\|_F = O\left(L\sqrt{\frac{p}{n}}\right).$$

NB: We've lost the log factor.

Entropy Integral: Lipschitz parameterized class

Recall that

$$n\mathbf{E}\|R_n\|_F = \mathbf{E}\sup_{F\cup -F}\langle\epsilon,\cdot\rangle = \mathbf{E}\sup_{F}\langle\epsilon,\cdot\rangle = \mathbf{E}\sup_{\theta}\langle\epsilon, f(\theta, X_1^n)\rangle,$$

which is sub-Gaussian wrt the Euclidean distance on $\mathbb{R}^n$. Also, recall that

$$N(\delta, f(\Theta, X_1^n), \|\cdot\|_2) \le N(\delta/(L\sqrt{n}), \Theta, \|\cdot\|_2) \le (1 + 2L\sqrt{n}/\delta)^p.$$

Entropy Integral: Lipschitz parameterized class

Hence,

$$
\begin{aligned}
\mathbf{E}\|R_n\|_F &\le \frac{8\sqrt{2}}{n}\int_0^\infty \sqrt{\log N\left(\frac{\epsilon}{L\sqrt{n}}, \Theta, \|\cdot\|_2\right)}\, d\epsilon \\
&= \frac{8\sqrt{2}L}{\sqrt{n}}\int_0^\infty \sqrt{\log N(\epsilon, \Theta, \|\cdot\|_2)}\, d\epsilon \\
&\le 8\sqrt{2}L\sqrt{\frac{p}{n}}\int_0^2 \sqrt{\log\left(1+\frac{2}{\epsilon}\right)}\, d\epsilon \\
&\le 8\sqrt{2}L\sqrt{\frac{p}{n}}\int_0^2 \sqrt{\log\left(\frac{4}{\epsilon}\right)}\, d\epsilon.
\end{aligned}
$$

Entropy Integral: Lipschitz parameterized class

Integrating by parts,

$$
\begin{aligned}
\mathbf{E}\|R_n\|_F &\le 8\sqrt{2}L\sqrt{\frac{p}{n}}\int_0^2 \sqrt{\log\left(\frac{4}{\epsilon}\right)}\,d\epsilon \\
&= 8\sqrt{2}L\sqrt{\frac{p}{n}}\left(\left[4e^{-y^2}y\right]_\infty^{\sqrt{\log 2}} - 4\int_\infty^{\sqrt{\log 2}} e^{-y^2}\,dy\right) \\
&\le 16\sqrt{2}\left(\sqrt{\log 2}+\sqrt{2\pi}\right)L\sqrt{\frac{p}{n}} \\
&< L\sqrt{\frac{8.7p}{n}}.
\end{aligned}
$$

Entropy Integral: VC-class

Theorem: For F a class of $\{0,1\}$-valued functions with VC-dimension d,

$$\mathbf{E}\|R_n\|_F = O\left(\sqrt{\frac{d}{n}}\right).$$

Compare with the consequence of Sauer's Lemma: $O(\sqrt{d\log(n/d)/n})$. We lose the log factor.

Note: This leads to a faster rate (without the log factor) in the proof of the Glivenko-Cantelli Theorem:

$$\Pr\left(\|F_n - F\|_\infty \geq \frac{c}{\sqrt{n}} + t\right) \leq 2\exp\left(-\frac{nt^2}{8}\right).$$

Entropy Integral: VC-class

We have

$$\begin{aligned}
\mathbf{E}\|R_n\|_F &\leq \frac{8\sqrt{2}}{n}\mathbf{E}\int_0^{2\sqrt{n}} \sqrt{\log N(\epsilon, F(X_1^n), \|\cdot\|_2)}\, d\epsilon \\
&\leq \frac{8\sqrt{2}}{n}\mathbf{E}\int_0^{2\sqrt{n}} \sqrt{\log N(\epsilon/\sqrt{n}, F, \|\cdot\|_{L_2(P_n)})}\, d\epsilon \\
&= \frac{8\sqrt{2}}{\sqrt{n}}\mathbf{E}\int_0^{2} \sqrt{\log N(\epsilon, F, \|\cdot\|_{L_2(P_n)})}\, d\epsilon,
\end{aligned}$$

where

$$\|f-g\|_{L_2(P_n)}^2 = \frac{1}{n}\sum_{i=1}^{n}(f(X_i)-g(X_i))^2.$$

Entropy Integral: VC-class

Fact (due to Haussler):

$$N(\epsilon, F, \|\cdot\|_{L_2(P_n)}) \leq cd(16e)^d \epsilon^{-2d}.$$

$$\begin{aligned}
\mathbf{E}\|R_n\|_F &\leq \frac{8\sqrt{2}}{n}\mathbf{E}\int_0^2 \sqrt{\log N(\epsilon, F, \|\cdot\|_{L_2(P_n)})}\, d\epsilon \\
&\leq \frac{8\sqrt{2}}{n}\mathbf{E}\int_0^2 \sqrt{\log\left(cd(16e)^d \epsilon^{-2d}\right)}\, d\epsilon \\
&= \cdots \\
&\leq c\sqrt{\frac{d}{n}}.
\end{aligned}$$

An aside: Generic Chaining

Theorem: Let X_θ be a zero-mean stochastic process that is sub-Gaussian wrt a pseudo-metric d on the indexing set T. Then for any probability distribution μ on T,

$$\mathbf{E} \sup_{\theta} X_\theta \leq c \sup_{\theta \in T} \int_0^\infty \sqrt{\log \frac{1}{\mu(B(\theta, \epsilon))}} \, d\epsilon.$$

An aside: Generic Chaining

Talagrand's γ_2:

Theorem: For X_θ as above and

$$\gamma_2(T,d) = \inf_{\mu} \sup_{\theta \in T} \int_0^\infty \sqrt{\log \frac{1}{\mu(B(\theta,\epsilon))}}\, d\epsilon,$$

we have

$$\mathbf{E} \sup_{\theta} X_\theta \le c\gamma_2(T,d).$$

Sudakov's Lower Bound

Theorem: For a zero-mean Gaussian process X_θ defined on T, define the variance pseudometric $d(\theta, \theta')^2 = \mathrm{Var}(X_\theta - X_{\theta'})$. Then

$$\mathbf{E} \sup_\theta X_\theta \geq \sup_{\epsilon > 0} \frac{\epsilon}{2} \sqrt{\log M(\epsilon, T, d)}.$$

Sudakov's Lower Bound

Compare with the Entropy integral:

Theorem: Let X_θ be a zero-mean stochastic process that is sub-Gaussian wrt a pseudo-metric d on the indexing set T. Then

$$\mathbf{E} \sup_\theta X_\theta \leq 8\sqrt{2} \int_0^\infty \sqrt{\log N(\epsilon, T, d)}\, d\epsilon.$$

Suppose that $\mathrm{Var}(X_\theta - X_{\theta'})$ is on the same scale as $d(\theta, \theta')^2$ (think of the Gaussian example of a sub-Gaussian process—this is precisely the variance). Then, modulo constants, the lower bound is the area of the largest rectangle that can fit under the curve $(\epsilon, \sqrt{\log N(\epsilon)})$, whereas the upper bound is the area under the curve.