Lecture 28 — October 25

Instructor: Stas Minsker *Scribe: Mose Wintner*

28.1 Talagrand's Contraction Inequality

Last time:

$$\mathbb{E}\phi\left(\sup_{t\in T}\sum_{j=1}^{n}\varepsilon_j\varphi(t_j)\right)\le \mathbb{E}\phi\left(\sup_{t\in T}\sum_{j=1}^{n}\varepsilon_j t_j\right).$$

Need: absolute values. To finish the proof, note that $|a| = a_+ + (-a)_+$, $a \mapsto a_+$ is convex, and

$$\begin{aligned}
\mathbb{E}\phi\left(\frac{1}{2}\sup_{t\in T}\left|\sum_{j=1}^{n}\varepsilon_j\varphi(t_j)\right|\right) &= \mathbb{E}\phi\left(\frac{1}{2}\sup_{t\in T}\sum\varepsilon_j\varphi(t_j))_+ + \frac{1}{2}\left(\sup_t\sum(-\varepsilon_j)\varphi(t_j)\right)_+\right)\\
&\le \frac{1}{2}\mathbb{E}\phi\left(\sup_{t\in T}\sum\varepsilon_j\varphi(t_j))_+\right) + \frac{1}{2}\mathbb{E}\phi\left(\sup_t\sum(-\varepsilon_j)\varphi(t_j))_+\right)\\
&= \mathbb{E}\phi(\sup\sum\varepsilon_j\varphi(t_j))_+
\end{aligned}$$

Since $a \mapsto \phi(a_+)$ is convex, apply what we proved last time to this map to get

$$\mathbb{E}\phi\left(\frac{1}{2}\sum\varepsilon_j\varphi(t_j)\right)\le \mathbb{E}\phi\left(\sup_t(\sum\varepsilon_j\varphi(t_j))_+\right)\le \mathbb{E}\phi\left(\sup_t\left|\sum\varepsilon_j t_j\right|\right).$$

This concludes the proof. Now let $\mathcal{F} = \{f : S \to \mathbb{R}\}$ be a class of functions and let $X_1, \dots, X_n \in S$ be iid. Define $Y_n(f) = \sum f(X_j)$, $||Y_n||_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |Y_n(f)|$.

Theorem 1. Talagrand's inequality. Assume that $|f| \le U$ for all $f \in \mathcal{F}$ and let

$$V = \mathbb{E}\sup_{f\in\mathcal{F}}\sum_{j=1}^{n} f^2(X_j).$$

Then there exist absolute constants $K_1, K_2 > 0$ such that

$$P\left(|Y_n(f) - \mathbb{E}Y_n(f)| \ge t\right) \le K_1 \exp\left(-\frac{t}{K_2 v}\log\left(1+\frac{tU}{V}\right)\right).$$

Note that $tU/V << 1$ iff $t << V/U$. Then the tail behaves like $\exp(-t^2/cV)$, a Gaussian. If tU/V isn't small then the tail behaves approximately like $\exp(-t/cV)$.

Example 1. $\mathcal{F} = \{f\}$ such that $\mathbb{E}f(X_1) = 0$, $V = n\text{Var}f(X_1)$,

$$P(|\sum f(X_j)| \geq t) \leq K_1 \exp(-\frac{t}{K_2 U}\log(1 + \frac{tU}{n\text{Var}f(X_1)})),$$

which is similar to Bernstein's inequality. Compare to bounded difference inequality:

$$P(|\sum f(X_j)| \geq t) \leq 2e^{-2t^2/nU^2}$$

but since usually $V << nU^2$, Talagrand's inequality is much better.

Note that

$$\begin{aligned} V &= \mathbb{E}\sup_f |\sum f^2(X_j)| \\ &\leq \mathbb{E}\sup_f \left|\sum f^2(X_j) - \mathbb{E}f^2(X_j)\right| + \sup_f \sum_{j=1}^n \mathbb{E}f^2(X_j) \\ \text{(symmetrization ineq)} &\leq 2(2U)\mathbb{E}\sup_f |\sum \varepsilon_j f^2(X_j)/2U| + \sup_f \sum \mathbb{E}f^2(X_j) \\ \text{(contraction ineq)} &\leq 8U\mathbb{E}\sup_f |\sum \varepsilon_j f(X_j)| + \sup_f \sum \mathbb{E}f^2(X_j). \end{aligned}$$

28.2 Generalization error bounds for Adaboost

Let $\mathcal{F}$ be a class of binary classifiers of VC dimension V, and let $(X_1, Y_1), \dots, (X_n, Y_n)$ be iid from P. Let

$$\widehat{g}_T = \frac{\sum_{j=1}^T \alpha_j f_j}{\sum_{j=1}^T \alpha_j}$$

be the output of Adaboost after T steps.

Theorem 2. With probability at least $1 - e^{-t}$ and for any $\theta > 0$,

$$P(Y \neq \text{sign}(\widehat{g}_T)) \leq \frac{1}{n}\sum_{j=1}^n I\{Y_j \widehat{g}_T(X_j)) \leq \theta\} + K\left(\frac{1}{\theta}\sqrt{\frac{V}{n}} + \frac{t}{\sqrt{n}}\right).$$

Proof: Let $\varphi_\theta(x)$ be 1 if $x \leq 0$, 0 if $x \geq \theta$, and $1 - x/\theta$ otherwise. It is Lipschitz continuous with Lipschitz constant $1/\theta$. Note that $I\{X \leq 0\} \leq \varphi_\theta(x) \leq I\{X \leq \theta\}$. Then

$$\begin{aligned} P(Y\widehat{g}_T(X) \leq 0 &= \mathbb{E}I\{Y\widehat{g}_T(X) \leq 0\} \\ &\leq \mathbb{E}\varphi_\theta(Y\widehat{g}_T(X)) \\ &= \frac{1}{n}\sum \varphi_\theta(Y_j\widehat{g}_T(X_j)) - \frac{1}{n}\sum \varphi_\theta(Y_j\widehat{g}_T(X_j)) - \mathbb{E}\varphi_\theta(Y\widehat{g}_T(X)) \\ &\leq \frac{1}{n}\sum I\{Y_j\widehat{g}_T(X_j) \leq \theta\} + \sup_{g \in CH\mathcal{F}} \left|\frac{1}{n}\sum \varphi_\theta(Y_j g(X_j)) - \mathbb{E}\varphi_\theta(Y g(X))\right|. \end{aligned}$$