36-788: Topics in High Dimensional Statistics I **Fall 2015**

Lecture 9: September 29

Lecturer: Alessandro Rinaldo *Scribe: Shashank Singh*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

9.1 Recap and Outline

Last time, we stated and proved the Efron-Stein Inequality:

Theorem 9.1 (Efron-Stein Inequality): *Let $X_1, \dots, X_n$ be independent random variables, let $Z = f(X_1, \dots, X_n)$ for some real-valued function f, and suppose $\mathbb{E}\left[Z^2\right] < \infty$. Then,*

$$\mathrm{Var}\left[Z\right] = \inf_{Z_1,\dots,Z_n} \sum_{i=1}^{n} \mathbb{E}\left[\left(Z - Z_i\right)^2\right],$$

where, for each $i \in \{1, \dots, n\}$, Z_i is a function of $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$ with $\mathbb{E}\left[Z_i^2\right] < \infty$.

We also applied the Efron-Stein Inequality to bound variances in a few examples:

1. Functions with the bounded difference property (BDP).
2. Kernel density estimates (at a point) in $\mathbb{R}$.
3. Worst-case empirical probability of a family of events (e.g., cumulative distribution functions).

Today, we begin with another class of functions, self-bounded functions, for which the Efron-Stein Inequality gives strong variance bounds. We then switch topics to discuss some general tools for deriving exponential concentration inequalities. In particular, we briefly discuss martingale methods (namely, Azuma's Inequality), before moving on to outline entropy methods.

9.2 Self-Bounding Functions

Definition 9.2 (Self Bounding Property) *Let $\mathcal{X}_1, \dots, \mathcal{X}_n$ be sets, let $\mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, and, for each $i \in \{1, \dots, n\}$, let $\mathcal{X}^{(i)} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_{i-1} \times \mathcal{X}_{i+1} \times \cdots \times \mathcal{X}_n$. A function $f : \mathcal{X} \to \mathbb{R}$ is said to have the* self-bounding property (SBP) *if there exist functions $f_i : \mathcal{X}_{(i)} \to \mathbb{R}$ such that, $\forall x \in \mathcal{X}$,*

SBP(a): $0 \le f(x) - f_i(x^{(i)}) \le 1, \quad \forall i \in \{1, \dots, n\}$

SBP(b): $\sum_{i=1}^{n} f(x) - f_i(x^{(i)}) \le f(x)$.

The Efron-Stein inequality gives the following strong variance bound on the value of a self-bounding function:

Lemma 9.3 *Let $X_1, \ldots, \mathcal{X}_n$ be independent random variables taking values in $\mathcal{X}_1, \ldots, X_n$, respectively, and let $Z := f(X_1, \ldots, X_n)$. If f has the SBP, then* $\operatorname{Var}[Z] \leq \mathbb{E}[Z]$.

For this and other reasons [1], when f has the SBP, Z can be compared to a Poisson random variable.

Proof: Applying the Efron-Stein inequality, SBP(a), and SBP(b) (in that order),

$$\operatorname{Var}[Z] \leq \mathbb{E}\left[\sum_{i=1}^{n}\left(f(X) - f_i\left(X^{(i)}\right)\right)^2\right] \leq \mathbb{E}\left[\sum_{i=1}^{n} f(X) - f_i\left(X^{(i)}\right)\right] \leq \mathbb{E}[f(X)] = \mathbb{E}[Z].$$

■

At this point, the SBP appears to be a rather artificial property defined for the purpose of proving the above bound. We now go on to give a few examples of interesting functions exhibiting the SBP. One such class is given by so-called *configuration functions*.

Definition 9.4 (Configuration Function) *Let $\mathcal{X}$ be a set, let $\Pi_1 \subseteq \mathcal{X}^1, \ldots, \Pi_n \subseteq \mathcal{X}^n$, and let $\Pi := \bigcup_{i=1}^{n} \Pi_i$. The* configuration function $f_\Pi : \mathcal{X}^n \to \mathbb{N}$ of Π *gives the length of the longest subsequence in Π:*

$$f_\Pi(x_1, \ldots, x_n) = \max\{k \in \mathbb{N} : \exists i_1 \leq \cdots \leq i_k \in \{1, \ldots, n\} \text{ s.t. } (x_{i_1}, \ldots, x_{i_k}) \in \Pi\}.$$

Lemma 9.5 *Suppose $\Pi \subseteq \bigcup_{i=1}^{n} \mathcal{X}^n$ is hereditary in the sense that, if $X \in \Pi$, then any subsequence of X is is Π. If f_Π is the configuration function of Π, then f is self-bounding, and hence, if $X_1, \ldots, X_n$ are drawn independently from $\mathcal{X}$, then*

$$\operatorname{Var}[f(X)] \leq \mathbb{E}[f(X)].$$

Proof: For each $\ell \in \{1, \ldots, n\}$, let

$$f_\ell(x_1, \ldots, x_n) = \max\{k \in \mathbb{N} : \exists i_1 \leq \cdots \leq i_k \in \{1, \ldots, n\} \backslash \{\ell\} \text{ s.t. } (x_{i_1}, \ldots, x_{i_k}) \in \Pi\}.$$

denote the length of the longest subsequence in Π not including x_ℓ. For each $\ell \in \{1, \ldots, n\}$, since Π is hereditary, $f_\ell(x) = f_\Pi(x) - 1$ if every length-$f_\Pi(x)$ subsequence of x in Π contains x_i, and $f_\ell(x) = f_\Pi(x)$ otherwise. Both SBP(a) and SBP(b) follow immediately. ■

Example: Suppose $\mathcal{X}$ is countable and $(X_1, \ldots, X_n)$ is drawn from a product distribution P^n on $\mathcal{X}^n$, then $f_\Pi(X_1, \ldots, X_n) := |\{X_1, \ldots, X_n\}|$ is a configuration function, where $X \in \Pi$ precisely when X contains only distinct elements. Since changing a single coordinate of x changes f by at most 1, f has the bounded difference property (BDP), and so $\operatorname{Var}[f(X)] \leq \frac{n}{4}$. On the other hand, note that, $\forall x \in \mathcal{X}^n$,

$$f_\Pi(x) = \sum_{i=1}^{n} 1_{\{x_i \notin \{x_1, \ldots, x_{i-1}\}\}},$$

and so, letting $p_j := \mathbb{P}[X_i = j]$ for each $j \in \mathcal{X}$,

$$\mathbb{E}[f_\Pi(X)] = \sum_{i=1}^{n} \sum_{j \in \mathcal{X}} (1 - p_j)^{i-1} p_j \in o(n).$$

Hence, in this case, the SBP implies a stronger bound than does the BDP.

[1] e.g., Z satisfies the sub-Poissonian inequality $\log \mathbb{E}\left[e^{\lambda(Z - \mathbb{E}[Z])}\right] \leq \left(e^\lambda - \lambda - 1\right) \mathbb{E}[Z]$. See, e.g., [BLM09] for details.

9.3 Martingale Methods for Deriving Concentration Inequalities

The concentration inequalities we have discussed so far all assume strict independence of the underlying sequence of random variables. Martingale methods allow us to weaken this slightly. Under these weaker assumptions, we will prove an exponential concentration inequality for a sum of bounded random variables (Azuma's Inequality).

Definition 9.6 (Filtration, Martingale) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space over which $\{X_i\}_{i=1}^{\infty}$ is a sequence of real-valued random variables, and let $\{\mathcal{F}\}_{i=1}^{\infty}$ be a sequence of σ-fields increasing to $\mathcal{F}$ (a* filtration*) on Ω. Then, $\{(X_i, \mathcal{F}_i)\}_{i=1}^{\infty}$ is a* martingale *if the following hold, $\forall i \in \mathbb{N}$:*

1. *X_i is $\mathcal{F}_i$ measurable.*
2. *$\mathbb{E}[|X_i|] < \infty$.*
3. *$\mathbb{E}[X_{i+1}|X_i] = X_i$.*

For notational convenience, we will often omit the filtration and simply refer to $\{X_i\}_{i=1}^{\infty}$ as a martingale. Also, when we refer to a finite sequence $\{(X_i, \mathcal{F}_i)\}_{i=1}^{n}$ as a martingale, the above definition is intended with $X_n = X_{n+1} = X_{n+2} = \cdots$ and $\mathcal{F}_n = \mathcal{F}_{n+1} = \mathcal{F}_{n+2} = \cdots$.

Examples:

1. Suppose $Y_1, Y_2, \ldots,$ are independent and integrable. If, for each $i \in \mathbb{N}$, $X_i = \sum_{i=1}^{\infty} Y_i - \mathbb{E}[Y_i]$, then $\{(X_i, \sigma(Y_1, \ldots, Y_i))\}_{i=1}^{\infty}$ and $\{(X_i/i, \sigma(Y_1, \ldots, Y_i))\}_{i=1}^{\infty}$ are martingales.
2. If X is integrable and $\{\mathcal{F}_i\}_{i=1}^{\infty}$ is a filtration, then $\{(\mathbb{E}[X|\mathcal{F}_i], \mathcal{F}_i)\}_{i=1}^{\infty}$ is a martingale.
3. If $f : \mathcal{X}^n \to \mathbb{R}$ is integrable and $\{X_i\}_{i=1}^{n}$ is a sequence of $\mathcal{X}$-valued random variables, then

$$\{(\mathbb{E}[f(X_1, \ldots, X_n)|X_1, \ldots, X_i], \sigma(X_1, \ldots, X_i))\}_{i=1}^{n}$$

   is a martingale (referred to as the *Doob martingale of f*).

Note that, by the Law of Iterated Expectation, for each $i \in \mathbb{N}$,

$$\mathbb{E}[X_{i+1}] = \mathbb{E}[\mathbb{E}[X_{i+1}|X_i]] = \mathbb{E}[X_i] = \cdots = \mathbb{E}[X_1].$$

Hence, for convenience, we define $X_0 := \mathbb{E}[X_1]$. Then, for martingales with bounded difference between terms, we can prove the following exponential concentration inequality:

Theorem 9.7 (Azuma's Inequality) *Suppose $\{X_i\}_{i=1}^{\infty}$ is a martingale, and suppose there exists a real-valued sequence $\{c_i\}_{i=1}^{\infty}$ such that, almost surely, each $|X_i - X_{i-1}| < c_i$. Then, $\forall n \in \mathbb{N}, t > 0$,*

$$\mathbb{P}[|X_n - X_0| > t] \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right).$$

Proof: We first bound the moment generating function. For any $\lambda > 0$, by the Law of Iterated Expectation,

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n} X_i - X_{i-1}\right)\right] &= \mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n-1} X_i - X_{i-1}\right)\mathbb{E}\left[e^{\lambda(X_n - X_{n-1})}\middle| X_1, \ldots, X_{n-1}\right]\right] \\ &\leq \mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n-1} X_i - X_{i-1}\right)\right]\exp\left(\frac{\lambda^2 c_n^2}{8}\right).\end{aligned}$$

where we used Hoeffding's Lemma, conditionally given $X_1, \ldots, X_{n-1}$. [2] Expanding this recurrence gives

$$\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n} X_i - X_{i-1}\right)\right] \leq \exp\left(\frac{\lambda^2\sum_{i=1}^{n} c_i^2}{8}\right) \tag{9.1}$$

We now follow the usual Chernoff technique. $\forall\lambda > 0$, expanding the telescoping sum and applying (9.1),

$$\mathbb{P}\left[X_n - X_0 > t\right] = \mathbb{P}\left[\sum_{i=1}^{n} X_i - X_{i-1} > t\right] \leq e^{-\lambda t}\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n} X_i - X_{i-1}\right)\right] \leq \exp\left(-\lambda t + \frac{\lambda^2\sum_{i=1}^{n} c_i^2}{8}\right).$$

The exponent is convex in λ, and is easily minimized via calculus. Doing so gives $\lambda = \frac{4t}{\sum_{i=1}^{n} c_i^2}$, and so

$$\mathbb{P}\left[X_n - X_0 > t\right] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right).$$

Plugging in $-X_0, \ldots, -X_n$ gives the corresponding left tail bound, and a union bound finishes the proof. ■

Applying Azuma's Inequality to a Doob martingale gives the widely used McDiarmid's Inequality [M89], also known as the Method of Bounded Differences:

Corollary 9.8 (McDiarmid's Inequality) *If $X_1, \ldots, X_n$ are independent $\mathcal{X}$-valued random variables and $f : \mathcal{X}^n \to \mathbb{R}$ satisfies the bounded difference property with constants $c_1, \ldots, c_n$, then, $\forall t > 0$,*

$$\mathbb{P}\left[|f\left(X_1, \ldots, X_n\right) - \mathbb{E}\left[f\left(X_1, \ldots, X_n\right)\right| > t\right]\right] \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right).$$

9.3.1 Improvements to Azuma's Inequality

See [CL06] and chapter 2 of [RS14] for surveys of martingale methods for deriving concentration inequalities.

Azuma's and McDiarmid's Inequalities are both easy to prove and widely applicable, resulting in many refinements and generalizations. If we can bound the quadratic variation of the martingale in question, we can derive a martingale analogue of Bernstein's Inequality. [R13] observes that McDiarmid's Inequality is loose for large t (e.g., on the order of the diameter of the underlying metric space) and provides tighter bounds for this regime. [K14] weakens the bounded difference assumption to a *bounded sub-Gaussian diameter* assumption. Finally, [S11] gives some refinements of Azuma's Inequality in terms of information-theoretic quantities.

[2] Hoeffding's Lemma states that, if a random variable X satisfies $\mathbb{E}\left[X\right] = 0$ and $|X| \leq c$ a.s. for some $c \in \mathbb{R}$, then, $\forall\lambda \in \mathbb{R}$, $\mathbb{E}\left[e^{\lambda X}\right] \leq \exp\left(\lambda^2 c^2/8\right)$. The proof follows from convexity of $x \mapsto e^{\lambda x}$, Jensen's Inequality, and a 2^{nd} order Taylor bound.

9.4 Entropy Methods for Deriving Concentration Inequalities

Entropy methods are a recent innovation (developed in the last 15 years), which tend to give very sharp concentration bounds, but rely on some highly non-trivial technical results. They rely on two steps to bound the log moment generating function:

1. Tensorization (sub-additivity) of entropy: decompose the entropy of the random variable into contributions of each coordinate (e.g., the Efron-Stein Inequality does this for variance instead of entropy).
2. Herbst's Argument: Re-express the entropy bound as a differential inequality, which can be implicitly solved via Taylor expansion to give concentration bounds.

Next time, after defining entropy and divergence and proving several technical lemmas, we will flesh out this approach and use it to prove actual concentration inequalities.

References

[BLM09] S. Boucheron, G. Lugosi, and P. Massart, "On concentration of self-bounding functions," *Electronic Journal of Probability*, 2009, pp. 1884–1899.

[M89] C. McDiarmid, "On the method of bounded differences," *Surveys in Combinatorics*, 1989, pp. 148–188.

[CL06] F. Chung and L. Lu, "Concentration inequalities and martingale inequalities: a survey," *Internet Mathematics*, 2006, pp. 79–127.

[RS14] M. Raginsky and I. Sason, "Concentration of Measure Inequalities in Information Theory, Communications and Coding," *Foundations and Trends in Communications and Information Theory*, 2014.

[R13] E. Rio, "On McDiarmid's concentration inequality," *Electronic Communications in Probability*, 2013.

[S11] I. Sason, "On refined versions of the Azuma-Hoeffding inequality with applications in information theory," *arXiv:1111.1977*, 2011.

[K14] A. Kontorovich, "Concentration in unbounded metric spaces and algorithmic stability," *International Conference on Machine Learning*, 2014.