MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.265/15.070J | Fall 2013
Lecture 12 | 10/21/2013

Martingale Concentration Inequalities and Applications

Content.

1. Exponential concentration for martingales with bounded increments
2. Concentration for Lipschitz continuous functions
3. Examples in statistics and random graph theory

1 Azuma-Hoeffding inequality

Suppose X_n is a martingale wrt filtration $\mathcal{F}_n$ such that $X_0 = 0$ The goal of this lecture is to obtain bounds of the form $\mathbb{P}(|X_n| \geq \delta n) \leq \exp(-\Theta(n))$ under some condition on X_n. Note that since $\mathbb{E}[X_n] = 0$, the deviation from zero is the "right" regime to look for rare events. It turns out the exponential bound of the form above holds under very simple assumption that the increments of X_n are bounded. The theorem below is known as Azuma-Hoeffding Inequality.

Theorem 1 (**Azuma-Hoeffding Inequality**). *Suppose $X_n, n \geq 1$ is a martingale such that $X_0 = 0$ and $|X_i - X_{i-1}| \leq d_i, 1 \leq i \leq n$ almost surely for some constants $d_i, 1 \leq i \leq n$. Then, for every $t > 0$,*

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

Notice that in the special case when $d_i = d$, we can take $t = xn$ and obtain an upper bound $2\exp\left(-x^2 n/(2d^2)\right)$ - which is of the form promised above. Note that this is consistent with the Chernoff bound for the special case X_n is the sum of i.i.d. zero mean terms, though it is applicable only in the special case of a.s. bounded increments.

Proof. $f(x) \triangleq \exp(\lambda x)$ is a convex function in x for any $\lambda \in \mathbb{R}$. Then we have $f(-d_i) = \exp(-\lambda d_i)$ and $f(d_i) = \exp(\lambda d_i)$. Using convexity we have that when $|x/d_i| \leq 1$

$$\begin{aligned}\exp(\lambda x) = f(x) &= f \quad \frac{1}{2}(\frac{x}{d_i}+1)d_i + \frac{1}{2}(1-\frac{x}{d_i})(-d_i) \\ &\leq \frac{1}{2} \quad \frac{x}{d_i}+1 \quad f(d_i) + \frac{1}{2} \quad 1-\frac{x}{d_i} \quad f(-d_i) \\ &= \frac{f(d_i)+f(-d_i)}{2} + \frac{f(d_i)-f(-d_i)}{2}x. \end{aligned} \tag{1}$$

Further, for every a

$$\begin{aligned}\frac{\exp(a)+\exp(-a)}{2} &= \sum_{k=0}^{\infty}\frac{a^k}{k!} + \sum_{k=0}^{\infty}\frac{(-1)^k a^k}{k!} = \sum_{k=0}^{\infty}\frac{a^{2k}}{(2k)!} \\ &\leq \sum_{k=0}^{\infty}\frac{a^{2k}}{2^k k!} \text{ (because } 2^k k! \leq (2k)!) \\ &= \sum_{k=0}^{\infty}\frac{(\frac{a^2}{2})^k}{k!} = \exp(\frac{a^2}{2}). \end{aligned} \tag{2}$$

We conclude that for every x such that $|x/d_i| \leq 1$

$$\exp(\lambda x) \leq \exp(\frac{d_i^2}{2}) + \frac{\exp(\lambda d_i) - \exp(-\lambda d_i)}{2}x. \tag{3}$$

We now turn to our martingale sequence X_n. For every $t > 0$ and every $\lambda > 0$ we have

$$\begin{aligned}\mathbb{P}(X_n \geq t) &= \mathbb{P}\left(\exp(\lambda X_n) \geq \exp(\lambda t)\right) \\ &\leq \exp(-\lambda t)\mathbb{E}[\exp(\lambda X_n)] \\ &= \exp(-\lambda t)\mathbb{E}[\exp(\lambda \sum_{1\leq i\leq n}(X_i - X_{i-1}))], \end{aligned}$$

where $X_0 = 0$ was used in the last equality. Applying the tower property of conditional expectation we have

$$\begin{aligned}&\mathbb{E}\left[\exp(\lambda \sum_{1\leq i\leq n}(X_i - X_{i-1}))\right] \\ &\quad = \mathbb{E}\left[\mathbb{E}\left[\exp(\lambda(X_n - X_{n-1}))\exp(\lambda \sum_{1\leq i\leq n-1}(X_i - X_{i-1}))|\mathcal{F}_{n-1}\right]\right]. \end{aligned}$$

Now, since $X_i, i \leq n-1$ are measurable wrt $\mathcal{F}_{n-1}$, then

$$\begin{aligned}
&\mathbb{E}\left[\exp(\lambda(X_n - X_{n-1}))\exp(\lambda \sum_{1\leq i\leq n-1}(X_i - X_{i-1}))|\mathcal{F}_{n-1}\right] \\
&= \exp(\lambda \sum_{1\leq i\leq n-1}(X_i - X_{i-1}))\mathbb{E}\left[\exp(\lambda(X_n - X_{n-1}))|\mathcal{F}_{n-1}\right] \\
&\leq \exp(\lambda \sum_{1\leq i\leq n-1}(X_i - X_{i-1}))\times \\
&\times \left(\exp\left(\frac{\lambda^2 d_n^2}{2}\right) + \frac{\exp(\lambda d_i) - \exp(-\lambda d_i)}{2}\mathbb{E}[X_n - X_{n-1}|\mathcal{F}_{n-1}]\right),
\end{aligned}$$

where (3) was used in the last inequality. Martingale property implies $\mathbb{E}[X_n - X_{n-1}|\mathcal{F}_{n-1}] = 0$, and we have obtained an upper bound

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

Iterating further we obtain the following upper bound on $\mathbb{P}(X_n \geq t)$:

$$\exp(-\lambda t)\exp\left(\frac{\sum_{1\leq i\leq n}\lambda^2 d_i^2}{2}\right)$$

Optimizing over the choice of λ, we see that the tightest bound is obtained by setting $\lambda = t/\sum_i d_i^2 > 0$, leading to an upper bound

$$\mathbb{P}(X_n \geq t) \leq \exp\left(-\frac{t^2}{2\sum_i d_i^2}\right).$$

A similar approach using $\lambda < 0$ gives for every $t > 0$

$$\mathbb{P}(X_n \leq -t) \leq \exp\left(-\frac{t^2}{2\sum_i d_i^2}\right).$$

Combining, we obtain the required result. □

2 Application to Lipschitz continuous functions of i.i.d. random variables

Suppose $X_1, ..., X_n$ are independent random variables. Suppose $g : \mathbb{R}^n \to \mathbb{R}$ is a function and $d_1, \ldots, d_n$ are constants such that for any two vectors $x_1, \ldots, x_n$

and $y_1, \dots, y_n$

$$|g(x_1, \dots, x_n) - g(y_1, \dots, y_n)| \le \sum_{i=1}^{n} d_i \mathbf{1}\{x_i \neq y_i\}. \tag{4}$$

In particular when a vector x changes value only in its i-th coordinate the amount of change in function g is at most d_i. As a special case, consider a subset of vectors $x = (x_1, \dots, x_n)$ such that $|x_i| \le c_i$ and suppose g is Lipschitz continuous with constant K. Namely, for for every x, y, $|g(x) - g(y)| \le K|x - y|$, where $|x - y| = \quad {}_i |x_i - y_i|$. Then for any two such vectors

$$|g(x) - g(y)| \le K|x - y| \le K \sum_i 2c_i |x_i - y_i|,$$

and therefore this fits into a previous framework with $d_i = Kc_i$.

Theorem 2. *Suppose $X_i, 1 \le i \le n$ are i.i.d. and function $g : \mathbb{R}^n \to \mathbb{R}$ satisfies (4). Then for every $t \ge 0$*

$$\mathbb{P}(|g(X_1, ..., X_n) - \mathbb{E}[g(X_1, ..., X_n)]| > t) \le 2 \exp\left(-\frac{t^2}{2\sum_i d_i^2}\right).$$

Proof. Let $\mathcal{F}_i$ be the σ-field generated by variables $X_1, \dots, X_i$: $\mathcal{F}_i = \sigma(X_1, ..., X_i)$. For convenience, we also set $\mathcal{F}_0$ to be the trivial σ-field consisting of $\emptyset, \Omega$, so that $E[Z|\mathcal{F}_0] = \mathbb{E}[Z]$ for every r.v. Z. Let $M_0 = \mathbb{E}[g(X_1, ..., X_n)]$, $M_1 = \mathbb{E}[g(X_1, ..., X_n)|\mathcal{F}_1]$,...,$M_n = \mathbb{E}[g(X_1, ..., X_n)|\mathcal{F}_n]$. Observe that M_n is simply $g(X_1, \dots, X_n)$, since $X_1, \dots, X_n$ are measurable wrt $\mathcal{F}_n$. Thus, we by tower property

$$\mathbb{E}[M_n|\mathcal{F}_{n-1}] = \mathbb{E}[\mathbb{E}[g(X_1, ..., X_n)|\mathcal{F}_n]|\mathcal{F}_{n-1}] = M_{n-1}.$$

Thus, M_i is a martingale. We have

$$\begin{aligned} M_{i+1} - M_i &= \mathbb{E}[\mathbb{E}[g(X_1, ..., X_n)|\mathcal{F}_{i+1}] - \mathbb{E}[g(X_1, ..., X_n)|\mathcal{F}_i]] \\ &= \mathbb{E}[\mathbb{E}[g(X_1, ..., X_n) - \mathbb{E}[g(X_1, ..., X_n)|\mathcal{F}_i]|\mathcal{F}_{i+1}]]. \end{aligned}$$

Since X_i's are independent, then M_i is a r.v. which on any vector $x = (x_1, \dots, x_n) \in \Omega$ takes value

$$M_i = \int_{x_{i+1}, \dots, x_n} g(x_1, ..., x_n) d\mathbb{P}(x_{i+1}) \cdots d\mathbb{P}(x_n),$$

(and in particular only depends on the first i coordinates of x). Similarly

$$M_{i+1} = \int_{x_{i+2},\ldots,x_n} g(x_1, ..., x_n) d\mathbb{P}(x_{i+2}) \cdots d\mathbb{P}(x_n).$$

Thus

$$\begin{aligned}|M_{i+1} - M_i| &= |\int_{x_{i+2},\ldots,x_n} (g(x_1, ..., x_n) - \int_{x_{i+1}} g(x_1, ..., x_n) d\mathbb{P}(x_{i+1})) d\mathbb{P}(x_{i+1}) \cdots d\mathbb{P}(x_n)|, \\ &\leq d_{i+1} \int_{x_{i+2},\ldots,x_n} d\mathbb{P}(x_{i+1}) \cdots d\mathbb{P}(x_n) \\ &= d_{i+1}.\end{aligned}$$

This derivation represents a simple idea that M_i and M_{i+1} only differ in "averaging out" X_{i+1} in M_i. Now defining $\hat{M}_i = M_i - M_0 = M_i - \mathbb{E}[g(X_1, \ldots, X_n)]$, we have that $\hat{M}_i$ is also a martingale with differences bounded by d_i, but with an additional property $M_0 = 0$. Applying Theorem 1 we obtain the required result. □

3 Two examples

We now consider two applications of the concentration inequalities developed in the previous sections. Our first example concerns convergence empirical distributions to the true distributions of random variables. Specifically, suppose we have a distribution function F, and i.i.d. sequence $X_1, \ldots, X_n$ with distribution F. From the sample $X_1, \ldots, X_n$ we can build an empirical distribution function $F_n(x) = n^{-1} \sum_{1 \leq i \leq n} \mathbf{1}\{X_i \leq x\}$. Namely, $F_n(x)$ is simply the frequency of observing values at most x in our sample. We should realize that F_n is a random function, since it depends on the sample $X_1, \ldots, X_n$. An important Theorem called Glivenko-Cantelli says that $\sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$ converges to zero and in expectation, the latter meaning of course that $\mathbb{E}[\sup_{x \in \mathbb{R}} |F_n(x) - F(x)|] \rightarrow 0$. Proving this result is beyond our scope. However, applying the martingale concentration inequality we can bound the deviation of $\sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$ around its expectation. For convenience let $L_n = L_n(X_1, \ldots, X_n) = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$, which is commonly called empirical risk in the statistics and machine learning fields. We need to bound $\mathbb{P}(|L_n - E[L_n]| > t)$. Observe that L satisfies property (4) with $d_i = 1/n$. Indeed changing one coordinate X_i to some X_i' changes F_n by at most $1/n$, and thus the same applies to

L_n. Applying Theorem 2 we obtain

$$\begin{aligned}\mathbb{P}(|L_n - E[L_n]| > t) &\leq 2\exp\left(-\frac{t^2}{2n(1/n)^2}\right)\\ &= 2\exp\left(-\frac{t^2 n}{2}\right).\end{aligned}$$

Thus, we obtain a large deviations type bound on the difference $L_n - \mathbb{E}[L_n]$.

For our second example we turn to combinatorial optimization on random graphs. We will use the so-called Max-Cut problem as an example, though the approach works for many other optimization and constraint satisfaction problems as well. Consider a simple undirected graph $G = (V, E)$. V is the set of nodes, denoted by $1, 2, \ldots, n$. And E is the set of edges which we describe as a list of pairs $(i_1, j_1), \ldots, (i_{|E|}, j_{|E|})$, where $i_1, \ldots, i_{|E|}, j_1, \ldots, j_{|E|}$ are nodes. The graph is undirected, which means that the edges (i_1, j_1) and (j_1, i_1) are identical. We can also represent the graph as an $n \times n$ zero-one matrix A, where $A_{i,j} = 1$ if $(i, j) \in E)$ and $A_{i,j} = 0$ otherwise. Then A is a symmetric matrix, namely $A^T = A$, where A^T is a transpose of A. A cut in this graph is a partition σ of nodes into two groups, encoded by function a function $\sigma : V \rightarrow \{0, 1\}$. The value $MC(\sigma)$ of the cut associated with σ is the number of edges between the two groups. Formally, $MC(\sigma) = |\{(i, j) \in E : \sigma(i) \neq \sigma(j)\}|$. Clearly $MC(\sigma) \leq |E|$. At the same time, a random assignment $\sigma(i) = 0$ with probability $1/2$ and $= 1$ with probability $1/2$ gives a cut with expected value $MC(\sigma) \geq (1/2)|E|$. In fact there is a simple algorithm to construct such a cut explicitly. Now denote by $MC(G)$ the maximum possible value of the cut: $MC(G) = \max_\sigma MC(\sigma)$. Thus $1/2 \leq MC(G)/|E| \leq 1$. Further, suppose we delete an arbitrary edge from the graph G and obtain a new graph G'. Observe that in this case $MC(G') \geq MG(G) - 1$ - the Max-Cut value either stays the same or goes down by at most one. Similarly, when we add an edge, the Max-Cut value increases by at most one. Putting this together, if we replace an arbitrary edge $e \in E$ by a different edge e' and leave all the other edges intact, the value of the Max-Cut changes by at most one.

Now suppose the graph $G = G(n, dn)$ is a random Erdös-Rényi graph with $|E| = dn$ edges. Specifically, suppose we choose every edges $E_1, \ldots, E_{dn}$ uniformly at random from the total set of $\binom{n}{2}$ edges, independently for these nd choices. Denote by MC_n the value of the maximum cut $MC(G(n, dn))$ on this random graph. Since the graph is random, we have that MC_n is a random variable. Furthermore, as we have just established, $d/2 \leq MC_n/n \leq d$. One of the major open problems in the theory of random graphs is computing the scaling limit $\mathbb{E}[MC_n]/n$ as $n \rightarrow \infty$. However, we can easily obtain bounds on the concentration of MC_n around its expectation, using Azuma-Hoeffding inequality. For this goal, think of random edges $E_1, \ldots, E_{dn}$ as i.i.d. random variables in the space $1, 2, \ldots, \big({n \atop 2}$ corresponding to the space of all possible edges on n nodes. Let $g(E_1, \ldots, E_{dn}) = MC_n$. Observe that indeed g is a function of dn i.i.d. random variables. By our observation, replacing one edge E_i by a different edge E_i' changes MC_n by at most one. Thus we can apply Theorem 2 which gives

$$\mathbb{P}\left(|MC_n - \mathbb{E}[MC_n]| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2dn}\right).$$

In particular, taking $t = rn$, where $r > 0$ is a constant, we obtain a large deviations type bound $2\exp(-\frac{r^2 n}{2d})$. Taking instead $t = r\sqrt{n}$, we obtain Gaussian type bound $2\exp(-\frac{r^2}{2d})$. Namely, $MC_n = \mathbb{E}[MC_n] + \Theta(\sqrt{n})$. This is a meaningful concentration around the mean since, as we have discussed above $\mathbb{E}[MC_n] = \Theta(n)$.