36-709: Advanced Statistical Theory **Fall 2019**

Lecture 5: February 12

Lecturer: Alessandro Rinaldo *Scribe: David Zhao*

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

In this lecture, we prove the bounded differences inequality and give an example of its application. We then begin our discussion of density estimation.

5.1 The bounded differences inequality

Recall the bounded differences property we defined in the previous lecture.

Definition: We say that a function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the bounded differences property (BDP) if $\exists L_1, \ldots, L_n > 0$ s.t. for any $(X_1, \ldots, X_n)$ in the domain of f, for any coordinate k,

$$\sup_{x', y'} |f(x_1, \ldots, x_{k-1}, y', x_{k+1}, \ldots, x_n) - f(x_1, \ldots, x_{k-1}, x', x_{k+1}, \ldots, x_n)| \leq L_k$$

Now, we show that functions of random variables that satisfy the BDP concentrate around their expectations.

Theorem (Bounded Differences Inequality): Let $X_1, \ldots, X_n$ be independent random variables, and let $f : \mathbb{R}^n \to \mathbb{R}$ be a function that satisfies the BDP as defined above, with constants $L_1, \ldots, L_n$. Then

$$\mathbb{P}(|Z - \mathbb{E}[Z]| > t) \leq 2 \exp\Big(- \frac{2t^2}{\sum_{k=1}^n L_k^2} \Big)$$

where $Z = f(X_1, \ldots, X_n)$.

Proof: Suppose $D_1, \ldots, D_n$ is a martingale difference sequence, and suppose that $a_k \leq D_k \leq b_k, \ \forall k$ a.e. for sequences of constants $\{a_k\}$ and $\{b_k\}$. Then by the Azuma-Hoeffding inequality, we have

$$\mathbb{P}\Big(\Big| \sum_{k=1}^n D_k \Big| > t \Big) \leq 2 \exp\Big(- \frac{2t^2}{\sum_{k=1}^n (b_k - a_k)^2} \Big), \ \forall t > 0$$

Let

$$D_k = \mathbb{E}[Z | X_1, \ldots, X_k] - \mathbb{E}[Z | X_1, \ldots, X_{k-1}], \ k > 1$$
$$D_0 = \mathbb{E}[Z]$$

Then $Z - \mathbb{E}[Z] = \sum_{k=1}^n D_k$. Note that $D_1, \ldots, D_n$ form a martingale difference sequence.

Now, for each $k = 1, \ldots, n$, define

$$A_n = \inf_x \Big\{ \mathbb{E}[Z | X_1, \ldots, X_{k-1}, x] - \mathbb{E}[Z | X_1, \ldots, X_{k-1}] \Big\}$$

$$B_n = \sup_x \Big\{ \mathbb{E}[Z|X_1, \ldots, X_{k-1}, x] - \mathbb{E}[Z|X_1, \ldots, X_{k-1}] \Big\}$$

Observe that

$$D_k - A_k = \mathbb{E}[Z|X_1, \ldots, X_k] - \inf_x \Big\{ \mathbb{E}[Z|X_1, \ldots, X_{k-1}, x] \Big\} \geq 0 \ a.e.$$

Similarly,

$$B_k - D_k = \sup_x \Big\{ \mathbb{E}[Z|X_1, \ldots, X_{k-1}, x] \Big\} - \mathbb{E}[Z|X_1, \ldots, X_k] \geq 0 \ a.e.$$

In other words, for sequences of random variables $\{A_k\}$ and $\{B_k\}$, it holds that $A_k \leq D_k \leq B_k, \ \forall k$ a.e.

Finally, we unpack $B_k - A_k$ for each k, and show that it is $\leq L_k$, using the independence assumption.

$$B_k - A_k = \sup_x \Big\{ \mathbb{E}[Z|X_1, \ldots, X_{k-1}, x] \Big\} - \inf_y \Big\{ \mathbb{E}[Z|X_1, \ldots, X_{k-1}, y] \Big\}$$

$$= \sup_x \Big\{ \int f(X_1, \ldots, X_{k-1}, x, x_{k+1}, \ldots, x_n) \ dP(x_{k+1}, \ldots, x_n | X_1, \ldots, X_{k-1}, x) \Big\}$$

$$- \inf_y \Big\{ \int f(X_1, \ldots, X_{k-1}, y, x_{k+1}, \ldots, x_n) \ dP(x_{k+1}, \ldots, x_n | X_1, \ldots, X_{k-1}, y) \Big\}$$

Let x^* and y^* be the (random) numbers that achieve the sup and inf over z, respectively, of

$$\int f(X_1, \ldots, X_{k-1}, z, x_{k+1}, \ldots, x_n) \ dP(x_{k+1}, \ldots, x_n | X_1, \ldots, X_{k-1}, z)$$

Note that in general,

$$dP(x_{k+1}, \ldots, x_n | X_1, \ldots, X_{k-1}, x^*) \neq dP(x_{k+1}, \ldots, x_n | X_1, \ldots, X_{k-1}, y^*)$$

unless $X_1, \ldots, X_n$ are independent, in which case both expressions are equal to

$$dP(x_{k+1}) \ldots dP(x_n)$$

But the theorem does assume that $X_1, \ldots, X_n$ are independent, so we have that

$$B_k - A_k = \sup_x \Big\{ \int f(X_1, \ldots, X_{k-1}, x, x_{k+1}, \ldots, x_n) \ dP(x_{k+1}, \ldots, x_n | X_1, \ldots, X_{k-1}, x) \Big\}$$

$$- \inf_y \Big\{ \int f(X_1, \ldots, X_{k-1}, y, x_{k+1}, \ldots, x_n) \ dP(x_{k+1}, \ldots, x_n | X_1, \ldots, X_{k-1}, y) \Big\}$$

$$= \sup_x \Big\{ \int f(X_1, \ldots, X_{k-1}, x, x_{k+1}, \ldots, x_n) \ dP(x_{k+1}) \ldots dP(x_n) \Big\}$$

$$- \inf_y \Big\{ \int f(X_1, \ldots, X_{k-1}, y, x_{k+1}, \ldots, x_n) \ dP(x_{k+1}) \ldots dP(x_n) \Big\}$$

$$\leq \sup_{x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n, x, y} \Big\{ \Big| f(x_1, \ldots, x_{k-1}, x, x_{k+1}, \ldots, x_n) - f(x_1, \ldots, x_{k-1}, y, x_{k+1}, \ldots, x_n) \Big| \Big\}$$

$$\leq L_k, \ \forall k \ a.e.$$

and the claim follows. ■

5.2 Application of the bounded differences inequality

Example: Let G_n be an Erdos-Renyi graph on n nodes. Let the $\binom{n}{2}$ potential edges be i.i.d. Bernoulli random variables, where 1 indicates the presence of an edge and 0 indicates its absence. Let C_n be the clique number, i.e. the size of the largest complete subgraph.

We show that C_n concentrates around its expectation. Note that C_n can be thought of as the result of a function applied to all of the $\binom{n}{2}$ potential edges of the graph. This function satisfies the BDP, because if we only change one edge at a time while keeping all other edges fixed, then C_n changes by at most 1. In other words, $L_k = 1,\ \forall k$. It follows that

$$\mathbb{P}\Big(\Big|\frac{C_n}{n} - E\Big[\frac{C_n}{n}\Big]\Big| \geq t\Big) = \mathbb{P}\Big(|C_n - E[C_n]| \geq nt\Big) \leq 2\exp\Big(-\frac{2n^2t^2}{\sum_{k=1}^n 1}\Big) = 2\exp(-2nt^2),\ \forall t > 0$$

Example: We could also consider T_n, the number of triangles in the graph. Note that if T_n is the result of a function applied to all of the $\binom{n}{2}$ potential edges, then changing one edge at a time can change the value of T_n by as much as $n-2$. So here, we cannot apply the BDP in a meaningful way.

5.3 Introduction to density estimation

We now turn our attention to a different type of problem. Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f$ where f is an unknown density. We seek to estimate f using our sample. Note that f is a point not in Euclidean space, but in a functional space, which makes density estimation a complex problem.

One type of density estimator is called the kernel density estimator.

Definition: Let $K : \mathbb{R} \to \mathbb{R}_+$ be a function such that $\int_{\mathbb{R}} K(x)dx = 1$. Then for a parameter $h > 0$ known as the bandwidth, define the kernel density estimator as

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{k=1}^n K\Big(\frac{X_k - x}{h}\Big),\ x \in \mathbb{R}$$

In other words, we are simply convolving the empirical measure (consisting of point masses at the observed data points) with a kernel function to "smooth it out".

Let us define a function $x \to f_h(x)$ as the pointwise expectation $\mathbb{E}[\hat{f}_h(x)]$ at every x. Then $f_h(x)$ is a deterministic probability density, determined by the kernel function K and the choice of bandwidth h.

Note that there are two levels of approximation here: the deviation $\hat{f}_h(x) - f_h(x)$ representing the variance, and the deviation $f_h - f$ representing the bias. As $h \to 0$, the bias $\to 0$ and the variance $\to \infty$.

We never explicitly observe the bias because we obviously do not know the true underlying function f. If we make certain smoothness assumptions about f belonging to a particular class of functions (e.g. a Holder class), we can derive expressions for both the bias and variance, which lead to well-known results on the optimal choice of bandwidth h as a function of n, and the resulting optimal rate of convergence.

Note that the above well-known results usually consider convergence rates in the sense of mean squared error, or L_2 error defined in terms of the L_2 norm:

$$\|\hat{f}_h(x) - f_h(x)\|_2 = \sqrt{\int (\hat{f}_h(x) - f_h(x))^2 dx}$$

The L_2 norm is popular because it is convenient to work with computationally, but a more natural way of thinking statistically about how "well" $\hat{f}_h$ approximates f_h is the L_1 norm:

$$\|\hat{f}_h(x) - f_h(x)\|_1 = \int |\hat{f}_h(x) - f_h(x)| dx = 2 d_{TV}(\hat{P}_h, P_h) = 2 \sup_{A \in \mathcal{A}} |\hat{P}_h(A) - P_h(A)|$$

where $d_{TV}(P, Q)$ is the TV distance between probability measures P and Q. Note that TV distance defines a very strong notion of similarity, as it considers the subset of the domain over which the two measures differ the most. So two measures that are close in TV distance are very similar over all possible subsets we might take over the domain.

Unfortunately, the L_1 norm is inconvenient to work with in practice because it is difficult to compute. For reference, see [DL01] for various methods to compute and use the L_1 norm in density estimation.

We conclude today by noting that the L_1 distance satisfies the BDP, with $L_k = \dfrac{2}{n},\ \forall k$, since

$$f(x_1, \ldots, x_{k-1}, x, x_{k+1}, \ldots, x_n) - f(x_1, \ldots, x_{k-1}, y, x_{k+1}, \ldots, x_n) \leq \frac{1}{nh} \int \left| K\Big(\frac{x - z}{h}\Big) - K\Big(\frac{y - z}{h}\Big) \right| dz$$

$$\leq \frac{1}{nh} \Big[h \int K(w) dw + h \int K(v) dv \Big] = \frac{2}{n}$$

using the substitutions $w = \dfrac{x - z}{h}$ and $v = \dfrac{y - z}{h}$.

It follows that

$$\mathbb{P}\Big(\Big| L_1(\hat{f}_h, f_h) - \mathbb{E}[L_1(\hat{f}_h, f_h)] \Big| > t \Big) \leq 2 \exp\Big(- \frac{nt^2}{2} \Big),\ \forall t > 0$$

The remarkable thing is that this concentration property holds regardless of the choice of bandwidth h.

Of course, in this case, it is easy to show that $L_1(\hat{f}_h, f_h)$ concentrates around its expectation, but, as we have noted, difficult to actually compute what that expectation is.

References

[DL01] L. Devroye and G. Lugosi, "Combinatorial Methods in Density Estimation," *Springer Series in Statistics*, 2001.