Analysis of Boolean Functions **(CMU 18-859S, Spring 2007)**

Lecture 21: Berry-Esseen Theorem

April. 3, 2007

Lecturer: Ryan O'Donnell *Scribe: Suresh Purini*

1 Berry-Esseen Theorem

In this class we study a simplified version of the Berry-Esseen theorem, which quantifies how "close" are the distributions of the two random variables $\boldsymbol{Q}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ and $\boldsymbol{Q}(\boldsymbol{g}_1, \cdots, \boldsymbol{g}_n)$, where $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ are uniform random bits ± 1 and $\boldsymbol{g}_1, \cdots, \boldsymbol{g}_n$ are Gaussian random variables with mean 0 and variance 1, and Q is a multivariate polynomial of degree 1 or in other words it is an affine function. In the next class, we study a version of Berry-Esseen theorem where Q is a multi-linear polynomial. This theorem is used to prove that "Majority is Stablest". The following is the main theorem of this class.

Theorem 1.1 *Let $Q(u_1, \cdots, u_n) = \sum_{i=1}^{n} \alpha_i u_i$, $\alpha_i \in \mathbb{R}$, be a linear polynomial over formal variables $u_1, \cdots, u_n$. Also assume that $\sum_{i=1}^{n} \alpha_i^2 = 1$ and $\alpha_i^2 \leq \tau, \forall i \in [n]$. Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ be i.i.d uniform random ± 1 bits. Let the random variable $\boldsymbol{g}$ be normally distributed with mean 0 and variance 1, i.e., $\boldsymbol{g} \sim N(0,1)$. Then the random variables $\boldsymbol{Q}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ and $\boldsymbol{g}$ are "close" in distribution. In particular*

1. $\forall t_o \in \mathbb{R}$, $|\mathbf{Pr}[\boldsymbol{Q}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n) < t_0] - \mathbf{Pr}[\boldsymbol{g} < t_0]| \leq O(\tau)$.
2. $|\mathbf{E}[|\boldsymbol{Q}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)|] - \mathbf{E}[|g|]| \leq O(\tau)$.

We first understand some of the ideas used in the proof of theorem 1.1.

Idea 1: The first idea is to replace the Gaussian random variable $\boldsymbol{g}$ by another Gaussian random variable with same mean and variance but looks very similar to the random variable $\boldsymbol{Q}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$. Let $\boldsymbol{g}_1, \cdots, \boldsymbol{g}_n$ be i.i.d Gaussian random variables with mean 0 and variance 1. Consider the random variable $\boldsymbol{Q}(\boldsymbol{g}_1, \cdots, \boldsymbol{g}_n) \sim \sum_{i=1}^{n} \alpha_i \boldsymbol{g}_i$. It turns out that $\boldsymbol{Q}(\boldsymbol{g}_1, \cdots, \boldsymbol{g}_n) \sim N(0, \sum_{i=1}^{n} \alpha_i^2) \sim N(0,1)$[1] is also Gaussianly distributed. Now we can compare the distributions $\boldsymbol{Q}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ and $\boldsymbol{Q}(\boldsymbol{g}_1, \cdots, \boldsymbol{g}_n)$ which at least appear to look alike.

[1] If $\boldsymbol{X} \sim N(0,1)$, then $\alpha \boldsymbol{X} \sim N(0, \alpha^2)$. If $\boldsymbol{X} \sim N(\mu, \sigma^2)$ and $\boldsymbol{Y} \sim N(\nu, \tau^2)$, then $\boldsymbol{X} + \boldsymbol{Y} \sim N(\mu + \nu, \sigma^2 + \tau^2)$. In other words, sum of Gaussian random variables is Gaussian whose mean and variance are respectively equal to sum of the individual means and variances of Gaussian random variables used in summation. Refer http://en.wikipedia.org/wiki/Sum_of_normally_distributed_random_variables

Idea 2: We shall try to see a generic way to say that two random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ are close in distribution. For all "nice" test functions $\psi : \mathbb{R} \to \mathbb{R}$,

$$|\mathbf{E}[\psi(\boldsymbol{X})] - \mathbf{E}[\psi(\boldsymbol{Y})]| \leq \text{"small"}.$$

First note that if the functions $\psi_{t_0}(t) = \begin{cases} 1 & \text{if } t < t_0 \\ 0 & \text{otherwise} \end{cases}$ and $\psi_2(t) = |t|$ fit in the definition of "nice", then by letting $\boldsymbol{X} = \boldsymbol{Q}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ and $\boldsymbol{Y} = \boldsymbol{Q}(\boldsymbol{g}_1, \cdots, \boldsymbol{g}_n)$ we shall get statements 1 and 2 of theorem 1.1 respectively. We shall later see that the above two functions ψ_1 and ψ_2 do not fit in our definition of "nice" but they can be approximated by "nice" functions.

Let us look at some examples of "nice" functions with respect to the random variables $\boldsymbol{X} = \boldsymbol{Q}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ and $\boldsymbol{Y} = \boldsymbol{Q}(\boldsymbol{g}_1, \cdots, \boldsymbol{g}_n)$. Note that if $\boldsymbol{g} \sim N(0,1)$, then $\mathbf{E}[\boldsymbol{g}^{2k+1}] = 0$ and $\mathbf{E}[\boldsymbol{g}^{2k}] = (2k-1) \cdot (2k-3) \cdots 5 \cdot 3 \cdot 1$. In particular $\mathbf{E}[\boldsymbol{g}^4] = 3$. These facts are repeatedly used in the following examples.

1. Let $\psi(t) = a + bt$. Then

$$\begin{aligned} \mathbf{E}[\psi(\boldsymbol{X})] - \mathbf{E}[\psi(\boldsymbol{Y})] &= \mathbf{E}[a + b\boldsymbol{X}] - \mathbf{E}[a + b\boldsymbol{Y}] \\ &= b(\mathbf{E}[\boldsymbol{X}] - \mathbf{E}[\boldsymbol{Y}]) \\ &= b(\mathbf{E}[\sum_{i=1}^{n} \alpha_i \boldsymbol{x}_i] - \mathbf{E}[\sum_{i=1}^{n} \alpha_i \boldsymbol{g}_i]) \\ &= 0 \end{aligned}$$

2. Let $\psi(t) = a + bt + ct^2$. Then

$$\begin{aligned} \mathbf{E}[\psi(\boldsymbol{X})] - \mathbf{E}[\psi(\boldsymbol{Y})] &= \mathbf{E}[a + b\boldsymbol{X} + c\boldsymbol{X}^2] - \mathbf{E}[a + b\boldsymbol{Y} + c\boldsymbol{Y}^2] \\ &= c(\mathbf{E}[(\sum_{i=1}^{n} \alpha_i \boldsymbol{x}_i)^2] - \mathbf{E}[(\sum_{i=1}^{n} \alpha_i \boldsymbol{g}_i)^2]) \\ &= 0 \end{aligned}$$

3. Let $\psi(t) = t^3$. Then

$$\begin{aligned} \mathbf{E}[\psi(\boldsymbol{X})] - \mathbf{E}[\psi(\boldsymbol{Y})] &= \mathbf{E}[(\sum_{i=1}^{n} \alpha_i \boldsymbol{x}_i)^3] - \mathbf{E}[(\sum_{i=1}^{n} \alpha_i \boldsymbol{g}_i)^3 \\ &= \sum_{i,j,k} \alpha_i \alpha_j \alpha_k \mathbf{E}[\boldsymbol{x}_i \boldsymbol{x}_j \boldsymbol{x}_k] - \sum_{i,j,k} \alpha_i \alpha_j \alpha_k \mathbf{E}[\boldsymbol{g}_i \boldsymbol{g}_j \boldsymbol{g}_k] \\ &= 0 \end{aligned}$$

At this point one might conjecture that all polynomials are nice functions. But it turns out it is not "quite" true as we see in the next example.

4. Let $\psi(t) = t^4$. Then

$$\begin{aligned}
\mathbf{E}[\psi(\boldsymbol{X})] - \mathbf{E}[\psi(\boldsymbol{Y})] &= \mathbf{E}[(\sum_{i=1}^{n} \alpha_i \boldsymbol{x}_i)^4] - \mathbf{E}[(\sum_{i=1}^{n} \alpha_i \boldsymbol{g}_i)^4] \\
&= \sum_{i,j,k,l} \alpha_i\alpha_j\alpha_k\alpha_l \mathbf{E}[\boldsymbol{x}_i\boldsymbol{x}_j\boldsymbol{x}_k\boldsymbol{x}_l] - \sum_{i,j,k,l} \alpha_i\alpha_j\alpha_k\alpha_l \mathbf{E}[\boldsymbol{g}_i\boldsymbol{g}_j\boldsymbol{g}_k\boldsymbol{g}_l] \\
&= \sum_{i=1}^{n} \alpha_i^4 - \sum_{i=1}^{n} \alpha_i^4 \mathbf{E}[\boldsymbol{g}_i^4]
\end{aligned}$$

Since the fourth moment of $N(0,1)$ is 3, we have that

$$\begin{aligned}
|\mathbf{E}[\psi(\boldsymbol{X})] - \mathbf{E}[\psi(\boldsymbol{Y})]| &\leq 2\sum_{i=1}^{n} \alpha_i^4 \\
&\leq 2\,(\max_i \alpha_i^2)(\sum_{i=1}^{n} \alpha_i^2) \\
&\leq \tau.
\end{aligned}$$

We now formally define “nice” functions.

Definition 1.2 *A function $\psi : \mathbb{R} \to \mathbb{R}$ is B-nice, $B \in \mathbb{R}^+$, if ψ is smooth and $|\psi''''(t)| \leq B$, $\forall t$.*

Remark 1.3 *By bounding the fourth derivative of a function at all points, we can see that we have an upper bound on the remainder term in the Taylor’s theorem. So if the bound on the fourth derivative of a function is small at all points, then we get a very good approximation for the function at any point by using the first four terms in the Taylor series expansion of the function.*

Idea 3: The final idea required to prove our main theorem is “hybridization” of random variables. Let $\boldsymbol{Z}_i = \alpha_1\boldsymbol{g}_1 + \cdots + \alpha_i\boldsymbol{g}_i + \alpha_{i+1}\boldsymbol{x}_1 + \cdots + \alpha_n\boldsymbol{x}_n$. So $\boldsymbol{Z}_0 = \boldsymbol{X}$ $(= \sum_{i=1}^{n} \alpha_i\boldsymbol{x}_i)$. And $\boldsymbol{Z}_n = \boldsymbol{Y}$ $(= \sum_{i=1}^{n} \alpha_i\boldsymbol{g}_i)$

We shall show that for any B-nice function ψ

$$|\mathbf{E}[\psi(\boldsymbol{Z}_{i-1}] - \mathbf{E}[\psi(\boldsymbol{Z}_i)]| \leq O(B\alpha_i^4) \;\; \forall i = 1 \cdot\cdot n$$

Then by telescoping together with triangle inequality we get

$$\begin{aligned}
|\mathbf{E}[\psi(\boldsymbol{X}) - \mathbf{E}[\psi(\boldsymbol{Y})]| &\leq \sum_{i=1}^{n} O(B\alpha_i^4) \\
&\leq O((\max_i \alpha_i^2)(\sum_{i=1}^{n} \alpha_i^2)) \\
&\leq O(B\tau)
\end{aligned}$$

We shall now recall Taylor’s theorem which we shall use in proving the next theorem.

Theorem 1.4 (Taylor's theorem) *For all smooth functions f and for any $r \in \mathbb{N}$, there exists $y \in [x, x+\epsilon]$, such that*

$$f(x+\epsilon) = f(x) + \epsilon f'(x) + \frac{\epsilon^2}{2!} f''(x) + \cdots + \frac{\epsilon^{r-1}}{(r-1)!} f^{(r-1)}(x) + \frac{\epsilon^r}{r!} f^{(r)}(y)$$

Theorem 1.5 *For all B-nice functions ψ*

$$|\mathbf{E}[\psi(\boldsymbol{Z}_{i-1})] - \mathbf{E}[\psi(\boldsymbol{Z}_i)]| \leq O(B\tau).$$

Proof: Write $\boldsymbol{R} = \alpha_1 \boldsymbol{g}_1 + \cdots + \alpha_{i-1}\boldsymbol{g}_{i-1} + \alpha_{i+1}\boldsymbol{x}_{i+1} + \cdots + \alpha_n \boldsymbol{x}_n$. Then $\boldsymbol{Z}_{i-1} = \boldsymbol{R} + \alpha_i \boldsymbol{x}_i$ and $\boldsymbol{Z}_i = \boldsymbol{R} + \alpha_i \boldsymbol{g}_i$ Note that $x_i, \boldsymbol{g}_i, \boldsymbol{R}$ are mutually independent. We want to bound $|\mathbf{E}[\psi(\boldsymbol{R} + \alpha_i \boldsymbol{x}_i)] - \mathbf{E}[\psi(\boldsymbol{R} + \alpha_i \boldsymbol{g}_i)]|$. Since ψ is a B-nice function we have $\psi''''(t) \leq B, \forall t$. This gives us the following

$$\forall t, \epsilon > 0,\ \psi(t+\epsilon) = \psi(t) + \psi'(t)\epsilon + \psi''(t)\frac{\epsilon^2}{2} + \frac{\psi'''(t)}{6}\epsilon^3 + \{\leq \frac{B}{24}\epsilon^4\}$$

Hence

$$\begin{aligned}
|\mathbf{E}[\psi(\boldsymbol{X})] - \mathbf{E}[\psi[\boldsymbol{Y}]]| &= |\mathbf{E}[\psi(\boldsymbol{R} + \alpha_i \boldsymbol{x}_i)] - \mathbf{E}[\psi(\boldsymbol{R} + \alpha_i \boldsymbol{g}_i)]| \\
&= |\mathbf{E}[\psi(R) + \psi'(R)(\alpha_i \boldsymbol{x}_i) + \psi''(R)\frac{(\alpha_i \boldsymbol{x}_i)^2}{2} + \frac{\psi'''(R)}{6}(\alpha_i \boldsymbol{x}_i)^3 + \{\leq \frac{B}{24}(\alpha_i \boldsymbol{x}_i)^4\}] \\
&- \mathbf{E}[\psi(R) + \psi'(R)(\alpha_i \boldsymbol{g}_i) + \psi''(R)\frac{(\alpha_i \boldsymbol{g}_i)^2}{2} + \frac{\psi'''(R)}{6}(\alpha_i \boldsymbol{g}_i)^3 + \{\leq \frac{B}{24}(\alpha_i \boldsymbol{g}_i)^4\}]| \\
&\leq \mathbf{E}[\{\frac{B}{24}(\alpha_i \boldsymbol{x}_i)^4\} + \{\frac{B}{24}(\alpha_i \boldsymbol{g}_i)^4\}] \\
&\leq \frac{B}{24}\alpha_i^4 + \frac{B}{24}3\alpha_i^4 \\
&\leq O(B\alpha_i^4) \\
&\leq O(B\tau)
\end{aligned}$$

□

Recall that we mentioned before that if the threshold function $\psi_{t_0}(t) = \begin{cases} 1 & \text{if } t < t_0 \\ 0 & \text{otherwise} \end{cases}$ and the absolute value function $\psi_2(t) = |t|$ fit in the definition of B-nice functions we have our Berry-Esseen theorem proved. We can see that they are not B-nice functions. However, they can be approximated by B-nice functions. We use this fact prove the Berry-Esseen theorem.

Claim 1.6 *$\forall t_o \in \mathbb{R}$ and $\forall \lambda, 0 < \lambda < \frac{1}{2}$, there exists a $O(\frac{1}{\lambda^4})$-nice function $\psi_{t_0,\lambda} : \mathbb{R} \to \mathbb{R}$ which approximates ψ_{t_0} in the following sense: $\psi_{t_0,\lambda} = 1$ for $t < t_0 - \lambda$; $\psi_{t_0,\lambda}(t) = 0$ for $t > t_0 + \lambda$ and $0 \leq \psi_{t_0,\lambda}(t) \leq 1$ for $|t - t_0| \leq \lambda$.*

We give a proof sketch for the first part of the Berry-Esseen theorem, $\forall t_0 \in \mathbb{R}$, $\mathbf{Pr}[\boldsymbol{X} < t_0] - \mathbf{Pr}[\boldsymbol{g} < t_0]| \leq O(\tau)$, where $\boldsymbol{X} = \sum\limits_{i=1}^{n} \alpha_i \boldsymbol{x}_i$, with $\boldsymbol{x}_i$'s as i.i.d uniform random ± 1 bits and $\boldsymbol{g} \sim N(0,1)$. Also $\sum_{i=1}^{n} \alpha_i^2 = 1$ and $\alpha_i^2 \leq \tau$, $\forall i \in [n]$.

Proof Sketch:

$$\mathbf{E}[\psi_{t_0-\lambda,\lambda}(\boldsymbol{X})] \leq \mathbf{Pr}[\boldsymbol{X} < t_0] \leq \mathbf{E}[\psi_{t_0+\lambda,\lambda}(\boldsymbol{X})]$$

By Berry-Essen theorem we have $\mathbf{E}[\psi_{t_0-\lambda,\lambda}(\boldsymbol{X})] = \mathbf{E}[\psi_{t_0-\lambda,\lambda}(\boldsymbol{g})] \pm O(\frac{\tau}{\lambda^4})$ and $\mathbf{E}[\psi_{t_0+\lambda,\lambda}(\boldsymbol{X})] = \mathbf{E}[\psi_{t_0+\lambda,\lambda}(\boldsymbol{g})] \pm O(\frac{\tau}{\lambda^4})$. But $\mathbf{E}[\psi_{t_0+\lambda,\lambda}(\boldsymbol{g})] = \mathbf{Pr}[\boldsymbol{g} < t_0 + \lambda]$, which is within $O(\lambda)$ of $\mathbf{Pr}[\boldsymbol{g} < t_0]$. Therefore we have

$$|\mathbf{Pr}[\boldsymbol{X} < t_0] - \mathbf{Pr}[\boldsymbol{g} < t_0]| \leq O(\frac{\tau}{\lambda^4}) + O(\lambda)$$

By taking $\lambda = \tau^{\frac{1}{5}}$, we have

$$|\mathbf{Pr}[\boldsymbol{X} < t_0] - \mathbf{Pr}[\boldsymbol{g} < t_0]| \leq O(\tau^{\frac{1}{5}}).$$