In this lecture we introduce the Chen-Stein method on Poisson approximation of a sequence of weakly correlated random variables.

11.1 Motivation

\section*{Scan statistics} Consider a matrix $y \in \mathbb{R}^{H \times W}$ with $HW$ random variables $y_{ij}$. We want to perform a hypothesis testing, with the null hypothesis $H_0$ being $y_{ij}$ i.i.d. distributed according to some known distribution $F$. Typically $F$ can be taken as a zero-mean Gaussian distribution $\mathcal{N}(0, 1)$. One particular test statistic is

$$\max_{i,j} \{T_{ij}\}_{i,j=1}^{H-h+1, W-w+1},$$

where $T_{ij}$ can be either the average or the maximum of $y_{ij}$ in a small scanning window of height $h$ and width $w$. Under the null hypothesis, for a fixed scanning window $T_{ij}$, the distribution of its average or its maximum can be easily worked out. However, the distribution of the maximum of all scan statistics $T_{ij}$ (or even its confidence interval) is difficulty to compute, because $T_{ij}$ are not independent from each other. Nevertheless, the correlation between the scan statistics are weak, and hence we expect them to behave close to independently distributed random variables.

\section*{Max correlation} Consider a matrix $X \in \mathbb{R}^{n \times p}$, where each columns of $X$, $X_i$, are i.i.d. distributed according to a multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ for some known parameters $\mu$ and $\Sigma$. We are interested in the following quantity

$$M = \max_{i_1 \neq i_2} \text{corr}(X_{i_1}, X_{i_2}),$$

where $\text{corr}(x, y)$ is the empirical correlation between two vectors $x$ and $y$, which is defined as

$$\text{corr}(x, y) = \frac{\sum_i x_i y_i}{\|x\|_2 \|y\|_2}.$$

Again, $\text{corr}(X_{i_1}, X_{i_2})$ are not independent from $\text{corr}(X_{i_1'}, X_{i_2'})$ as long as $\{i_1, i_2\} \cap \{i_1', i_2'\} \neq \emptyset$. However, we expect the mutual correlation is weak.

11.2 The Chen-Stein Poisson Approximation

We have the following classic result, which states that the sum of a sequence of weakly correlated random variables behave close to a Poisson random variable.
Theorem 11.1 ([AGG89]) Let \( \{X_\alpha\}_{\alpha \in I} \) be a sequence of (correlated) Bernoulli random variables indexed by \( \alpha \), where the index set \( I \) is countable. Suppose \( X_\alpha \sim \text{Bernoulli}(p_\alpha) \), \( W = \sum_\alpha X_\alpha \) and \( \mathbb{E}[W] = \sum_\alpha p_\alpha = \lambda \).

For each \( \alpha \in I \), fix its neighborhood \( B_\alpha \subseteq I \) and define
\[
\begin{align*}
  b_1 &= \sum_{\alpha \in I} \sum_{\beta \in B_\alpha} p_\alpha p_\beta, \\
  b_2 &= \sum_{\alpha \in I} \sum_{\beta \not\in B_\alpha \setminus \{\alpha\}} p_\alpha p_\beta, \\
  b_3 &= \sum_{\alpha \in I} \mathbb{E} \left[ \mathbb{E}[X_\alpha - p_\alpha | X_\beta : \beta \not\in B_\alpha] \right],
\end{align*}
\]

where \( p_{\alpha \beta} = \text{Pr}[X_\alpha = 1 \land X_\beta = 1] \). Note that if \( X_\alpha \) is independent of \( X_\beta \) for all \( \beta \) in the complement of \( B_\alpha \) then \( b_3 = 0 \). Suppose \( Z \sim \text{Poisson}(\lambda) \). We then have
\[
\|W - Z\|_{TV} \leq 2 \left( (b_1 + b_2) \frac{1 - e^{-\lambda}}{\lambda} + b_3 \min \left( 1, \frac{1}{\sqrt{\lambda}} \right) \right) \leq 2(b_1 + b_2 + b_3),
\]

where \( \|p - q\|_{TV} = \int |p(x) - q(x)| dx \) is the total variation between two probability distributions.

As a simple corollary, the following proposition characterizes the probability of \( W = 0 \) (i.e., no bad event happens), which is very useful in many applications.

Corollary 11.2 Assuming the same notations in Theorem 11.1. We then have
\[
\left| \text{Pr}(W = 0) - e^{-\lambda} \right| \leq (b_1 + b_2 + b_3) \cdot \frac{1 - e^{-\lambda}}{\lambda} \leq (b_1 + b_2 + b_3) \min \left( 1, \frac{1}{\sqrt{\lambda}} \right).
\]

11.3 Application: the birthday problem

Consider \( N \) people, each with birthday sampled uniformly at random from \( \{1, \cdots, 365\} \). We are interested in (approximately) computing the probability that no two people share the same birthday. Let \( \{X_{ij}\}_{1 \leq i < j \leq N} \) be random variables with \( X_{ij} = 1 \) if person \( i \) and person \( j \) has the same birthday and \( X_{ij} = 0 \) otherwise. Define \( W = \sum_{i < j} X_{ij} \). The event that no two people share the same birthday is equivalent to the event that \( W = 0 \).

For notational convenience define \( I = \{(i, j) : 1 \leq i < j \leq N\} \) and \( \lambda = \mathbb{E}[W] = \binom{N}{2} \frac{1}{d} \), where \( d = 365 \) is the number of days in a year. For \( \alpha = (i, j) \), define its neighborhood \( B_\alpha \) as
\[
B_\alpha = B_{ij} = \{(k, \ell) : \{k, \ell\} \cap \{i, j\} \neq \emptyset\}.
\]

Clearly, \( X_\alpha \) is independent of \( X_\beta \) for all \( \beta \notin B_\alpha \) and hence \( b_3 = 0 \). For the other two quantities \( b_1 \) and \( b_2 \), we have
\[
\begin{align*}
  b_1 &= |I| \cdot |B_\alpha| \cdot \frac{1}{d^2}, \\
  b_2 &= |I| \cdot (|B_\alpha| - 1) \cdot \frac{1}{d^2},
\end{align*}
\]
where in the last equation we used the fact that $X_\alpha$ and $X_\beta$ are pairwise independent for all $\alpha \neq \beta$ and hence $p_{\alpha\beta} = p_\alpha p_\beta = 1/d^2$. Consequently, applying Corollary 11.2 and noting that $|I| = \binom{N}{2}$ and $|B_\alpha| = 2(N-2)+1$, we have

$$\left| \Pr(W = 0) - \exp \left( -\binom{N}{2} \frac{1}{d^2} \right) \right| \leq \frac{1}{N} \binom{N}{2} \frac{1}{d^2} (4N - 7) = \frac{4N - 7}{d}.$$

References