Concentration Inequalities with Exchangeable Pairs

Sourav Chatterjee
Concentration inequalities

• What are concentration inequalities? From an application perspective, they give useful bounds on

\[ \mathbb{P}\{|f(X) - \mathbb{E}f(X)| \geq x\} \]

where \( X \) is some random variable, usually high dimensional, and \( f \) is a well behaved (usually Lipschitz) function.

• Useful in a variety of fields. Important tool in combinatorics, machine learning and theoretical computer science.

• Huge literature. Very deep results about functions of independent random variables.
• No single good method for dependent problems. Logarithmic Sobolev inequalities and their modifications have been successful. But explicit constants are hard to get.

• Same holds for transportation cost inequalities. Moreover, not very useful for discrete settings.

• Neither of the above methods is “probabilistic” in the true sense of the word. Rather, like Fourier analysis, they are analytical.

• We shall look for a “probabilistic” alternative through exchangeable pairs.
Stein’s method

- $X$, $X'$, and $Y$ are random variables on some general space. $(X, X')$ is an exchangeable pair. Want to bound $|\mathbb{E} f(X) - \mathbb{E} f(Y)|$.

- Stein’s method:

  (a) Find $F$ such that $F(x, y) = -F(y, x)$, and $\mathbb{E}(F(X, X')|X) = f(X) - \mathbb{E} f(X)$.

  (b) Construct $Z$ such that $\mathcal{L}(Z|Y = y) = \mathcal{L}(X'|X = y)$.

  (c) Use:

  $$\mathbb{E} f(Y) - \mathbb{E} f(X) = \mathbb{E}(\mathbb{E}(F(Y, Z)|Y))$$
  $$= \mathbb{E}(F(Y, Z)),$$

  and show that this is small by constructing $Y'$ such that $(Y, Y')$ is an exchangeable pair (so that $\mathbb{E}(F(Y, Y')) = 0$) and $(Y, Y')$ is “close” to $(Y, Z)$. 
• **Question:** Can we use exchangeable pairs to get concentration inequalities?

• Suppose we have \( f \) and \( F \) as before, with \( \mathbb{E} f(X) = 0 \). Observe that for any \( g \),

\[
\mathbb{E}(g(X)f(X)) = \mathbb{E}(g(X)F(X,X')).
\]

• **Exchangeability** \( \Rightarrow \)

\[
\mathbb{E}(g(X)F(X,X')) = \mathbb{E}(g(X')F(X',X)).
\]

• **Antisymmetry** \( \Rightarrow \)

\[
\mathbb{E}(g(X')F(X',X)) = -\mathbb{E}(g(X')F(X,X')).
\]
Combining, we get

\[ \mathbb{E}(g(X)f(X)) = \frac{1}{2} \mathbb{E}((g(X) - g(X'))F(X, X')). \]

Taking \( g = f \), we have the variance formula:

\[ \text{Var}(f(X)) = \frac{1}{2} \mathbb{E}((f(X) - f(X'))F(X, X')). \]

Example: Magnetization in Curie-Weiss model.
Curie-Weiss model

• Configuration space of $n$ spins: $\{-1, 1\}^n$.

• Probability mass on this space:

\[ p_\beta(\sigma) = Z(\beta)^{-1} e^{\frac{\beta}{n} \sum_{i<j} \sigma_i \sigma_j}. \]

Here $\beta$ is a parameter and $Z(\beta)$ is the normalizing constant.

• Magnetization: $m(\sigma) = \frac{1}{n} \sum_{i=1}^{n} \sigma_i$.

• Well-known: $m(\sigma) \approx \tanh(\beta m(\sigma))$ with high probability.
• Say two configurations are “neighbors” if they differ at exactly one site.

• Get $\sigma'$ from $\sigma$ by taking one step in the Metropolis chain along a random neighbor.

• Set $F(\sigma, \sigma') = n(m(\sigma) - m(\sigma'))$. Then

$$f(\sigma) := \mathbb{E}(F(\sigma, \sigma')|\sigma)$$

$$= m(\sigma) - \frac{1}{n} \sum_{i=1}^{n} \tanh(\beta m_i(\sigma)),$$

where $m_i(\sigma) = \frac{1}{n} \sum_{j \neq i} \sigma_j$.

• Then $|F(\sigma, \sigma')| \leq 2$ and $|f(\sigma) - f(\sigma')| \leq 2(1 + \beta)/n$. 
• Thus,

\[
\text{Var} \left( m(\sigma) - \frac{1}{n} \sum_{i=1}^{n} \tanh(\beta m_i(\sigma)) \right) \\
= \frac{1}{2} \mathbb{E} \left( (f(\sigma) - f(\sigma')) F(\sigma, \sigma') \right) \\
\leq \frac{2(1 + \beta)}{n}.
\]

• Finally, note that \(|m_i(\sigma) - m(\sigma)| \leq 1/n\). Combining, we get

\[
\mathbb{E}(m(\sigma) - \tanh(\beta m(\sigma)))^2 \leq \frac{2(1 + \beta)}{n} + \frac{\beta^2}{n^2}.
\]
Concentration inequalities

We have the following moment and tail inequality version of the earlier variance formula:

**Theorem 1** Define

\[ v(X) := \frac{1}{2} \mathbb{E}(|(f(X) - f(X'))F(X, X')| |X). \]

Then for any positive integer \( p \), we have

\[ \|f(X) - \mathbb{E}f(X)\|_{2p}^2 \leq (2p - 1)\|v(X)\|_p. \]

Moreover, if \( |v(X)| \leq C \) almost surely, then

\[ \mathbb{P}\{|f(X) - \mathbb{E}f(X)| \geq x\} \leq 2e^{-x^2/2C} \]

for each \( x \geq 0 \).

(Recall that: \((X, X')\) is an exchangeable pair, \( F \) is antisymmetric, and \( \mathbb{E}(F(X, X') | X) = f(X) - \mathbb{E}f(X) \).)
Proof

• Let \( \varphi(\theta) = \mathbb{E}(e^{\theta f(X)}) \).

• Using the same trick as before, we have
  \[
  \varphi'(\theta) = \mathbb{E}(e^{\theta f(X)} f(X)) = \frac{1}{2} \mathbb{E}((e^{\theta f(X)} - e^{\theta f(X')})F(X, X')).
  \]

• Using \(|e^x - e^y| \leq \frac{1}{2}(e^x + e^y)|x - y|\), we get
  \[
  \varphi'(\theta) \leq |\theta| \mathbb{E}(e^{\theta f(X)} v(X)),
  \]
  where
  \[
  v(X) = \frac{1}{2} \mathbb{E}
  \left( |(f(X) - f(X'))F(X, X')| |X| \right).
  \]

• Similarly,
  \[
  \mathbb{E}(f(X)^{2p}) \leq (2p - 1) \mathbb{E}(f(X)^{2p-2}v(X))
  \]
  and apply Hölder’s inequality.
Before we give an example, let us state a refinement of the previous tail bound:

**Theorem 2** Suppose \( v(X) \leq B f(X) + C \) a.s. Then

\[
P\{|f(X) - E f(X)| \geq x\} \leq 2e^{-x^2/(2C+2Bx)}
\]

for each \( x \geq 0 \).

While the moment bounds are generalizations of the Burkholder-Gundy-Davis inequalities, and the first tail bound generalizes the Hoeffding inequality, the above can be seen as an exchangeable pair version of Bernstein’s inequality.
Example: Random permutations

**Proposition 1** Let \( \{a_{ij}\} \) be an \( n \) by \( n \) array of elements of \([0, 1]\). Let \( \pi \) be a random (uniform) permutation of \( \{1, \ldots, n\} \), and let \( W = \sum_{i=1}^{n} a_{i\pi(i)} \). Then for any \( x \geq 0 \),

\[
P\{|W - \mathbb{E}(W)| \geq x\} \leq 2e^{-x^2/(4\mathbb{E}(W)+2x)}.
\]

For instance, if \( a_{ij} = \mathbb{I}\{i = j\} \), then \( W \) is the number of fixed points of \( \pi \) and \( \mathbb{E}(W) = 1 \).

**Sketch of Proof:**

- Obtain \( \pi' \) by applying a random transposition to \( \pi \). Let \( W' = W(\pi') \).

- Let \( F(\pi, \pi') = \frac{1}{2}n(W - W') \).

- Easy: \( \mathbb{E}(F(\pi, \pi')|\pi) = W - \mathbb{E}(W) =: f(\pi). \)
Proof (contd.):

• Using $0 \leq a_{ij} \leq 1$, we can show

$$v(\pi) = \frac{1}{2} \mathbb{E} \left( |(f(\pi) - f(\pi'))F(\pi, \pi')| \right|_{\pi}$$

$$= \frac{n}{4} \mathbb{E} ((W - W')^2 | \pi)$$

$$\leq f(\pi) + 2\mathbb{E}(W).$$

• Use Theorem ??.
• **Next question:** Given $f$, how to obtain $F$, or at least, information about $F$?

• **A coupling method:** Construct a coupling $\{(X_k, X'_k)\}_{k\geq 0}$ with the following properties:

1. The marginal chains are Markovian from the kernel defined by $(X, X')$ and are weakly ergodic.

2. $(X_0, X'_0)$ has the same distribution as $(X, X')$.

3. For each $k$, the distribution of $X_k$ given $(X_0, X'_0)$ depends only on $X_0$, and the distribution of $X'_k$ given $(X_0, X'_0)$ depends only on $X'_0$. 

15
Then we have the following theorem:

**Theorem 3** Suppose $\sum_{k=0}^{\infty} \mathbb{E}|f(X_k) - f(X'_k)| < \infty$. If we define

$$F(X_0, X'_0) := \sum_{k=0}^{\infty} \mathbb{E}(f(X_k) - f(X'_k)|X_0, X'_0)$$

then $F$ is antisymmetric and $\mathbb{E}(F(X, X')|X) = f(X) - \mathbb{E}f(X)$.

Tradeoff: Size of steps vs. coupling time.
Proof:

• Antisymmetry is easy.

• Let \( f_k(X_0) = \mathbb{E}(f(X_k)|X_0) \).

• Note that
  \[
  \mathbb{E}(f(X'_k)|X_0) = \mathbb{E}(f_k(X'_0)|X_0) = f_{k+1}(X_0).
  \]

• Thus,
  \[
  \sum_{k=0}^{N} \mathbb{E}(f(X_k) - f(X'_k)|X_0) = f(X_0) - f_{N+1}(X_0).
  \]

• Using given conditions, get \( f_{N+1}(X_0) \to \mathbb{E}f(X) \).
Combining the coupling result with the exchangeable pair tail bound theorem, we get the following:

**Theorem 4** *For any positive integer* $p$, *we have*

\[
\|f(X) - \mathbb{E}f(X)\|_{2p}^2 \\
\leq \frac{2p - 1}{2} \sum_{k=0}^{\infty} \|\mathbb{E}((f(X_0) - f(X_0'))(f(X_k) - f(X_k'))|X_0)\|_p.
\]

*Moreover, if*

\[
\frac{1}{2} \sum_{k=0}^{\infty} \mathbb{E}(|(f(X_0) - f(X_0'))(f(X_k) - f(X_k'))|X_0) \leq C \text{ a.s.},
\]

*then*

\[
P\{|f(X) - \mathbb{E}f(X)| \geq x\} \leq e^{-x^2/2C}
\]

*for each* $x \geq 0$. 

18
• When $X$ is a random $n$-vector with independent components, the natural Markov step is to choose a coordinate uniformly at random, and replace by an independent copy.

• The natural coupling is to choose the same coordinate and substitute the same value in the second chain as in the first.

• Then $X_0$ and $X'_0$ differ at one coordinate, and so the coupling time is like $n$.

• All the while, $X_k$ and $X'_k$ differ at most at one coordinate.
• With this coupling, the moment bounds can be combined to give the exponential Efron-Stein of Boucheron, Lugosi & Massart (Ann. Probab. 2003).

• For independent variables, the moment inequalities themselves are the generalized Burkholder inequalities of Boucheron, Bousquet, Lugosi & Massart (To appear in the Annals.)

• They give many interesting applications.

• Talagrand’s famous convex distance inequality follows from the exponential Efron-Stein.
Concentration under weak dependence

- Configuration space: $\Omega^n$, endowed with a probability $\mu$.

- Convention: $\bar{x}^i = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$.

- $\mu_i(\cdot|\bar{x}^i)$ denotes conditional distribution of the $i^{th}$ coordinate given all others.

- Assume: For each $i, x, y$,

$$\|\mu_i(\cdot|\bar{x}^i) - \mu_i(\cdot|\bar{y}^i)\|_{TV} \leq \sum_{j=1}^{n} a_{ij} \mathbb{I}\{x_j \neq y_j\},$$

where $a_{ij} \geq 0$ for all $i, j$ and $a_{ii} = 0$ for all $i$. The matrix $A := (a_{ij})$ is called "Do-brushin's interdependence matrix".
We have the following result:

**Theorem 5** Suppose \( f : \Omega^n \to \mathbb{R} \) satisfies

\[
|f(x) - f(y)| \leq \sum_{i=1}^{n} c_i \mathbb{I}\{x_i \neq y_i\}.
\]

and also \( \mathbb{E}|f(X)| < \infty \). If \( \|A\|_2 < 1 \), we have

\[
P\{|f(X) - \mathbb{E}f(X)| \geq x\} \leq 2 \exp\left(-\frac{(1 - \|A\|_2)x^2}{\sum_i c_i^2}\right)
\]

for each \( x \geq 0 \).

- Note: The condition \( \|A\|_2 < 1 \), where \( \|A\|_2 \) is the \( L^2 \)-norm of \( A \), is a relaxed version of “Dobrushin’s condition of weak dependence” (which demands \( \|A\|_\infty < 1 \)).

- The bound depends on \( n \) only through \( \sum c_i^2 \), thus giving satisfactory bounds for lower dimensional marginals also.
Further notes:

- Stroock and Zegarlinski, in a series of papers in 1992, proved the equivalence of Dobrushin’s condition and the log-Sobolev inequality for spin systems on a lattice. The lecture notes by Guionnet and Zegarlinski have a readable version.

- Their treatment seems to be heavily using the lattice. I don’t know if it’s generalizable.

- Explicit constants are very hard or impossible to get for these log Sobolev inequalities.

- Marton (*Ann. Probab.* 2004) has the Wasserstein distance version of our theorem, under the added assumption that the conditionals satisfy log-Sobolev inequalities.
Brief sketch of proof:

- The natural Markov chain is the Gibbs' sampler: Choose a coordinate $i$ at random and generate from the conditional distribution.

- Coupling rule: Choose the same coordinate for $X_k$ and $X_k'$, and generate $X_{k+1,i}, X'_{k+1,i}$ so that

$$
\mathbb{P}\{X_{k+1,i} \neq X'_{k+1,i} | X_k, X'_k\} = \| \mu_i(\cdot | \bar{X}_k^i) - \mu_i(\cdot | \bar{X}'_k^i) \|_{TV}.
$$

- Proceed by induction to get bounds.
Example: Graph colorings

- Suppose $G = (V, E)$ is a graph with $n$ vertices and maximum degree $r$.

- Let $X = (X_i, i \in V)$ be a coloring of $G$ chosen uniformly from the set of all proper colorings (i.e. no two adjacent vertices have the same color) with $k$ colors.

- Let $f : \{1, \ldots, k\}^G \to \mathbb{R}$ be a map satisfying $|f(x) - f(y)| \leq \sum_{i=1}^{n} c_i \mathbb{I}\{x_i \neq y_i\}$.

- If $k > 2r$, then it’s easy to show that we can take $a_{ij} = 1/(k - r)$ for $(i, j) \in E$ and 0 otherwise. Thus,

$$
P\{\left|f(X) - \mathbb{E}f(X)\right| \geq x\} \leq 2 \exp\left(-\frac{k-2r}{k-r} \frac{x^2}{\sum_{i=1}^{n} c_i^2}\right).$$
Example: Densities on $[-1, 1]^n$.

Suppose we have a product measure $\nu^n$ on $[-1, 1]^n$, and suppose $\mu$ has density $Z^{-1}e^{H(x)}$ with respect to $\nu^n$.

**Proposition 2** For each pair $(i, j)$ with $i \neq j$, define

$$a_{ij} := 4 \sup \left| \frac{\partial^2 H}{\partial x_i \partial x_j} \right|$$

and let $a_{ii} = 0$ for each $i$. Then for each $i$ and $x, y \in [-1, 1]^n$, we have

$$d_{TV}(\mu_i(\cdot | \bar{x}^i), \mu_i(\cdot | \bar{y}^i)) \leq \sum_{j=1}^{n} a_{ij} \mathbb{1}\{x_j \neq y_j\}.$$

This covers, for example, two famous models: The Ising model and the xy model. The Dobrushin condition is satisfied at high temperatures.
Selected references:


