Concentration inequalities by Stein couplings

Daniel Paulin

National University of Singapore, Department of Statistics and Applied Probability

May 21, 2015

Workshop on New Directions in Stein's Method, Singapore

Joint work with Adrian Röllin.

Main idea

Definition 1

Let (W, W', G) *be a coupling of square integrable random variables. We call* (W, W', G) *a Stein coupling if*

$$\mathbb{E}\{Gf(W') - Gf(W)\} = \mathbb{E}\{Wf(W)\},$$

for all functions for which the expectation exists.

▷ With the choice $f(x) = e^{\theta x}$, we have

$$\begin{aligned} m'(\theta) &= \mathbb{E}\{Wf(W)\} = \mathbb{E}\{Gf(W') - Gf(W)\} \\ &= \mathbb{E}\left\{G\left(e^{\theta W'} - e^{\theta W}\right)\right\}. \end{aligned} \tag{1}$$

▷ This can be bounded using information about the typical size of G and $W - W'$, and a bound on $m'(\theta)$ leads to concentration inequalities.

Comparison with the literature

▷ The case when

- $W = f(X), W' = f(X')$, and (X, X') is an exchangeable pair, and
- $G = -F(X, X')$ such that F is an antisymmetric function

was studied by Chatterjee (2007), and further extended in Chatterjee & Dey (2010).

▷ Chatterjee (2012) proves concentration inequalities using a non-exchangeable coupling construction, that is not a Stein-coupling, but similarly implies bounds on the moment generating function.

▷ Application: sharp bounds for the number of triangles in an Erdős-Rényi graph.

▷ However, his main theorem is optimised for this particular problem, and it is not applicable to our examples.

▷ Ghosh & Goldstein (2011) and Goldstein & Islak (2013) use size biasing and zero biasing to obtain concentration inequalities.

A Stein coupling for sums, motivated by local dependence

▷ Suppose that $X_1, \ldots, X_n$ are dependent random variables, and

$$W = X_1 + \ldots + X_n.$$

▷ Let

$$G = n \cdot X_I,$$

and I be uniformly distributed on $[n] := \{1, \ldots, n\}$.

▷ Suppose that we can construct W' such that

- $\mathbb{E}(W') = 0$, and
- W' is independent of X_I.

▷ Then (W, W', G) is a Stein coupling.

▷ Such couplings can exist even when $X_1, \ldots, X_n$ are not defined as functions of independent random variables, a situation that is difficult to handle with other methods in the literature of concentration inequalities.

Result: Proposition 1

▷ Let

$$G^{(-)} := \text{ess sup}(G) - G. \tag{2}$$

Proposition 1

If W and W' have the same distribution, then for any $\theta \in \mathbb{R}$,

$$|m'(\theta)| \leq \mathbb{E}\left(|\theta| G^{(-)} |W - W'| \left(\frac{e^{\theta W} + e^{\theta W'}}{2}\right)\right). \tag{3}$$

Proof.

We can bound $\mathbb{E}\left\{G\left(e^{\theta W'} - e^{\theta W}\right)\right\}$ by the right hand side using the fact that W and W' has the same distribution, and the inequality $|e^x - e^y| \leq \frac{e^x + e^y}{2} \cdot |x - y|$. □

▷ The following lemma can be used together with Proposition 1 to obtain concentration inequalities.

Lemma 1

Let W be a centered random variable with moment generating function $m(\theta)$. Let $C, D \geq 0$, suppose that $m(\theta)$ is finite, and continuously differentiable in $[0, 1/C)$, and satisfies

$$m'(\theta) \leq C\theta m'(\theta) + D\theta m(\theta).$$

Then for $0 \leq \theta < 1/C$,

$$\log(m(\theta)) \leq \frac{D\theta^2}{2(1 - C\theta)}, \tag{4}$$

and for every $t \geq 0$,

$$\mathbb{P}(W \geq t) \leq \exp\left(-\frac{t^2}{2(D + Ct)}\right). \tag{5}$$

Example: Large subgraphs of huge graphs

▷ Consider a fixed graph with N vertices, called *host graph*.

▷ Vertices of the graph: $[N] := \{1, \ldots, N\}$.

▷ Edges of the graph: $(E_{i,j})_{1\leq i<j\leq N}$.

▷ Graph: $\mathcal{G} := ([N], (E_{i,j})_{1\leq i<j\leq N})$.

▷ Let $I(1), \ldots, I(n)$ be random variables chosen from $[N]$ by sampling without replacement, uniformly from the $N \cdot \ldots \cdot (N - n + 1)$ possibilities.

▷ Consider a *random subgraph* with vertices $I(1), \ldots, I(n)$, denoted $\mathcal{H} := (\{I(1), \ldots, I(n)\}, (E_{I(i),I(j)})_{1\leq i<j\leq n}))$.

▷ *A natural question*: if $\mathcal{F}$ a small fixed subgraph with k vertices, then how many copies of $\mathcal{F}$ are in our subgraph $\mathcal{H}$, and how is this related to the total number of such copies in the host graph $\mathcal{G}$?

▷ Let $N_{\mathcal{F}}(\mathcal{H})$ denote the number of *full copies* of a fixed graph $\mathcal{F} := \{[k], (F_{i,j})_{1\le i<j\le k}\}$ in $\mathcal{H}$.
▷ The following proposition shows a concentration inequality for this quantity, in terms of $N_{\mathcal{F}}(\mathcal{G})$, the number of full copies of $\mathcal{F}$ in the host graph $\mathcal{G}$.

Theorem 1

For any $t \ge 0$, we have

$$\begin{aligned}&\mathbb{P}\left(|N_{\mathcal{F}}(\mathcal{H}) - \mathbb{E}(N_{\mathcal{F}}(\mathcal{H}))| \ge t\right)\\ &\le 2\exp\left(-\frac{t^2}{2k^2n^{k-1}\cdot\mathbb{E}(N_{\mathcal{F}}(\mathcal{H})) + k^2n^{k-1}t}\right),\end{aligned}$$

where $\mathbb{E}(N_{\mathcal{F}}(\mathcal{H})) = N_{\mathcal{F}}(\mathcal{G}) \cdot \frac{n(n-1)\dots(n-k+1)}{N(N-1)\dots(N-k+1)}$.

▷ This theorem can be viewed as a non-asymptotic law of large numbers.
▷ When N and n are large, and k is small, and $\mathcal{F}$ is quite frequent in $\mathcal{G}$ in the sense that $N_{\mathcal{F}}(\mathcal{G}) = \mathcal{O}(N^k)$, then $\mathbb{E}(N_{\mathcal{F}}(\mathcal{H})) = \mathcal{O}(n^k)$, while the typical deviation of $N_{\mathcal{F}}(\mathcal{H})$ is of $\mathcal{O}(kn^{k-1/2})$.
▷ This implies that $N_{\mathcal{F}}(\mathcal{H})$ is concentrated around its mean, which is determined by $\mathcal{G}$.
▷ Thus we can read the structure of $\mathcal{G}$, in the sense of subgraph frequencies, and make small error with high probability, from just one large sample $\mathcal{H}$.

Proof of Theorem 1.

$\triangleright$ Firstly, we construct a Stein coupling (W, W', G).
$\triangleright$ W will correspond to $N_{\mathcal{F}}(\mathcal{H}) - \mathbb{E}N_{\mathcal{F}}(\mathcal{H})$.
$\triangleright$ For notational simplicity, we define W' first, then G and finally W.
$\triangleright$ Let $I'(1), \dots, I'(n)$ be sampled without replacement from $[N]$, and define W' as the centered version of the number of full copies of $\mathcal{F}$ in the subgraph $\mathcal{H}'$ of $\mathcal{G}$ with vertices $I'(1), \dots, I'(n)$.
$\triangleright$ Let $J(1), J(2), \dots, J(k)$ be sampled without replacement from $[N]$, independently of $I'(1), \dots, I'(n)$, and let $\mathcal{G}_J$ be the subgraph of $\mathcal{G}$ with these vertices, and

$$G := -n \cdot \ldots \cdot (n-k+1) \cdot \left(\mathbb{1}\left[\mathcal{G}_J = \mathcal{F}\right] - \mathbb{P}\left[\mathcal{G}_J = \mathcal{F}\right]\right),$$

$\triangleright$ This is a rescaled, centered version of the indicator function corresponding to whether the subgraph of $\mathcal{G}$ with vertices $J(1), \dots, J(k)$ equals to $\mathcal{F}$. $\square$

▷ Because of the independence, it follows that

$$\mathbb{E}(G|W') = 0.$$

▷ We define $I(1), \ldots, I(n)$ as follows. First, set

$$I(1) := I'(1), \ldots, I(n) := I'(n).$$

▷ Whenever an element of the sequence $I(1), \ldots, I(n)$ is also a member of the sequence $J(1), \ldots, J(k)$, we mark it in both sequences.

▷ Suppose that there are r non-marked elements left in the sequence $J(1), \ldots, J(k)$.

▷ We choose r elements at random from the non-marked elements of $I(1), \ldots, I(n)$, and replace them with the corresponding non-marked element of $J(1), \ldots, J(k)$.

▷ This ensures that the sequence $J(1), \ldots, J(k)$ is distributed as if it were sampled without replacement from $I(1), \ldots, I(n)$.

$\triangleright$ Let $\mathcal{H}$ be the subgraph of $\mathcal{G}$ with vertices $I(1), \ldots, I(n)$, and W be the centered version of the number of full copies of $\mathcal{F}$ in $\mathcal{H}$.
$\triangleright$ Then $\mathbb{E}(G|W) = -W$, thus (W, W', G) is a Stein coupling.
$\triangleright$ Here W' and W have the same distribution (also exchangeable). Moreover, there are at most k indices i in $[n]$ such that $I(i)$ differs from $I'(i)$, therefore

$$|W - W'| \leq n \cdot \ldots \cdot (n-k+1) - (n-k) \cdot \ldots \cdot (n-2k+1) \leq k^2 n^{k-1}.$$

$\triangleright$ The result now follows from Proposition 1 and Lemma 1.

Result: Proposition 2

Proposition 2

Let (W, W', G) be a Stein coupling. Let

$$G^{(-)} := \mathrm{ess\ sup}(G) - G, \tag{6}$$

where $\mathrm{ess\ sup}(G)$ *denotes the supremum of* G *in the almost sure sense. Suppose that* W *and* W' *have the same distribution. Suppose that* $W_{\max}$ *and* $W_{\min}$ *are random variables such that* $|W - W'| \leq W_{\max} - W_{\min}$*, and conditioned on some* σ*-field* $\mathcal{F}$*,* G *is independent of* $W_{\max} - W_{\min}$ *and* W'*. Suppose that* $W_{\max} - W_{\min} \leq M < \infty$ *almost surely. Then*

$$m'(\theta) \leq \mathbb{E}\left(\mathbb{E}\left(G^{(-)} \middle| \mathcal{F}\right)\left(e^{\theta(W_{\max}-W_{\min})} - 1\right)e^{\theta W'}\right) \text{ for } \theta > 0, \text{ thus} \tag{7}$$

$$m'(\theta) \leq \mathbb{E}\left(2\theta\mathbb{E}\left(G^{(-)} \middle| \mathcal{F}\right)(W_{\max} - W_{\min})e^{\theta W'}\right) \text{ for } 0 \leq \theta \leq 1/M, \text{ and} \tag{8}$$

$$m'(\theta) \geq \mathbb{E}\left(\theta\mathbb{E}\left(G^{(-)} \middle| \mathcal{F}\right)(W_{\max} - W_{\min})e^{\theta W'}\right) \text{ for } \theta < 0. \tag{9}$$

▷ The following lemma allows us to bound expectations of the form $\mathbb{E}(e^{\theta W} V)$.

Lemma 2 (Massart (2000))

For real valued random variables V and W, any $L > 0$, for every $\theta \in \mathbb{R}$, we have

$$\mathbb{E}(e^{\theta W} V) \leq L^{-1} \log \mathbb{E}(e^{LV}) m(\theta) + L^{-1} \theta m'(\theta) - L^{-1} m(\theta) \log(m(\theta)),$$

if the expectations on both sides exist.

Example: Number of edges in geometric random graphs

$\triangleright$ We define a $\mathrm{Geo}(n, c)$ as follows.
$\triangleright$ Let $\Omega = [0, 1]^2$, and $X_1, \ldots, X_n$ be i.i.d. uniform in Ω.
$\triangleright$ Define the distance function $d : \Omega^2 \to \mathbb{R}_+$ as the torus distance between two points (this assumption is made to avoid edge effects).
$\triangleright$ For some $c > 0$, we put an edge between two points X_i and X_j if their distance is less than c.
$\triangleright$ We call the resulting graph $\mathrm{Geo}(n, c)$, the *random geometric graph*.

Theorem 2

Denote by $\mathcal{E}$ *the number of edges in the geometric random graph* $\mathrm{Geo}(n, c)$*. Let*

$$C_L := \sqrt{6\pi}nc, \quad D_L := 12(\log(1/c) + nc^2\pi)n^2c^2\pi$$
$$C_U := \max(\sqrt{12\pi}nc, 2n), \quad D_U := 24(\log(1/c) + nc^2\pi)n^2c^2\pi.$$

Then for any $t \geq 0$*,*

$$\mathbb{P}(\mathcal{E} - \mathbb{E}(\mathcal{E}) \geq t) \leq \exp\left(-\frac{t^2}{2(D_U + C_U t)}\right), \text{ and}$$
$$\mathbb{P}(\mathcal{E} - \mathbb{E}(\mathcal{E}) \leq -t) \leq \exp\left(-\frac{t^2}{2(D_L + C_L t)}\right).$$

▷ Applying McDiarmid's bounded differences inequalities would only give a concentration inequality of order $\exp(-t^2/n^3)$, independent of c. Our result depends on c, thus it is better when c is much smaller than 1.

Proof of Theorem 2.

▷ Denote by $\mathcal{E}_{i,j}$ the indicator function of the edge between the points X_i and X_j, then the total number of edges is

$$\mathcal{E} = \sum_{1 \le i < j \le n} \mathcal{E}_{i,j}.$$

▷ We have $\mathbb{E}(\mathcal{E}_{i,j}) = c^2\pi$, so $\mathbb{E}(\mathcal{E}) = \binom{n}{2} c^2\pi$.
▷ Let I and J be random indices such that $I < J$, uniformly chosen among the $\binom{n}{2}$ possibilities.
▷ Let

$$G := \binom{n}{2}\left(-\mathcal{E}_{I,J} + c^2\pi\right),$$

then

$$G^{(-)} = \binom{n}{2}\mathcal{E}_{I,J}.$$

▷ Let $W = \mathcal{E} - \mathbb{E}(\mathcal{E})$. We create W' by replacing X_I and X_J by an independent copy and evaluating W on the resulting graph.

▷ Let $\mathcal{E}_{\max}$ be the maximum number of edges in the geometric random graph that only differs from our graph in X_I and X_J (i.e. we move them to the most dense areas).
▷ Similarly, let $\mathcal{E}_{\min}$ be the number of edges of the graph created by removing X_I and X_J.
▷ Let $W_{\max} := \mathcal{E}_{\max} - \mathbb{E}(\mathcal{E})$, and $W_{\min} := \mathcal{E}_{\min} - \mathbb{E}(\mathcal{E})$.
▷ Conditions of Proposition 2 hold if $\mathcal{F}$ is σ-field generated by I, J. For $\theta < 0$,

$$\begin{aligned} m'(\theta) &\geq \mathbb{E}\left(\theta\mathbb{E}(G^{(-)}|\mathcal{F})(W_{\max} - W_{\min})e^{\theta W'}\right) \\ &\geq \theta\binom{n}{2}c^2\pi \cdot \mathbb{E}\left((W_{\max} - W_{\min})e^{\theta W'}\right). \end{aligned}$$

▷ Moreover, we have

$$W_{\max} - W_{\min} \leq 2 \cdot \text{ max number of points in a circle of size } c.$$

▷ The square can be cut into $1/(4c^2)$ small squares of edge length $2c$.
▷ By putting a circle of radius c centered in the middle of each square and on the vertices of each square, we cover the original square with $1/(2c^2)$ circles.
▷ Since any circle of radius c can cross at most 6 of these circles, we have

$$W_{\max} - W_{\min} \leq 12 \cdot \text{ max no. of points in a circ. among the } 1/(2c^2) \text{ circ.}$$

▷ Since the number of points in a circle of radius c is just the sum of n independent Bernoulli random variables with parameter $c^2\pi$, we have that for any $L > 0$,

$$\mathbb{E}\left(e^{L(W_{\max}-W_{\min})}\right) \leq \frac{1}{2c^2}\left(1 - c^2\pi + c^2\pi \cdot e^{12L}\right)^n,$$

and the results follow by Lemma 2 and Proposition 2. $\square$

Result: Proposition 3

Proposition 3

Suppose that $W \geq W'$ almost surely. Then for any $\theta \geq 0$,

$$m'(\theta) = \mathbb{E}(-G(e^{\theta W} - e^{\theta W'})) \leq \mathbb{E}(\theta G_{-}(W - W') \cdot e^{\theta W}). \tag{10}$$

Similarly, if $W' \geq W$ almost surely, then for any $\theta \leq 0$,

$$m'(\theta) = \mathbb{E}(-G(e^{\theta W} - e^{\theta W'})) \geq \mathbb{E}(\theta G_{+}(W' - W) \cdot e^{\theta W}). \tag{11}$$

Here $G_{-} := -G \cdot \mathbb{1}[G < 0]$ and $G_{+} := G \cdot \mathbb{1}[G > 0]$ denotes the negative, and positive parts of G.

$\triangleright$ Note that if $\mathbb{E}(G|W') = 0$, then we can shift W' by a constant and ensure that the conditions of this theorem hold.

Example: Isolated vertices in Erdős-Rényi graphs

▷ Let $G(n, p)$ be an Erdős-Rényi graph, with edges $X := (X_{i,j})_{1 \leq i < j \leq n}$ being i.i.d. Bernoulli random variables with parameter p. ▷ Denote the number of its isolated vertices (i.e. the vertices with zero incurring edges) by $\mathcal{I}(X)$. Then the following proposition bounds the lower tail of $\mathcal{I}(X)$.

Proposition 4

For any $t \geq 0$, we have

$$\mathbb{P}(\mathcal{I}(X) \leq \mathbb{E}(\mathcal{I}(X)) - t) \leq \exp\left(-\frac{t^2}{4n(1-p)^{n-1}}\right). \tag{12}$$

Remark 1

Ghosh et al. *(2011) have shown the same bound using size biasing.*

Proof.

▷ $\mathbb{E}(\mathcal{I}(X)) = n(1-p)^{n-1}$, thus we set

$$W := \mathcal{I}(X) - n(1-p)^{n-1}.$$

▷ X' is defined picking a vertex I uniformly from $[n]$, and removing all the edges connected to it.

$$W' := \mathcal{I}(X') - n(1-p)^{n-1}.$$

$$G := -n\mathbb{1}[I \text{ is an isolated vertex}] + n(1-p)^{n-1}.$$

▷ Then (W, W', G) is a Stein coupling, $\mathbb{E}(G|W') = 0$, and $W' \geq W$ almost surely.
▷ From Proposition 3, we obtain that for $\theta < 0$,

$$m'(\theta) \geq \mathbb{E}(G_+\theta(W' - W)e^{\theta W}) \geq n(1-p)^{n-1}\theta\mathbb{E}((W' - W)e^{\theta W}).$$

$\triangleright$ Now we are left to bound $\mathbb{E}(W' - W|W)$. We will show that for any graph X,

$$\mathbb{E}(W' - W|W) \leq 2.$$

$\triangleright$ Here $W' - W$ expresses the number of new isolated vertices created by erasing all of the edges of a randomly picked vertex from X. $\triangleright$ This operation can only create new isolated vertices from those that only had one incurring edge.
$\triangleright$ Such vertices are organised into groups of two (two vertices are connected to each other and isolated from the rest) or groups of $k \geq 3$ ($k - 1$ vertices have their only edge connected to the kth vertex, which we call *root vertex*).
$\triangleright$ Let $N_2(X)$ denote the number of groups of two, and N_k denote the number of groups of $3 \leq k \leq n$. Since the total number of vertices n, we must have $\sum_{k\geq 2} kN_k \leq n$.

$\triangleright$ If we pick the vertex I from a group of two, that will create two new isolated vertices. If we pick a root vertex from a group of $k \geq 3$, we create k new isolated vertices, while if we pick any other vertex, we create only one new isolated vertex.
$\triangleright$ Therefore, we have

$$\mathbb{E}(W' - W|X) \leq \frac{2N_2}{n} \cdot 2 + \sum_{k=3}^{n} \left(\frac{N_k}{n} k + \frac{(k-1)N_k}{n} \right) \leq \frac{\sum_{k=2}^{n} 2kN_k}{n} \leq 2.$$

$\triangleright$ This implies that $\mathbb{E}(W' - W|W) \leq 2$, and by substituting this into our bound on the moment generating function, we get that for $\theta \leq 0$, $m'(\theta) \geq 2n(1-p)^{n-1}\theta m(\theta)$.
$\triangleright$ From this, we obtain our concentration bound by a standard argument.

Conclusion

- Stein-type couplings can be used to show concentration inequalities for sums of dependent random variables.
- These inequalities are non-asymptotic. They can be applied even when there is no limiting distribution, or the limiting distribution is not known.
- Unlike most of the other methods in the literature, they can be also applied in situations when the random variables are not defined in terms of underlying independent random variables.
- By appropriate construction of the coupling, model specific information can be taken into account, and good bounds can be obtained.
- The arguments can be extended to obtain moment bounds as well.

THANK YOU!

Chatterjee, Sourav. 2007. Stein's method for concentration inequalities. *Probab. Theory Related Fields*, **138**(1-2), 305–321.

Chatterjee, Sourav. 2012. The missing log in large deviations for triangle counts. *Random Structures Algorithms*, **40**(4), 437–451.

Chatterjee, Sourav, & Dey, Partha S. 2010. Applications of Stein's method for concentration inequalities. *Ann. Probab.*, **38**(6), 2443–2485.

Ghosh, Subhankar, & Goldstein, Larry. 2011. Concentration of measures via size-biased couplings. *Probab. Theory Related Fields*, **149**(1-2), 271–278.

Ghosh, Subhankar, Goldstein, Larry, & Raič, Martin. 2011. Concentration of measure for the number of isolated vertices in the Erdős–Rényi random graph by size bias couplings. *Statistics & Probability Letters*, **81**(11), 1565–1570.

Goldstein, Larry, & Islak, Umit. 2013. Concentration inequalities via zero bias couplings. *arXiv preprint arXiv:1304.5001*.

Massart, Pascal. 2000. About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.*, **28**(2), 863–884.