Coupling and Poisson Approximations

Lee De Zhang

March 16, 2021

Introduction

Coupling

Stein's Method

Stein-Chen Method

Conclusion

Introduction

- Given the following:
  - A finite index set $\Gamma = \{1, 2, ..., n\}$
  - A collection of 0-1 random variables I_α, $\alpha \in \Gamma$
- Suppose $p_\alpha := \mathbb{P}(I_\alpha = 1)$'s are small (and not necessarily identical).
  - What is the behaviour of random variable $W := \sum_{\alpha \in \Gamma} I_\alpha$?
  - Convergence when $n \to \infty$?
- Stein-Chen method: W is approximated by Poisson(λ)
  - $\lambda := \sum_{\alpha \in \Gamma} p_\alpha$
  - How good is this approximation? Justified using probabilistic coupling
  - Detailed treatment in Barbour et al. [1992]

Coupling

- From Wikipedia,
  - In probability theory, coupling is a proof technique that allows one to compare two unrelated variables by 'forcing' them to be related in some way.

Definition (*Coupling*)

Given a measurable space $(\Omega, \mathcal{F})$, and two probability measures μ and ν on this space. A coupling of μ and ν is a measure γ on the space $(\Omega \times \Omega, \mathcal{F} \times \mathcal{F})$, such that the marginals of γ coincides with μ and ν, in the sense that for all $A \in \Omega$, $\gamma(A \times S) = \mu(A)$, and $\gamma(S \times A) = \nu(A)$.

- In practice, given some random variable X, coupling constructs a random variable Y on the same probability space as X
- Very useful tool to derive upper bounds

Example - Biased Coin Toss

- Given two coins X and Y, denote $H = 1, T = 0$
- $p = \mathbb{P}(X = 1) < \mathbb{P}(Y = 1) = q$
- Intuitively, $\#$ heads of $X < \#$ heads of Y a.s.
  - Can be proved using a coupling argument

- Let $X_1, ..., X_n$ be the outcome of the first n flips of coin X
- Define $Y_1, ..., Y_n$ such that,
  - If $X_i = 1$, then $Y_i = 1$
  - If $X_i = 0$, then $Y_i = 1$ with probability $(q - p)/(1 - p)$
    - Probability obtained by solving

$$\mathbb{P}(Y_i = 1 | X_i = 0)\mathbb{P}(X_i = 0) + \mathbb{P}(Y_i = 1 | X_i = 1)\mathbb{P}(X_i = 1) = q$$

    - Recall the definition, need to preserve marginal distribution of Y

- Using this choice of coupling, the sequence Y_i has the same marginal distribution as Y
- But now, (X_i, Y_i) forms a coupling
  - In the sense that Y_i depends on X_i
  - More specifically, $Y_i \geq X_i$
- Therefore, for any $k < n$,

$$\mathbb{P}(X_1 + ... + X_n > k) \leq \mathbb{P}(Y_1 + ... + Y_n > k)$$

Example - Convergence of Positive Recurrent Markov Chain

- Given an irreducible, aperiodic Markov chain X, with a countable state space S, with transition kernel Π.
- Let μ be the stationary distribution of X
  - That is,
  $$\mu(x) = \sum_{y \in S} \mu(y)\Pi(y,x),$$
  for any $x \in S$
  - Equivalently, $\langle \mu, \Pi f \rangle = \langle \mu, \Pi \rangle$ for any bounded f
- If X is positive recurrent, then for any pair $x, y \in S$,
$$\lim_{n\to\infty} \Pi^n(x,y) = \mu(y),$$
- Standard proof using renewal theorem (but we need to first prove the renewal theorem)
- Can be proved by making a coupling on X

- Let X^1 be a copy of X with initial distribution μ
  - That is, we pick the starting point of X^1 from S using μ
- Let X^2 be a copy of X with initial distribution δ_x for some $x \in S$
  - δ_x is the Dirac measure
  - Equivalently, this means we start X^2 at x with probability 1
- We claim (and prove) the following
  - X^1 and X^2 meet at some time $\tau < \infty$ a.s.
  - After they meet, they will follow the same probability measure

X^1 and X^2 meet at $\tau < \infty$ a.s.

- Let $X_n^i :=$ state of X^i at time n
  - $\tau = \inf\{n : X_n^1 = X_n^2\}$
- Checking $\tau < \infty$ is equivalent to checking if (X^1, X^2) forms an irreducible Markov chain on $S \times S$
  - Irreducibility means, for any $x, y, z \in S$,

$$\mathbb{P}[(X_n^1, X_n^2) = (z, z) | (X_0^1, X_0^2) = (y, x)] > 0,$$

  for sufficiently large $n < \infty$
- Since X^1, X^2 are aperiodic and recurrent,

$$\begin{aligned} &\mathbb{P}[(X_n^1, X_n^2) = (z, z) | (X_0^1, X_0^2) = (y, x)] \\ &= \mathbb{P}[X_n^1 = z | X_0^1 = y]\mathbb{P}[X_n^2 = z | X_0^2 = x)] \\ &> 0, \end{aligned}$$

  which proves the irreducibility of (X^1, X^2)
- Therefore $\tau < \infty$ a.s.

After the Markov chains meet

- Define two new Markov chains $\tilde{X}^1, \tilde{X}^2$
  - When $n \leq \tau$, $\tilde{X}_n^i = X_n^i$
  - When $n > \tau$, $\tilde{X}_n^i = X_n^1$
    - i.e. when both meet, the second MC follows the first MC
- By strong Markov property (present only depends on immediate past),
  - $\tilde{X}^i$ and X^i have the same distribution
- Since $\tau < \infty$ a.s., $\mathbb{P}(\tilde{X}_n^1 \neq \tilde{X}_n^2) \downarrow 0$
- $\mu(y) = \mathbb{P}(\tilde{X}_n^1)$ for all $n \in \mathbb{N}$
- $\Pi^n(x, y) = \mathbb{P}(\tilde{X}_n^2 = y)$ for all $y \in S, n \in \mathbb{N}$

- We have the total variation distance between $\Pi^n(x, y)$ and $\mu(y)$,

$$\begin{aligned}\frac{1}{2}\sum_{y\in S}|\Pi^n(x,y)-\mu(y)| &= \frac{1}{2}\sum_{y\in S}|\mathbb{E}[1_{\{\tilde{X}_n^2=y\}}-1_{\{\tilde{X}_n^1=y\}}]| \\ &\le \mathbb{P}(\tilde{X}_n^1\neq\tilde{X}_n^2) \\ &= \mathbb{P}(\tau>n),\end{aligned}$$

  which is 0 a.s. for sufficiently large n.
- Convergence in total variation distance $\Rightarrow$ pointwise convergence
- Therefore,

$$\lim_{n\to\infty}\Pi^n(x,y)=\mu(y),$$

  for all $y \in S$

- Probabilistic coupling allows sleek and efficient proving methods
- The Markov chain example can be extended to other types of random walks on graphs
- A prelude to bounding the error of Poisson approximations

Stein's Method

- Introduced by Charles Stein in his seminal paper [Stein et al., 1972]
- Produces a Berry-Essen like bound for normal approximations
- Its appeal is in its abstractness. Can generalize it for other distributions
  - Chen-Stein method generalizes it to Poisson approximation
  - Peköz [1996] extends it to geometric distributions

General Framework of Stein's Method

The metric used for bounding approximation errors is the total variation distance

Definition (*Total Variation Distance*)

Let P and Q be probability measures on the same probability space $(S, \mathcal{S})$. The total variation distance between P and Q is given by,

$$d_{TV}(P, Q) = \sup_{s \in \mathcal{S}} |P(s) - Q(s)|.$$

If $\mathcal{S}$ is countable, then d_{TV} can be rewritten as,

$$d_{TV}(P, Q) = \frac{1}{2} \sum_{s \in \mathcal{S}} |P(s) - Q(s)|.$$

The proof of the latter expression is given in Levin and Peres [2017]

General Framework of Stein's Method (cont'd)

Lemma (*Stein's Lemma*)

Define a functional operator $\mathcal{A}$, such that,

$$\mathcal{A}f(x) = f'(x) - xf(x).$$

Given a random variable W on probability space $(\Omega, \mathcal{F}, \mathcal{P})$, f is absolutely continuous, and $f' \in L_1(\Omega, \mathcal{F}, \mathcal{P})$, then $\mathcal{A}f(W) = 0$ if and only if W follows a standard normal distribution.
The operator $\mathcal{A}$ is known as the characterizing operator of the normal distribution.

Lemma (*'Solution' to Stein's Lemma*)

Let $\Phi(x)$ denote the CDF of the standard normal distribution, then the unique bounded solution of,

$$f'(w) - wf(w) = \mathbb{I}[w \leq x] - \Phi(x),$$

is given by,

$$f(w) = e^{-w^2/2} \int_{-\infty}^{w} e^{t^2/2} (\Phi(x) - \mathbb{I}(t \leq x))\, dt.$$

Theorem (*Error of Normal Approximation*)

Using f defined above, for any random variable W,

$$|\mathbb{P}[W \leq x] - \Phi(x)| = |\mathbb{E}(f'(W) - Wf(W))|$$

Final Step

- Using couplings, we can find simple bounds of $\mathbb{E}(f'(W))$ and $\mathbb{E}(Wf(W))$
  - Simple in the sense that a closed form solution is available
- Examples of such couplings, and the proofs of the previous claims, are given in [Ross et al., 2011]

Stein-Chen Method

- An extension of Stein's method by his student, Louis Chen in his paper Chen [1975]
- Using the same techniques as the Stein method, but using Poisson distribution instead of Normal
- A good introduction given in Janson [1994]
  - Rest of this presentation summarizes this paper

Poisson Approx. for Sum of Bernoulli Variables

- Recall what we want to approximate
  - $W = \sum_{\alpha \in \Gamma} I_\alpha$
  - I_α are 0-1 random variables, with $\mathbb{P}(I_\alpha = 1) = p_\alpha$
- Claim: If the individual p_α's are small, W is well approximated by Poisson(λ)
  - $\lambda = \sum_{\alpha \in \Gamma} p_\alpha$
- The Stein-Chen method is a way to justify such approximations
  - Bound $d_{TV}(\mathcal{L}(W), Po(\lambda))$
  - Show this bound $\to 0$
- General framework given in next slide

General Framework of Stein-Chen Method

1. For any $\lambda > 0$, $A \subset \mathbb{Z}$, we define the Stein equation,

$$\lambda g_{\lambda,A}(j+1) - j g_{\lambda,A}(j) = I(j \in A) - Po(\lambda)(A),$$

   for convenience, we write $g := g_{\lambda,A}$.

2. Taking expectations, we have,

$$\mathbb{E}(\lambda g(j+1) - j g(j)) = P(W \in A) - Po(\lambda)(A).$$

3. By deriving,

$$|\lambda g(j+1) - jg(j)| \leq \min(1, 1/\lambda),$$

we obtain,

$$d_{TV}(\mathcal{L}(W), Po(\lambda)) \leq \min(1, 1/\lambda) \sup_{g \in G} \mathbb{E}(\lambda g(W+1) - Wg(W)),$$

where G is the class of functions satisfying the Stein equation.

4. Letting $\lambda = \sum p_i$, we derive,

$$\begin{aligned} |\mathbb{P}(W \in A) - Po(\lambda)(A)| &= |\mathbb{E}(\lambda g(W+1) - Wg(W))| \\ &= \sum_i p_i \left(\mathbb{E}(g(W+1)) - \mathbb{E}(g(W)|I_i = 1)\right). \end{aligned}$$

- This often does not have a nice analytical solution
- Use coupling in this step

Constructing a Coupling

- Recall that we wish to bound,

$$\mathbb{E}(\lambda g(W+1) - Wg(W)) = \sum_{\alpha} p_\alpha \left(\mathbb{E}(g(W+1)) - \mathbb{E}(g(W)|I_\alpha = 1)\right).$$

- Suppose we can construct a random variable W_α on the same probability space as W
    - $\mathcal{L}(W_\alpha)$ matches that of the conditional distribution $\mathcal{L}(W - I_\alpha | I_\alpha = 1)$
- Then we derive the bound,

$$\begin{aligned} &\sum_{\alpha} p_\alpha \left(\mathbb{E}(g(W+1)) - \mathbb{E}(g(W)|I_\alpha = 1)\right) \\ &= \sum_{\alpha} p_\alpha \left(\mathbb{E}(g(W+1)) - \mathbb{E}(g(W_\alpha + 1))\right) \\ &\leq \sum_{\alpha} p_\alpha \mathbb{E}|W - W_\alpha|. \end{aligned}$$

- With the coupling coupling (W, W_α),

$$d_{TV}(\mathcal{L}(W), Po(\lambda)) \leq \min\left(1, \frac{1}{\lambda}\right) \sum_{\alpha} p_\alpha \mathbb{E}|W - W_\alpha|.$$

- Therefore, to justify a Poisson approximation, we need a coupling (W, W_α) such that $\mathbb{E}|W - W_\alpha|$ is small.
    - Such that $\sum_\alpha p_\alpha \mathbb{E}|W - W_\alpha| \to 0$.

General Method to Obtain a Coupling

- Recall that we want to construct W_α
  - $\mathcal{L}(W_\alpha) = \mathcal{L}(W - I_\alpha | I_\alpha = 1)$
- A natural way to construct this coupling is the following
  - Define a random variable $J_{\beta\alpha}$
  - $\mathcal{L}(J_{\beta\alpha}) = \mathcal{L}(I_\beta | I_\alpha = 1)$
  - Then, $W_\alpha := \sum_{\beta \neq \alpha} J_{\beta\alpha}$
- Now, we have a coupling (W, W_α)
- Recall that we want to bound $\mathbb{E}|W - W_\alpha|$, and now,

$$W - W_\alpha = I_\alpha + \sum_{\beta \neq \alpha} (I_\beta - J_{\beta\alpha}).$$

- If there is some special relationship between I_β and $J_{\beta\alpha}$, then nicer bounds can be obtained
  - For example, if $I_\beta \geq J_{\beta\alpha}$, then $(I_\beta - J_{\beta\alpha}) \leq I_\beta$
  - Such approximations may simplify the upper bound

An Upper Bound using (W, W_α)

We obtain the following upper bound using the coupling (W, W_α).

Proposition

Using the above couplings, we have,

$$d_{TV}(\mathcal{L}(W), Po(\lambda)) \leq \min(1, 1/\lambda) \left(\sum_{\alpha\in\Gamma} p_\alpha^2 + \sum_{\alpha\in\Gamma} \sum_{\beta\neq\alpha} p_\alpha \mathbb{E}|I_\beta - J_{\beta\alpha}| \right).$$

The following corollary if the I_α's are pairwise independent is immediate.

Corollary

If the I_α's are pairwise independent, then the above reduces to

$$d_{TV}(\mathcal{L}(W), Po(\lambda)) \leq \min(1, 1/\lambda) \sum_{\alpha\in\Gamma} p_\alpha^2.$$

Example - Occupancy Problem

- r balls are thrown independently of each other into some boxes
  - How many boxes will be empty at the end?
- Probability of a ball hitting box α is p_α
- Let W be the number of empty boxes.
  - Define $I_\alpha = I(\text{box } \alpha \text{ is empty})$.
  - Hence $W = \sum_\alpha I_\alpha$.
  - Goal is to approximate W

- Goal is to construct coupling (W, W_α)
  - $W_\alpha = \sum_{\beta \neq \alpha} J_{\beta\alpha}$
- How to construct $J_{\beta\alpha}$?
  - Recall that $\mathcal{L}(J_{\beta\alpha}) = \mathcal{L}(I_\beta | I_\alpha = 1)$
- Here's a way we can do this
  - We iterate through every box. Let the current box be α
  - If box α is empty ($I_\alpha = 1$), then $J_{\beta\alpha} = I_\beta$
  - If box α is occupied
    - Take all the balls from box α
    - Redistribute it to the other boxes under the conditional distribution that box α is empty.
    - i.e. redistribute to box β with probability $\mathbb{P}(I_\beta = 1 | I_\alpha = 1)$
    - Let $J_{\beta\alpha} = I(\text{box } \beta \text{ is empty after this redistribution})$.

- Under this construction, $J_{\beta\alpha} \leq I_\beta$
  - If $I_\beta = 0$ (box β is already occupied), redistribution changes nothing
  - If $I_\beta = 1$, $J_{\beta\alpha} \leq 1$ since we may redistribute a ball inside box β
- This is known as a monotone coupling
  - We get an elegant representation of the upper bound of d_{TV}!
- Since

$$p_\alpha \mathbb{E}|I_\beta - J_{\beta\alpha}| = p_\alpha \mathbb{E}(I_\beta - J_{\beta\alpha}) = -Cov(I_\alpha, I_\beta),$$

We get the following

$$
\begin{aligned}
& d_{TV}(\mathcal{L}(W), Po(\lambda)) \\
& \leq \min(1, 1/\lambda)\left(\sum_{\alpha\in\Gamma} p_\alpha^2 + \sum_{\alpha\in\Gamma}\sum_{\beta\neq\alpha} p_\alpha \mathbb{E}|I_\beta - J_{\beta\alpha}|\right) \\
& = \min(1, 1/\lambda)\left(\sum_{\alpha\in\Gamma} p_\alpha^2 - \sum_{\alpha\in\Gamma}\sum_{\beta\neq\alpha} Cov(I_\alpha, I_\beta)\right) \\
& = \min(1, 1/\lambda)\left(\sum_{\alpha\in\Gamma}(p_\alpha^2 + Var(I_\alpha)) - \sum_{\alpha\in\Gamma}\sum_{\beta\in\Gamma} Cov(I_\alpha, I_\beta)\right) \\
& = \min(1, 1/\lambda)\left(\sum_{\alpha\in\Gamma}(p_\alpha^2 + p_\alpha - p_\alpha^2) - Var(W)\right) \\
& = \min(1, 1/\lambda)\left(\lambda - Var(W)\right).
\end{aligned}
$$

- If we have sufficiently many boxes, the boxes are pairwise weakly dependent
- We can approximate $Var(W) \approx \sum_\alpha p_\alpha(1-p_\alpha)$,
  - Where p_α is the probability of a ball going into box α.
- Therefore,

$$d_{TV}(\mathcal{L}(W), Po(\lambda)) \leq \lambda - Var(W) = \sum_\alpha p_\alpha[1-(1-p_\alpha)] = \sum_\alpha p_\alpha^2,$$

Conclusion

- Probabilistic coupling is a powerful tool in probability theory
  - Enables sleek and elegant proofs!
- Coupling is an indispensable tool in the Stein-Chen method
  - A good coupling choice can be used to justify using Poisson approximations

A. D. Barbour, L. Holst, and S. Janson. *Poisson approximation*, volume 2. The Clarendon Press Oxford University Press, 1992.

L. H. Chen. Poisson approximation for dependent trials. *The Annals of Probability*, pages 534–545, 1975.

S. Janson. Coupling and poisson approximation. *Acta Applicandae Mathematica*, 34(1-2):7–15, 1994.

D. A. Levin and Y. Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

E. A. Peköz. Stein's method for geometric approximation. *Journal of applied probability*, pages 707–713, 1996.

N. Ross et al. Fundamentals of stein's method. *Probability Surveys*, 8: 210–293, 2011.

C. Stein et al. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California, 1972.