Stein's Method for Stationary Distributions of Markov Chains and Application to Ising Models

Guy Bresler and Dheeraj Nagaraj; Gesine Reinert and Nathan Ross

Arthur Gretton's notes

March 20, 2018

Summary

Tasks

- In continuous spaces, we can define a Stein operator which makes expectations zero under some distribution P. We use this to compare to a model *without normalising*
- Can we do the same for distributions on discrete spaces?

Why should we care?

- Test goodnes of fit for *complicated* models in discrete spaces (MRFs)
  - application to speech and text modelling
- Can we guarantee that a simple model is "close" to a more complex one (and in what sense)?

The basic result

Consider the vector $x \in \{-1, 1\}^n$.
Define reference probability μ and candidate probability ν. Write

$$\mu_i(\cdot|x^{(\sim i)})$$

the conditional probability of the ith entry given the remaining coordinate values $x^{(\sim i)}$.
The basic result is:

$$|E_\mu f - E_\nu f| \leq E_\nu \left(\frac{1}{n} \sum_{i=1}^{n} \alpha_i(f,\mu) \left| \mu_i(1|x^{(\sim i)}) - \nu_i(1|x^{(\sim i)}) \right| \right)$$

for any function $f : \mathcal{X} \to \mathbb{R}$.

- Comparing conditional probabilities easier than comparing full probabilities.
- The challenge: what is $\alpha_i(f,\mu)$?

Proof of the basic result: preliminaries

Define the *Glauber dynamics* P with respect to μ, and Q with respect to ν.

- This refers to a Gibbs sampler where you randomly pick which coordinate i to sample

Proof of the basic result: preliminaries

Define the *Glauber dynamics* P with respect to μ, and Q with respect to ν.

- This refers to a Gibbs sampler where you randomly pick which coordinate i to sample

We are able to define a function h associated with μ and f as

$$h - Ph = (I - P)h = f - E_\mu f$$

(the Poisson equation). The Stein operator is $(I - P)$.

Proof of the basic result: preliminaries

Define the *Glauber dynamics* P with respect to μ, and Q with respect to ν.

- This refers to a Gibbs sampler where you randomly pick which coordinate i to sample

We are able to define a function h associated with μ and f as

$$h - Ph = (I - P)h = f - E_\mu f$$

(the Poisson equation). The Stein operator is $(I - P)$.
By definition of the transition operator,

$$E_\mu(h - Ph) = 0$$

Solving for h,

$$h = (I - P)^\dagger (f - E_\mu f),$$

where we use the pseudoinverse.

Proof of the basic result

By definition of the Gibbs transition,

$$E_\nu h = E_\nu Q h.$$

Therefore

$$E_\nu f - E_\mu f = E_\nu (f - E_\mu f)$$

Proof of the basic result

By definition of the Gibbs transition,

$$E_{\nu} h = E_{\nu} Q h.$$

Therefore

$$\begin{aligned} E_{\nu} f - E_{\mu} f &= E_{\nu}(f - E_{\mu} f) \\ &= E_{\nu}(h - Ph) \end{aligned}$$

Proof of the basic result

By definition of the Gibbs transition,

$$E_{\nu} h = E_{\nu} Q h.$$

Therefore

$$\begin{aligned} E_{\nu} f - E_{\mu} f &= E_{\nu}(f - E_{\mu} f) \\ &= E_{\nu}(h - Ph) \\ &= E_{\nu}(Qh - Ph) \end{aligned}$$

Proof of the basic result

Write $x^{i,+1}$ as the vector with ith coordinate set to 1. Then

$$
\begin{aligned}
&Qh - Ph \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(h(x^{i,+1})\nu_i(1|x^{(\sim i)}) - h(x^{i,-1})\nu_i(-1|x^{(\sim i)}) \\
&\quad - h(x^{i,+1})\mu_i(1|x^{(\sim i)}) - h(x^{i,-1})\underbrace{\mu_i(-1|x^{(\sim i)})}_{=1-\mu_i(1|x^{(\sim i)})}\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n}\Delta_i(h)\left(\nu_i(1|x^{(\sim i)}) - \mu_i(1|x^{(\sim i)})\right)
\end{aligned}
$$

where we denote the ith "derivative" as

$$\Delta_i(h) = h(x^{(i,+)}) - h(x^{(i,-)}).$$

Proof of the basic result

Therefore by the above results and triangle inequality,

$$|E_{\mu}f - E_{\nu}f| \leq E_{\nu}\left(\frac{1}{n}\sum_{i=1}^{n}|\Delta_i(h)|\left|\mu_i(1|x^{(\sim i)}) - \nu_i(1|x^{(\sim i)})\right|\right).$$

The advanced result

Assume P is α-contractive: given two independent X_t, Y_t with transition P, and $\alpha \in [0, 1)$,

$$E\left[d_h(X_t, Y_t) | X_0 = x | Y_0 = y\right] \leq \alpha^t d_H(x, y).$$

Assume a smooth f:

$$f(X_t) - f(Y_t) \leq L d_H(X_t, Y_t).$$

The advanced result

Assume P is α-contractive: given two independent X_t, Y_t with transition P, and $\alpha \in [0,1)$,

$$E\left[d_h(X_t, Y_t) | X_0 = x | Y_0 = y\right] \leq \alpha^t d_H(x, y).$$

Assume a smooth f:

$$f(X_t) - f(Y_t) \leq L d_H(X_t, Y_t).$$

The advanced result is:

$$|E_\mu f - E_\nu f| \leq \frac{L}{1-\alpha} E_\nu \left(\frac{1}{n} \sum_{i=1}^{n} \left| \mu_i(1|x^{(\sim i)}) - \nu_i(1|x^{(\sim i)}) \right| \right)$$

Proof of the advanced result

Recall the definition

$$h = (I - P)^{\dagger}(f - E_{\mu} f).$$

A more interpretable way to write this is:

$$h(x) = \sum_{t=0}^{\infty} E\left[f(X_t) - E_{\mu} f | X_0 = x\right].$$

Proof of the advanced result

Using the new expression for $h(x)$,

$$\Delta_i h(x) = \sum_{t=0}^{\infty} E\left[f(X_t) - f(Y_t) - \underbrace{E_{\mu} f + E_{\mu} f}_{=0} \middle| X_0 = x^{i,+1}, Y_0 = x^{i,-1} \right]$$

$x^{i,+1}$ means ith coordinate of x set to $+1$

Proof of the advanced result

Using the new expression for $h(x)$,

$$\Delta_i h(x) = \sum_{t=0}^{\infty} E\left[f(X_t) - f(Y_t) - \underbrace{E_\mu f + E_\mu f}_{=0} \middle| X_0 = x^{i,+1}, Y_0 = x^{i,-1} \right]$$

$$\leq \sum_{t=0}^{\infty} E\left[L d_H(X_t, Y_t) | X_0 = x, Y_0 = Y \right]$$

smooth $f(X_t) - f(Y_t) \leq L d_H(X_t, Y_t)$ where Hamming distance

$$d_H(x, y) = \sum_{i=1}^{n} I_{x^i \neq y^i}.$$

Proof of the advanced result

Using the new expression for $h(x)$,

$$\Delta_i h(x) = \sum_{t=0}^{\infty} E\left[f(X_t) - f(Y_t) - \underbrace{E_\mu f + E_\mu f}_{=0} | X_0 = x^{i,+1}, Y_0 = x^{i,-1} \right]$$

$$\leq \sum_{t=0}^{\infty} E\left[L d_H(X_t, Y_t) | X_0 = x, Y_0 = Y \right]$$

$$\leq L \sum_{t=0}^{\infty} \alpha^t$$

using the α-contractive property

$$E\left[d_h(X_t, Y_t) | X_0 = x | Y_0 = y \right] \leq \alpha^t d_H(x, y).$$

Proof of the advanced result

Using the new expression for $h(x)$,

$$\Delta_i h(x) = \sum_{t=0}^{\infty} E\left[f(X_t) - f(Y_t) - \underbrace{E_\mu f + E_\mu f}_{=0} | X_0 = x^{i,+1}, Y_0 = x^{i,-1} \right]$$

$$\leq \sum_{t=0}^{\infty} E\left[L d_H(X_t, Y_t) | X_0 = x, Y_0 = Y \right]$$

$$\leq L \sum_{t=0}^{\infty} \alpha^t$$

$$= \frac{L}{1-\alpha}$$

The result applied to Ising models

Define two Ising models

$$\mu \propto \exp\left(\frac{1}{2}x^\top L x\right) \qquad \nu \propto \exp\left(\frac{1}{2}x^\top M x\right)$$

Define the a-Lipschitz function class

$$|f(x) - f(y)| \leq \sum_{i=1}^{n} a_i \mathbb{I}_{x_i \neq y_i} := a^\top \Delta_{x,y}$$

(different Lipschitz constant a_i for each coordinate i).

The result applied to Ising models

Define two Ising models

$$\mu \propto \exp\left(\frac{1}{2}x^\top Lx\right) \qquad \nu \propto \exp\left(\frac{1}{2}x^\top Mx\right)$$

Define the a-Lipschitz function class

$$|f(x) - f(y)| \leq \sum_{i=1}^{n} a_i \mathbb{I}_{x_i \neq y_i} := a^\top \Delta_{x,y}$$

(different Lipschitz constant a_i for each coordinate i).
Then

$$|E_{\pi_L} f - E_{\pi_M} f| \leq \frac{\|a\|_2 \sqrt{n}}{2(1 - \|(|L|)\|_2)} \|L - M\|_2 .$$

Unfortunately this result may be wrong...