Stein's Method in Bayesian Computation

François-Xavier Briol
University College London & The Alan Turing Institute

jISBA Seminar Series: Junior Bayes Beyond Borders (JB^3)
@ Bocconi University (virtually)

Outline

- Computational Problem: Computing or approximating integrals (or expectations) $\int_{\mathcal{X}} f(x)\mathbb{P}(dx)$ against some arbitrary target $\mathbb{P}$ is a very hard task!

- Sketch Solution: Design a very expressive class of functions $\mathcal{G}$ such that $\int_{\mathcal{X}} g(x)\mathbb{P}(dx)$ can be computed in closed form $\forall g \in \mathcal{G}$.

- Aim: Discuss the journey of Stein's method from an analytical tool in probability theory to a useful trick for computational statistics.

- This will be done with a focus on contributions of my PhD thesis (and some more recent work). Warning: It is a (very) biased overview.

Outline

- Computational Problem: Computing or approximating integrals (or expectations) $\int_{\mathcal{X}} f(x)\mathbb{P}(dx)$ against some arbitrary target $\mathbb{P}$ is a very hard task!
- Sketch Solution: Design a very expressive class of functions $\mathcal{G}$ such that $\int_{\mathcal{X}} g(x)\mathbb{P}(dx)$ can be computed in closed form $\forall g \in \mathcal{G}$.
- Aim: Discuss the journey of Stein's method from an analytical tool in probability theory to a useful trick for computational statistics.
- This will be done with a focus on contributions of my PhD thesis (and some more recent work). Warning: It is a (very) biased overview.

Outline

- Computational Problem: Computing or approximating integrals (or expectations) $\int_{\mathcal{X}} f(x)\mathbb{P}(dx)$ against some arbitrary target $\mathbb{P}$ is a very hard task!
- Sketch Solution: Design a very expressive class of functions $\mathcal{G}$ such that $\int_{\mathcal{X}} g(x)\mathbb{P}(dx)$ can be computed in closed form $\forall g \in \mathcal{G}$.
- Aim: Discuss the journey of Stein's method from an analytical tool in probability theory to a useful trick for computational statistics.
- This will be done with a focus on contributions of my PhD thesis (and some more recent work). Warning: It is a (very) biased overview.

Outline

- Computational Problem: Computing or approximating integrals (or expectations) $\int_{\mathcal{X}} f(x)\mathbb{P}(dx)$ against some arbitrary target $\mathbb{P}$ is a very hard task!
- Sketch Solution: Design a very expressive class of functions $\mathcal{G}$ such that $\int_{\mathcal{X}} g(x)\mathbb{P}(dx)$ can be computed in closed form $\forall g \in \mathcal{G}$.
- Aim: Discuss the journey of Stein's method from an analytical tool in probability theory to a useful trick for computational statistics.
- This will be done with a focus on contributions of my PhD thesis (and some more recent work). Warning: It is a (very) biased overview.

Fundamentals of Stein's Method

Stein's Method

- Stein's method allows us to characterise a probability distribution $\mathbb{P}$ through a pair $(\mathcal{U}, \mathcal{S})$ consisting of a function space $\mathcal{U}$ called Stein class and an operator $\mathcal{S}$ called Stein operator:

$$\int_{\mathcal{X}} \mathcal{S}[u](x)\mathbb{Q}(dx) = 0 \quad \forall u \in \mathcal{U} \qquad \Leftrightarrow \qquad \mathbb{P} = \mathbb{Q}$$

In particular, the space of functions $\mathcal{G}$ where $g \in \mathcal{G}$ is given by $g = \mathcal{S}[u] + c$ for $u \in \mathcal{U}$ all integrate to $c \in \mathbb{R}$ against $\mathbb{P}$.

- Example: Suppose we want to characterise $\mathbb{P} = N(0,1)$. In this case, we can take the operator $\mathcal{S}[u](x) = u'(x) - xu(x)$ and a class $\mathcal{U}$ which contains absolutely continuous functions u such that $\int_{\mathbb{R}} |u'(x)|\mathbb{P}(dx) < \infty$.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In Proceedings of 6th Berkeley Symposium on Mathematical Statistics and Probability (pp. 583-602). University of California Press.

Stein's Method

- Stein's method allows us to characterise a probability distribution $\mathbb{P}$ through a pair $(\mathcal{U}, \mathcal{S})$ consisting of a function space $\mathcal{U}$ called Stein class and an operator $\mathcal{S}$ called Stein operator:

$$\int_{\mathcal{X}} \mathcal{S}[u](x)\mathbb{Q}(dx) = 0 \quad \forall u \in \mathcal{U} \qquad \Leftrightarrow \qquad \mathbb{P} = \mathbb{Q}$$

  In particular, the space of functions $\mathcal{G}$ where $g \in \mathcal{G}$ is given by $g = \mathcal{S}[u] + c$ for $u \in \mathcal{U}$ all integrate to $c \in \mathbb{R}$ against $\mathbb{P}$.

- Example: Suppose we want to characterise $\mathbb{P} = N(0,1)$. In this case, we can take the operator $\mathcal{S}[u](x) = u'(x) - xu(x)$ and a class $\mathcal{U}$ which contains absolutely continuous functions u such that $\int_{\mathbb{R}} |u'(x)|\mathbb{P}(dx) < \infty$.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In Proceedings of 6th Berkeley Symposium on Mathematical Statistics and Probability (pp. 583-602). University of California Press.

Stein's Method

- Stein's method allows us to characterise a probability distribution $\mathbb{P}$ through a pair $(\mathcal{U}, \mathcal{S})$ consisting of a function space $\mathcal{U}$ called Stein class and an operator $\mathcal{S}$ called Stein operator:

$$\int_{\mathcal{X}} \mathcal{S}[u](x)\mathbb{Q}(dx) = 0 \quad \forall u \in \mathcal{U} \qquad \Leftrightarrow \qquad \mathbb{P} = \mathbb{Q}$$

  In particular, the space of functions $\mathcal{G}$ where $g \in \mathcal{G}$ is given by $g = \mathcal{S}[u] + c$ for $u \in \mathcal{U}$ all integrate to $c \in \mathbb{R}$ against $\mathbb{P}$.

- Example: Suppose we want to characterise $\mathbb{P} = N(0,1)$. In this case, we can take the operator $\mathcal{S}[u](x) = u'(x) - xu(x)$ and a class $\mathcal{U}$ which contains absolutely continuous functions u such that $\int_{\mathbb{R}} |u'(x)|\mathbb{P}(dx) < \infty$.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In Proceedings of 6th Berkeley Symposium on Mathematical Statistics and Probability (pp. 583-602). University of California Press.

The Generator Approach to Stein's Method

- Barbour proposed the generator approach to Stein's method.
  [1] A. D. Barbour. Stein's Method and Poisson Process Convergence. Journal of Applied Probability, 25:175-184, 1988.

- Let $\{Z_t\}_{t\in\mathbb{R}}$ be a stationnary and reversible Markov process with invariant distribution $\mathbb{P}$. Then, it's infinitesimal generator (defined over suitable functions) is given by:

$$\mathcal{A}[u](x) = \lim_{s\to 0}\left(\frac{1}{s}\mathbb{E}[u(Z_s)|Z_0 = x] - u(x)\right)$$

  This describes the behaviour of the Markov process over an infinitesimal amount of time.

- In particular, note: $\mathbb{E}\left[\mathbb{E}[u(Z_s)|Z_0 = x] - u(x)\right] = \mathbb{E}[Z_0] - \mathbb{E}[Z_0] = 0$ so we may use any such operator as a Stein operator.

The Generator Approach to Stein's Method

- Barbour proposed the generator approach to Stein's method.
  [1] A. D. Barbour. Stein's Method and Poisson Process Convergence. Journal of Applied Probability, 25:175-184, 1988.

- Let $\{Z_t\}_{t\in\mathbb{R}}$ be a stationnary and reversible Markov process with invariant distribution $\mathbb{P}$. Then, it's infinitesimal generator (defined over suitable functions) is given by:

$$\mathcal{A}[u](x) = \lim_{s\to 0}\left(\frac{1}{s}\mathbb{E}[u(Z_s)|Z_0 = x] - u(x)\right)$$

  This describes the behaviour of the Markov process over an infinitesimal amount of time.

- In particular, note: $\mathbb{E}\left[\mathbb{E}[u(Z_s)|Z_0 = x] - u(x)\right] = \mathbb{E}[Z_0] - \mathbb{E}[Z_0] = 0$ so we may use any such operator as a Stein operator.

The Generator Approach to Stein's Method

- Barbour proposed the generator approach to Stein's method.
  [1] A. D. Barbour. Stein's Method and Poisson Process Convergence. Journal of Applied Probability, 25:175-184, 1988.

- Let $\{Z_t\}_{t\in\mathbb{R}}$ be a stationnary and reversible Markov process with invariant distribution $\mathbb{P}$. Then, it's infinitesimal generator (defined over suitable functions) is given by:

$$\mathcal{A}[u](x) = \lim_{s\to 0}\left(\frac{1}{s}\mathbb{E}[u(Z_s)|Z_0 = x] - u(x)\right)$$

  This describes the behaviour of the Markov process over an infinitesimal amount of time.

- In particular, note: $\mathbb{E}\left[\mathbb{E}[u(Z_s)|Z_0 = x] - u(x)\right] = \mathbb{E}[Z_0] - \mathbb{E}[Z_0] = 0$ so we may use any such operator as a Stein operator.

Langevin Stein Operator

- In the rest of this talk, we use the generator of a Langevin diffusion. Denote by p the density of $\mathbb{P}$, then:
  - Acting on vector-valued functions $u : \mathcal{X} \to \mathbb{R}^d$:

$$\mathcal{S}_L[u] := \nabla \log p \cdot u + \nabla \cdot u$$

  - Acting on scalar-valued functions $u : \mathcal{X} \to \mathbb{R}$:

$$\mathcal{S}_{\mathrm{SL}}[u] := \nabla \log p \cdot \nabla u + \Delta u$$

- We are back in familiar territory for computational statisticians...

Langevin Stein Operator

- In the rest of this talk, we use the generator of a Langevin diffusion. Denote by p the density of $\mathbb{P}$, then:
  - Acting on vector-valued functions $u : \mathcal{X} \to \mathbb{R}^d$:

$$\mathcal{S}_L[u] := \nabla \log p \cdot u + \nabla \cdot u$$

  - Acting on scalar-valued functions $u : \mathcal{X} \to \mathbb{R}$:

$$\mathcal{S}_{\mathrm{SL}}[u] := \nabla \log p \cdot \nabla u + \Delta u$$

- We are back in familiar territory for computational statisticians...

Langevin Stein Operator

- In the rest of this talk, we use the generator of a Langevin diffusion. Denote by p the density of $\mathbb{P}$, then:
  - Acting on vector-valued functions $u : \mathcal{X} \to \mathbb{R}^d$:

$$\mathcal{S}_L[u] := \nabla \log p \cdot u + \nabla \cdot u$$

  - Acting on scalar-valued functions $u : \mathcal{X} \to \mathbb{R}$:

$$\mathcal{S}_{\mathrm{SL}}[u] := \nabla \log p \cdot \nabla u + \Delta u$$

- We are back in familiar territory for computational statisticians...

Langevin Stein Operator

- We now have a class of functions $\mathcal{G}$ (of the form $g = \mathcal{S}_{\mathsf{SL}}[u]$) which integrate to zero:

$$\int_{\mathcal{X}} g(x)\mathbb{P}(dx) = \int_{\mathcal{X}} \mathcal{S}_{\mathsf{SL}}[u](x)\mathbb{P}(dx)$$
$$= \int_{\mathcal{X}} (\nabla \log p(x) \cdot \nabla u(x) + \Delta u(x))\mathbb{P}(dx) = 0$$

- Unlike in the previous case, this operator can be used for a very large class of distributions!

- Important Remark: Evaluating $\nabla \log p$ does not require any knowledge of normalisation constant of p.

(I am hiding some technical conditions!)

Langevin Stein Operator

- We now have a class of functions $\mathcal{G}$ (of the form $g = \mathcal{S}_{\mathsf{SL}}[u]$) which integrate to zero:

$$
\begin{aligned}
\int_{\mathcal{X}} g(x)\mathbb{P}(dx) &= \int_{\mathcal{X}} \mathcal{S}_{\mathsf{SL}}[u](x)\mathbb{P}(dx) \\
&= \int_{\mathcal{X}} (\nabla \log p(x) \cdot \nabla u(x) + \Delta u(x))\mathbb{P}(dx) = 0
\end{aligned}
$$

- Unlike in the previous case, this operator can be used for a very large class of distributions!

- Important Remark: Evaluating $\nabla \log p$ does not require any knowledge of normalisation constant of p.

(I am hiding some technical conditions!)

Langevin Stein Operator

- We now have a class of functions $\mathcal{G}$ (of the form $g = \mathcal{S}_{\mathsf{SL}}[u]$) which integrate to zero:

$$\int_{\mathcal{X}} g(x)\mathbb{P}(dx) = \int_{\mathcal{X}} \mathcal{S}_{\mathsf{SL}}[u](x)\mathbb{P}(dx)$$
$$= \int_{\mathcal{X}} (\nabla \log p(x) \cdot \nabla u(x) + \Delta u(x))\mathbb{P}(dx) = 0$$

- Unlike in the previous case, this operator can be used for a very large class of distributions!

- Important Remark: Evaluating $\nabla \log p$ does not require any knowledge of normalisation constant of p.

(I am hiding some technical conditions!)

Langevin Stein Operator

- We now have a class of functions $\mathcal{G}$ (of the form $g = \mathcal{S}_{\mathsf{SL}}[u]$) which integrate to zero:

$$\begin{aligned}\int_{\mathcal{X}} g(x)\mathbb{P}(dx) &= \int_{\mathcal{X}} \mathcal{S}_{\mathsf{SL}}[u](x)\mathbb{P}(dx) \\ &= \int_{\mathcal{X}} (\nabla \log p(x) \cdot \nabla u(x) + \Delta u(x))\mathbb{P}(dx) = 0\end{aligned}$$

- Unlike in the previous case, this operator can be used for a very large class of distributions!

- Important Remark: Evaluating $\nabla \log p$ does not require any knowledge of normalisation constant of p.

(I am hiding some technical conditions!)

Useful Quantity: Stein Discrepancies

- Integral probability metric (e.g. TV, Wasserstein, MMD...):

$$D(\mathbb{P},\mathbb{Q}) := \sup_{f \in \mathcal{F}} \left| \int_{\mathcal{X}} f(x)\mathbb{P}(dx) - \int_{\mathcal{X}} f(x)\mathbb{Q}(dx) \right|$$

- We call Stein discrepancy:

$$D_{\mathcal{U},\mathcal{S}}(\mathbb{P}\|\mathbb{Q}) := \sup_{u \in \mathcal{U}} \left| \underbrace{\int_{\mathcal{X}} \mathcal{S}[u](x)\mathbb{P}(dx)}_{=0 \text{ since } u \in \mathcal{U}} - \int_{\mathcal{X}} \mathcal{S}[u](x)\mathbb{Q}(dx) \right|$$

$$:= \sup_{u \in \mathcal{U}} \left| \int_{\mathcal{X}} \mathcal{S}[u](x)\mathbb{Q}(dx) \right|$$

- Let $\mathcal{V} \subseteq \mathcal{U}$. In particular:

$$D_{\mathcal{V},\mathcal{S}}\left(\mathbb{P}\middle\|\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}\right) := \sup_{u \in \mathcal{V}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathcal{S}[u](x_i) \right|$$

Useful Quantity: Stein Discrepancies

- Integral probability metric (e.g. TV, Wasserstein, MMD...):

$$D(\mathbb{P},\mathbb{Q}) := \sup_{f\in\mathcal{F}} \left| \int_{\mathcal{X}} f(x)\mathbb{P}(dx) - \int_{\mathcal{X}} f(x)\mathbb{Q}(dx) \right|$$

- We call Stein discrepancy:

$$D_{\mathcal{U},\mathcal{S}}(\mathbb{P}\|\mathbb{Q}) := \sup_{u\in\mathcal{U}} \left| \underbrace{\int_{\mathcal{X}} \mathcal{S}[u](x)\mathbb{P}(dx)}_{=0 \text{ since } u\in\mathcal{U}} - \int_{\mathcal{X}} \mathcal{S}[u](x)\mathbb{Q}(dx) \right|$$

$$:= \sup_{u\in\mathcal{U}} \left| \int_{\mathcal{X}} \mathcal{S}[u](x)\mathbb{Q}(dx) \right|$$

- Let $\mathcal{V} \subseteq \mathcal{U}$. In particular:

$$D_{\mathcal{V},\mathcal{S}}\left(\mathbb{P}\middle\|\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}\right) := \sup_{u\in\mathcal{V}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathcal{S}[u](x_i) \right|$$

Useful Quantity: Stein Discrepancies

- Integral probability metric (e.g. TV, Wasserstein, MMD...):

$$D(\mathbb{P},\mathbb{Q}) := \sup_{f\in\mathcal{F}} \left| \int_{\mathcal{X}} f(x)\mathbb{P}(dx) - \int_{\mathcal{X}} f(x)\mathbb{Q}(dx) \right|$$

- We call Stein discrepancy:

$$D_{\mathcal{U},\mathcal{S}}(\mathbb{P}\|\mathbb{Q}) := \sup_{u\in\mathcal{U}} \left| \underbrace{\int_{\mathcal{X}} \mathcal{S}[u](x)\mathbb{P}(dx)}_{=0 \text{ since } u\in\mathcal{U}} - \int_{\mathcal{X}} \mathcal{S}[u](x)\mathbb{Q}(dx) \right|$$

$$:= \sup_{u\in\mathcal{U}} \left| \int_{\mathcal{X}} \mathcal{S}[u](x)\mathbb{Q}(dx) \right|$$

- Let $\mathcal{V} \subseteq \mathcal{U}$. In particular:

$$D_{\mathcal{V},\mathcal{S}}\left(\mathbb{P}\middle\|\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}\right) := \sup_{u\in\mathcal{V}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathcal{S}[u](x_i) \right|$$

Kernel Stein Discrepancies

- Example: Let $\mathcal{H}_k$ to be the unit-ball of some reproducing kernel Hilbert space (RKHS) with kernel k. Take $\mathcal{V} = \mathcal{H}_k^d$ and $\mathcal{S} = \mathcal{S}_{\mathsf{L}}$. Then, we get the kernel Stein discrepancy (KSD):

$$D_{\mathcal{V},\mathcal{S}}\left(\mathbb{P} \Big\| \frac{1}{n}\sum_{i=1}^n \delta_{x_i}\right) := \sqrt{\frac{1}{n^2}\sum_{i,j=1}^n k_0(x_i, x_j)}$$

$$\begin{aligned} k_0(x,x) := {} & k(x,x')\nabla_x \log p(x)^\top \nabla_{x'} \log p(x') + \mathrm{Tr}(\nabla_x \nabla_{x'} k(x,x')) \\ & + \nabla_{x'} k(x,x')^\top \nabla_x \log p(x) + \nabla_x k(x,x')^\top \nabla_{x'} \log p(x') \end{aligned}$$

where for example $k(x,x') = \exp(-\|x-y\|_2^2/l^2)$ for $l > 0$.

Kernel Stein Discrepancies

- Example: Let $\mathcal{H}_k$ to be the unit-ball of some reproducing kernel Hilbert space (RKHS) with kernel k. Take $\mathcal{V} = \mathcal{H}_k^d$ and $\mathcal{S} = \mathcal{S}_{\mathsf{L}}$. Then, we get the kernel Stein discrepancy (KSD):

$$D_{\mathcal{V},\mathcal{S}}\left(\mathbb{P} \middle\| \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}\right) := \sqrt{\frac{1}{n^2}\sum_{i,j=1}^{n} k_0(x_i, x_j)}$$

$$\begin{aligned} k_0(x,x) := {} & k(x,x')\nabla_x \log p(x)^\top \nabla_{x'} \log p(x') + \mathrm{Tr}(\nabla_x \nabla_{x'} k(x,x')) \\ & + \nabla_{x'} k(x,x')^\top \nabla_x \log p(x) + \nabla_x k(x,x')^\top \nabla_{x'} \log p(x') \end{aligned}$$

where for example $k(x,x') = \exp(-\|x-y\|_2^2/l^2)$ for $l > 0$.

Application #1: Approximation of Posterior Distributions

Stein Points

- Task: We want to approximate a posterior $\mathbb{P}$ with points $\{x_i\}_{i=1}^n$.
- Solution: Minimise a Stein discrepancy:

$$\underset{\{x_i\}_{i=1}^n \subset \mathcal{X}}{\arg\min}\, D_{\mathcal{V},\mathcal{S}}\left(\mathbb{P} \middle\| \frac{1}{n}\sum_{i=1}^n \delta_{x_i}\right)$$

- In general, this is an intractable optimisation problem (it is very high-dimensional and non-convex), but we can solve it approximately.
- We call any point sets approximating this objective Stein Points.

Chen, W. Y., Mackey, L., Gorham, J., Briol, F.-X., and Oates, C. J. (2018). Stein points. International Conference on Machine Learning, PMLR 80 (pp. 843-852).

Chen, W. Y., Barp, A., Briol, F.-X., Gorham, J., Girolami, M., Mackey, L., and Oates, C. J. (2019). Stein point Markov chain Monte Carlo. International Conference on Machine Learning, PMLR 97 (pp. 1011-1021).

Stein Points

- Task: We want to approximate a posterior $\mathbb{P}$ with points $\{x_i\}_{i=1}^n$.
- Solution: Minimise a Stein discrepancy:

$$\operatorname*{arg\,min}_{\{x_i\}_{i=1}^n \subset \mathcal{X}} D_{\mathcal{V},\mathcal{S}}\left(\mathbb{P} \middle\| \frac{1}{n}\sum_{i=1}^n \delta_{x_i}\right)$$

- In general, this is an intractable optimisation problem (it is very high-dimensional and non-convex), but we can solve it approximately.
- We call any point sets approximating this objective Stein Points.

Chen, W. Y., Mackey, L., Gorham, J., Briol, F.-X., and Oates, C. J. (2018). Stein points. International Conference on Machine Learning, PMLR 80 (pp. 843-852).

Chen, W. Y., Barp, A., Briol, F.-X., Gorham, J., Girolami, M., Mackey, L., and Oates, C. J. (2019). Stein point Markov chain Monte Carlo. International Conference on Machine Learning, PMLR 97 (pp. 1011-1021).

Stein Points

- Task: We want to approximate a posterior $\mathbb{P}$ with points $\{x_i\}_{i=1}^n$.
- Solution: Minimise a Stein discrepancy:

$$\underset{\{x_i\}_{i=1}^n \subset \mathcal{X}}{\arg\min} \; D_{\mathcal{V},\mathcal{S}}\left(\mathbb{P} \middle\| \frac{1}{n}\sum_{i=1}^n \delta_{x_i}\right)$$

- In general, this is an intractable optimisation problem (it is very high-dimensional and non-convex), but we can solve it approximately.
- We call any point sets approximating this objective Stein Points.

Chen, W. Y., Mackey, L., Gorham, J., Briol, F.-X., and Oates, C. J. (2018). Stein points. International Conference on Machine Learning, PMLR 80 (pp. 843-852).

Chen, W. Y., Barp, A., Briol, F.-X., Gorham, J., Girolami, M., Mackey, L., and Oates, C. J. (2019). Stein point Markov chain Monte Carlo. International Conference on Machine Learning, PMLR 97 (pp. 1011-1021).

Stein Points

- Task: We want to approximate a posterior $\mathbb{P}$ with points $\{x_i\}_{i=1}^n$.
- Solution: Minimise a Stein discrepancy:

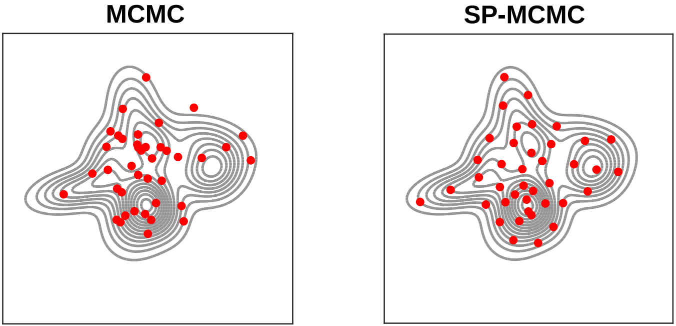
$$\underset{\{x_i\}_{i=1}^n \subset \mathcal{X}}{\arg\min} \, D_{\mathcal{V},\mathcal{S}} \left(\mathbb{P} \middle\| \frac{1}{n} \sum_{i=1}^n \delta_{x_i} \right)$$

- In general, this is an intractable optimisation problem (it is very high-dimensional and non-convex), but we can solve it approximately.
- We call any point sets approximating this objective Stein Points.

Chen, W. Y., Mackey, L., Gorham, J., Briol, F.-X., and Oates, C. J. (2018). Stein points. International Conference on Machine Learning, PMLR 80 (pp. 843-852).

Chen, W. Y., Barp, A., Briol, F.-X., Gorham, J., Girolami, M., Mackey, L., and Oates, C. J. (2019). Stein point Markov chain Monte Carlo. International Conference on Machine Learning, PMLR 97 (pp. 1011-1021).

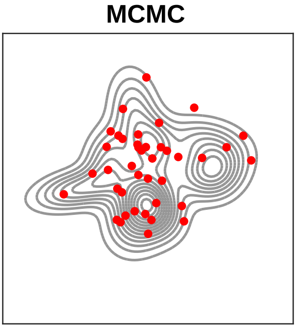
Stein Point MCMC

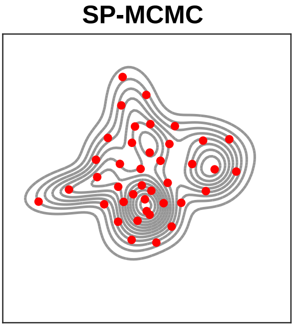


SP-MCMC:

- Greedy approximation of the KSD over the the path of a Markov chain (with an adaptive restart strategy).
- More expensive than MCMC, but gives "better" point sets! Particularly useful when $\nabla_x \log p$ is expensive.

Stein Point MCMC



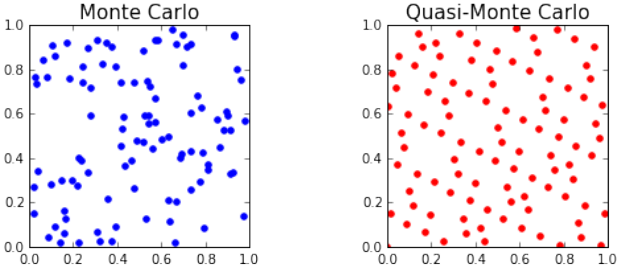


SP-MCMC:

- Greedy approximation of the KSD over the the path of a Markov chain (with an adaptive restart strategy).
- More expensive than MCMC, but gives "better" point sets! Particularly useful when $\nabla_x \log p$ is expensive.

Stein Points: Connections with QMC

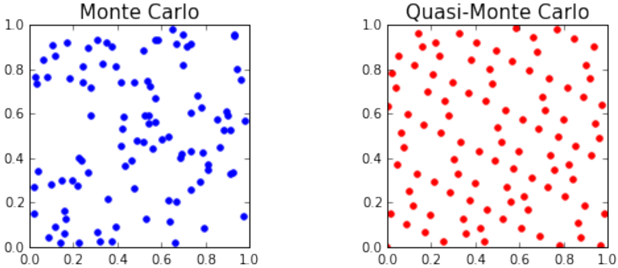
- There are some close parallels with quasi-Monte Carlo (QMC):



- There, the aim is to minimise the star-discrepancy.
- Of course, the big difference is that QMC is restricted to $\mathcal{X}$ being the unit cube and uniform $\mathbb{P}$. In comparison, Stein Points can be used when $\mathbb{P}$ is a posterior distribution.

Stein Points: Connections with QMC

- There are some close parallels with quasi-Monte Carlo (QMC):



- There, the aim is to minimise the star-discrepancy.
- Of course, the big difference is that QMC is restricted to $\mathcal{X}$ being the unit cube and uniform $\mathbb{P}$. In comparison, Stein Points can be used when $\mathbb{P}$ is a posterior distribution.

Application #2: Estimators for Unnormalised Models

Minimum Stein Discrepancy Estimators

- Task: We have $p_\theta(x) = \bar{p}_\theta(x)/C_\theta$ ($\bar{p}_\theta(x)$ can be evaluated pointwise), and our aim is to recover θ^* given iid realisations $\{x_i\}_{i=1}^n$ from $\mathbb{P}_{\theta^*}$.
- Solution: Estimators called Minimum Stein discrepancy estimators:

$$\hat{\theta}_n := \arg\min_\theta D_{\mathcal{V},\mathcal{S}}\left(\mathbb{P}_\theta \middle\| \frac{1}{n}\sum_{i=1}^n \delta_{x_i}\right)$$

A. Barp, F.-X. Briol, A. B. Duncan, M. Girolami, and L. Mackey. Minimum Stein discrepancy estimators. Neural Information Processing Systems, pages 12964–12976, 2019

- We showed many algorithms are special cases, including contrastive divergence ($\sim$ 4500 citations), score-matching and ratio matching ($\sim$ 600 citations), minimum probability flow ($\sim$ 150 citations).
- It is also possible to create new algorithms by changing $\mathcal{V}$ and $\mathcal{S}$!

Minimum Stein Discrepancy Estimators

- Task: We have $p_\theta(x) = \bar{p}_\theta(x)/C_\theta$ ($\bar{p}_\theta(x)$ can be evaluated pointwise), and our aim is to recover θ^* given iid realisations $\{x_i\}_{i=1}^n$ from $\mathbb{P}_{\theta^*}$.
- Solution: Estimators called Minimum Stein discrepancy estimators:

$$\hat{\theta}_n := \arg\min_\theta D_{\mathcal{V},\mathcal{S}}\left(\mathbb{P}_\theta \middle\| \frac{1}{n}\sum_{i=1}^n \delta_{x_i}\right)$$

A. Barp, F.-X. Briol, A. B. Duncan, M. Girolami, and L. Mackey. Minimum Stein discrepancy estimators. Neural Information Processing Systems, pages 12964-12976, 2019

- We showed many algorithms are special cases, including contrastive divergence ($\sim$ 4500 citations), score-matching and ratio matching ($\sim$ 600 citations), minimum probability flow ($\sim$ 150 citations).
- It is also possible to create new algorithms by changing $\mathcal{V}$ and $\mathcal{S}$!

Minimum Stein Discrepancy Estimators

- Task: We have $p_\theta(x) = \bar{p}_\theta(x)/C_\theta$ ($\bar{p}_\theta(x)$ can be evaluated pointwise), and our aim is to recover θ^* given iid realisations $\{x_i\}_{i=1}^n$ from $\mathbb{P}_{\theta^*}$.
- Solution: Estimators called Minimum Stein discrepancy estimators:

$$\hat{\theta}_n := \arg\min_\theta D_{\mathcal{V},\mathcal{S}} \left(\mathbb{P}_\theta \middle\| \frac{1}{n} \sum_{i=1}^n \delta_{x_i} \right)$$

A. Barp, F.-X. Briol, A. B. Duncan, M. Girolami, and L. Mackey. Minimum Stein discrepancy estimators. Neural Information Processing Systems, pages 12964-12976, 2019

- We showed many algorithms are special cases, including contrastive divergence ($\sim$ 4500 citations), score-matching and ratio matching ($\sim$ 600 citations), minimum probability flow ($\sim$ 150 citations).
- It is also possible to create new algorithms by changing $\mathcal{V}$ and $\mathcal{S}$!

Minimum Stein Discrepancy Estimators

- Task: We have $p_\theta(x) = \bar{p}_\theta(x)/C_\theta$ ($\bar{p}_\theta(x)$ can be evaluated pointwise), and our aim is to recover θ^* given iid realisations $\{x_i\}_{i=1}^n$ from $\mathbb{P}_{\theta^*}$.
- Solution: Estimators called Minimum Stein discrepancy estimators:

$$\hat{\theta}_n := \arg\min_\theta D_{\mathcal{V},\mathcal{S}} \left(\mathbb{P}_\theta \middle\| \frac{1}{n} \sum_{i=1}^n \delta_{x_i} \right)$$

A. Barp, F.-X. Briol, A. B. Duncan, M. Girolami, and L. Mackey. Minimum Stein discrepancy estimators. Neural Information Processing Systems, pages 12964-12976, 2019

- We showed many algorithms are special cases, including contrastive divergence ($\sim$ 4500 citations), score-matching and ratio matching ($\sim$ 600 citations), minimum probability flow ($\sim$ 150 citations).
- It is also possible to create new algorithms by changing $\mathcal{V}$ and $\mathcal{S}$!

Application #3: Control Variates

Control Variates for Monte Carlo Methods

- Task: We would like to approximate some integral $\int_{\mathcal{X}} f(x)\mathbb{P}(dx)$ using a Monte Carlo estimator $\frac{1}{n}\sum_{i=1}^{n} f(x_i)$ where $\{x_i\}_{i=1}^{n}$ is iid from $\mathbb{P}$.
- From the CLT, we know that the speed of convergence of Monte Carlo estimators depends on $\sigma_f^2 = \mathrm{Var}_{\mathbb{P}}[f]$:

$$\sqrt{n}\left(\int_{\mathcal{X}} f(x)\mathbb{P}(dx) - \frac{1}{n}\sum_{i=1}^{n} f(x_i)\right) \to \mathcal{N}(0, \sigma_f^2)$$

- For MCMC: $\sigma_f^2 = \mathrm{Var}[f(X_1)] + 2\sum_{k=1}^{\infty} \mathrm{Cov}(f(X_1), f(X_{1+k}))$.
- Similar expressions exists for QMC...

Control Variates for Monte Carlo Methods

- Task: We would like to approximate some integral $\int_{\mathcal{X}} f(x)\mathbb{P}(dx)$ using a Monte Carlo estimator $\frac{1}{n}\sum_{i=1}^{n} f(x_i)$ where $\{x_i\}_{i=1}^{n}$ is iid from $\mathbb{P}$.
- From the CLT, we know that the speed of convergence of Monte Carlo estimators depends on $\sigma_f^2 = \text{Var}_{\mathbb{P}}[f]$:

$$\sqrt{n}\left(\int_{\mathcal{X}} f(x)\mathbb{P}(dx) - \frac{1}{n}\sum_{i=1}^{n} f(x_i)\right) \to \mathcal{N}(0, \sigma_f^2)$$

- For MCMC: $\sigma_f^2 = \text{Var}[f(X_1)] + 2\sum_{k=1}^{\infty} \text{Cov}(f(X_1), f(X_{1+k}))$.
- Similar expressions exists for QMC...

Control Variates for Monte Carlo Methods

- Task: We would like to approximate some integral $\int_{\mathcal{X}} f(x)\mathbb{P}(dx)$ using a Monte Carlo estimator $\frac{1}{n}\sum_{i=1}^{n} f(x_i)$ where $\{x_i\}_{i=1}^{n}$ is iid from $\mathbb{P}$.
- From the CLT, we know that the speed of convergence of Monte Carlo estimators depends on $\sigma_f^2 = \mathrm{Var}_{\mathbb{P}}[f]$:

$$\sqrt{n}\left(\int_{\mathcal{X}} f(x)\mathbb{P}(dx) - \frac{1}{n}\sum_{i=1}^{n} f(x_i)\right) \to \mathcal{N}(0, \sigma_f^2)$$

- For MCMC: $\sigma_f^2 = \mathrm{Var}[f(X_1)] + 2\sum_{k=1}^{\infty} \mathrm{Cov}(f(X_1), f(X_{1+k}))$.
- Similar expressions exists for QMC...

Variance Reduction with Control Variates

- Solution: Use a control variate (CV), which is a function g such that:

$$\int_{\mathcal{X}} f(x)\mathbb{P}(dx) = \int_{\mathcal{X}} f(x) - g(x)\mathbb{P}(dx),$$
$$\mathrm{Var}_{\mathbb{P}}[f - g] \ll \mathrm{Var}_{\mathbb{P}}[f].$$

(1) To satisfy the first criterion, we can build CVs using Stein method by taking $g = \mathcal{S}[u]$ for some $u \in \mathcal{U}$, a Stein space.

(2) To satisfy the second criterion, we can choose the "best" CV in some approximation space $\mathcal{V} \subseteq \mathcal{U}$ in the sense of minimising the variance:

$$u^* = \arg\inf_{u \in \mathcal{V} \subseteq \mathcal{U}} \mathrm{Var}_{\mathbb{P}}[f - \mathcal{S}[u]]$$

In particular if $\mathcal{V} = \mathcal{U}$, this would give a zero-variance CV!

Variance Reduction with Control Variates

- Solution: Use a control variate (CV), which is a function g such that:

$$\int_{\mathcal{X}} f(x)\mathbb{P}(dx) = \int_{\mathcal{X}} f(x) - g(x)\mathbb{P}(dx),$$
$$\mathrm{Var}_{\mathbb{P}}[f - g] \ll \mathrm{Var}_{\mathbb{P}}[f].$$

(1) To satisfy the first criterion, we can build CVs using Stein method by taking $g = \mathcal{S}[u]$ for some $u \in \mathcal{U}$, a Stein space.

(2) To satisfy the second criterion, we can choose the "best" CV in some approximation space $\mathcal{V} \subseteq \mathcal{U}$ in the sense of minimising the variance:

$$u^* = \arg\inf_{u \in \mathcal{V} \subseteq \mathcal{U}} \mathrm{Var}_{\mathbb{P}}[f - \mathcal{S}[u]]$$

In particular if $\mathcal{V} = \mathcal{U}$, this would give a zero-variance CV!

Variance Reduction with Control Variates

- Solution: Use a control variate (CV), which is a function g such that:

$$\int_{\mathcal{X}} f(x)\mathbb{P}(dx) = \int_{\mathcal{X}} f(x) - g(x)\mathbb{P}(dx),$$
$$\mathrm{Var}_{\mathbb{P}}[f - g] \ll \mathrm{Var}_{\mathbb{P}}[f].$$

(1) To satisfy the first criterion, we can build CVs using Stein method by taking $g = \mathcal{S}[u]$ for some $u \in \mathcal{U}$, a Stein space.

(2) To satisfy the second criterion, we can choose the "best" CV in some approximation space $\mathcal{V} \subseteq \mathcal{U}$ in the sense of minimising the variance:

$$u^* = \arg\inf_{u \in \mathcal{V} \subseteq \mathcal{U}} \mathrm{Var}_{\mathbb{P}}[f - \mathcal{S}[u]]$$

In particular if $\mathcal{V} = \mathcal{U}$, this would give a zero-variance CV!

Some Approximations

A few approximations are needed to solve this problem:

(1) We look for the best CV in some parametric subspace $\mathcal{V}_\Theta \subset \mathcal{U}$.

$$\theta^* = \arg\min_{\theta \in \Theta} \mathrm{Var}_{\mathbb{P}}[f - \mathcal{S}[u_\theta]].$$

(2) We approximate the variance with a subset of size $m \ll n$ of samples:

$$\widehat{\mathrm{Var}}_m[f - \mathcal{S}[u_\theta]] \approx \mathrm{Var}_{\mathbb{P}}[f - \mathcal{S}[u_\theta]].$$

The literature has a variety of special cases with different combinations of

$$\widehat{\mathrm{Var}}_m[f - \mathcal{S}[u_\theta]] \quad \text{and} \quad \mathcal{V}_\Theta.$$

Some Approximations

A few approximations are needed to solve this problem:

(1) We look for the best CV in some parametric subspace $\mathcal{V}_\Theta \subset \mathcal{U}$.

$$\theta^* = \arg\min_{\theta \in \Theta} \mathrm{Var}_{\mathbb{P}}[f - \mathcal{S}[u_\theta]].$$

(2) We approximate the variance with a subset of size $m \ll n$ of samples:

$$\widehat{\mathrm{Var}}_m[f - \mathcal{S}[u_\theta]] \approx \mathrm{Var}_{\mathbb{P}}[f - \mathcal{S}[u_\theta]].$$

The literature has a variety of special cases with different combinations of

$$\widehat{\mathrm{Var}}_m[f - \mathcal{S}[u_\theta]] \quad \text{and} \quad \mathcal{V}_\Theta.$$

Special Cases

- $\mathcal{V}_\Theta$ is a space of polynomials with parameters in Θ:

  Assaraf, R., & Caffarel, M. (1999). Zero-variance principle for Monte Carlo algorithms. Physical Review Letters, 83(23), 4682.

  Mira, A., Solgi, R., & Imparato, D. (2013). Zero variance Markov chain Monte Carlo for Bayesian estimators. Statistics and Computing, 23(5), 653-662.

  South, L. F., Oates, C. J., Mira, A., & Drovandi, C. (2019). Regularised zero-variance control variates for high-dimensional variance reduction. arXiv:1811.05073.

- $\mathcal{V}_\Theta$ is a weighted sum of kernel evaluations with weights in Θ:

  Oates, C. J., Girolami, M., & Chopin, N. (2017). Control functionals for Monte Carlo integration. Journal of the Royal Statistical Society B, 79(3), 695-718.

  Oates, C. J., Cockayne, J., Briol, F.-X., & Girolami, M. (2019). Convergence rates for a class of estimators based on Stein's identity. Bernoulli, 25(2), 1141-1159.

The Genz Functions

Integrand f	MC	Poly. CV	Ker. CV	Poly.+Ker. CV
Continuous	2.77e-03	3.21e-03	3.28e-04	**1.85e-04**
Corner Peak	5.76e-03	1.07e-03	9.27e-06	**6.05e-06**
Discontinuous	2.04e-02	1.32e-02	3.91e-03	**2.65e-03**
Gaussian Peak	1.47e-03	1.40e-03	1.24e-05	**1.05e-05**
Oscillatory	4.17e-03	1.06e-03	4.63e-06	**3.90e-06**
Product Peak	1.37e-03	1.32e-03	2.12e-05	**2.52e-06**

Genz functions: Computed the mean absolute error (over 20 runs) for a set of 6 test functions with challenging features for integration (e.g. fast oscillations, peaks, discontinuities, etc...)

We took $m = 1000$ and $d = 1$. Polynomial CVs were of order 2.

Computing these CVs significantly improves the performance but can be very computationally expensive.

The Genz Functions

Integrand f	MC	Poly. CV	Ker. CV	Poly.+Ker. CV
Continuous	2.77e-03	3.21e-03	3.28e-04	**1.85e-04**
Corner Peak	5.76e-03	1.07e-03	9.27e-06	**6.05e-06**
Discontinuous	2.04e-02	1.32e-02	3.91e-03	**2.65e-03**
Gaussian Peak	1.47e-03	1.40e-03	1.24e-05	**1.05e-05**
Oscillatory	4.17e-03	1.06e-03	4.63e-06	**3.90e-06**
Product Peak	1.37e-03	1.32e-03	2.12e-05	**2.52e-06**

Genz functions: Computed the mean absolute error (over 20 runs) for a set of 6 test functions with challenging features for integration (e.g. fast oscillations, peaks, discontinuities, etc...)

We took $m = 1000$ and $d = 1$. Polynomial CVs were of order 2.

Computing these CVs significantly improves the performance but can be very computationally expensive.

Computational Complexity

- Problem: These linear systems quickly become enormous when the number of samples m is large, or the dimension d is large. The computational cost is cubic in the number of parameters.

Si, S., Oates, C. J., Duncan, A. B., Carin, L., & Briol, F.-X. (2020). Scalable control variates for Monte Carlo methods via stochastic optimization. arXiv:2006.07487.

- Significant speed-ups can be obtained by minimising the following objective through stochastic optimisation:

$$\arg\min_{\theta\in\Theta} \hat{J}_m(\theta) = \arg\min_{\theta\in\Theta} \widehat{\mathrm{Var}}_m[f - \mathcal{S}[u_\theta]] + \lambda_m \|\theta\|^2$$

$$\widehat{\mathrm{Var}}_m[f - \mathcal{S}[u_\theta]] = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - \mathcal{S}[u_\theta] - \theta_0)^2$$

Stochastic gradient descent (SGD): Given some $\{\alpha_t\}_{t\in\mathbb{N}_+}$, we:

- Initialise $\theta_0 \in \Theta$.
- For $t \in \mathbb{N}_+$, $\theta_{t+1} = \theta_t - \alpha_t \nabla_\theta \hat{J}_m(\theta)$

Computational Complexity

- Problem: These linear systems quickly become enormous when the number of samples m is large, or the dimension d is large. The computational cost is cubic in the number of parameters.

  Si, S., Oates, C. J., Duncan, A. B., Carin, L., & Briol, F.-X. (2020). Scalable control variates for Monte Carlo methods via stochastic optimization. arXiv:2006.07487.

- Significant speed-ups can be obtained by minimising the following objective through stochastic optimisation:

$$\arg\min_{\theta \in \Theta} \hat{J}_m(\theta) = \arg\min_{\theta \in \Theta} \widehat{\mathrm{Var}}_m[f - \mathcal{S}[u_\theta]] + \lambda_m \|\theta\|^2$$

$$\widehat{\mathrm{Var}}_m[f - \mathcal{S}[u_\theta]] = \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - \mathcal{S}[u_\theta] - \theta_0)^2$$

  Stochastic gradient descent (SGD): Given some $\{\alpha_t\}_{t \in \mathbb{N}_+}$, we:
  - Initialise $\theta_0 \in \Theta$.
  - For $t \in \mathbb{N}_+$, $\theta_{t+1} = \theta_t - \alpha_t \nabla_\theta \hat{J}_m(\theta)$

Computational Complexity

- Problem: These linear systems quickly become enormous when the number of samples m is large, or the dimension d is large. The computational cost is cubic in the number of parameters.

  Si, S., Oates, C. J., Duncan, A. B., Carin, L., & Briol, F.-X. (2020). Scalable control variates for Monte Carlo methods via stochastic optimization. arXiv:2006.07487.

- Significant speed-ups can be obtained by minimising the following objective through stochastic optimisation:

$$\arg\min_{\theta\in\Theta} \hat{J}_m(\theta) = \arg\min_{\theta\in\Theta} \widehat{\mathrm{Var}}_m[f - \mathcal{S}[u_\theta]] + \lambda_m \|\theta\|^2$$

$$\widehat{\mathrm{Var}}_m[f - \mathcal{S}[u_\theta]] = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - \mathcal{S}[u_\theta] - \theta_0)^2$$

Stochastic gradient descent (SGD): Given some $\{\alpha_t\}_{t\in\mathbb{N}_+}$, we:
  - Initialise $\theta_0 \in \Theta$.
  - For $t \in \mathbb{N}_+$, $\theta_{t+1} = \theta_t - \alpha_t \nabla_\theta \hat{J}_m(\theta)$

Computational Complexity

- Problem: These linear systems quickly become enormous when the number of samples m is large, or the dimension d is large. The computational cost is cubic in the number of parameters.

  Si, S., Oates, C. J., Duncan, A. B., Carin, L., & Briol, F.-X. (2020). Scalable control variates for Monte Carlo methods via stochastic optimization. arXiv:2006.07487.

- Significant speed-ups can be obtained by minimising the following objective through stochastic optimisation:

$$\arg\min_{\theta \in \Theta} \hat{J}_m(\theta) = \arg\min_{\theta \in \Theta} \widehat{\mathrm{Var}}_m[f - \mathcal{S}[u_\theta]] + \lambda_m \|\theta\|^2$$

$$\widehat{\mathrm{Var}}_m[f - \mathcal{S}[u_\theta]] = \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - \mathcal{S}[u_\theta] - \theta_0)^2$$

Stochastic gradient descent (SGD): Given some $\{\alpha_t\}_{t \in \mathbb{N}_+}$, we:

- Initialise $\theta_0 \in \Theta$.
- For $t \in \mathbb{N}_+$, $\theta_{t+1} = \theta_t - \alpha_t \nabla_\theta \hat{J}_m(\theta)$

Computational Complexity

- Problem: These linear systems quickly become enormous when the number of samples m is large, or the dimension d is large. The computational cost is cubic in the number of parameters.

  Si, S., Oates, C. J., Duncan, A. B., Carin, L., & Briol, F.-X. (2020). Scalable control variates for Monte Carlo methods via stochastic optimization. arXiv:2006.07487.

- Significant speed-ups can be obtained by minimising the following objective through stochastic optimisation:

$$\arg\min_{\theta \in \Theta} \hat{J}_m(\theta) = \arg\min_{\theta \in \Theta} \widehat{\mathrm{Var}}_m[f - \mathcal{S}[u_\theta]] + \lambda_m \|\theta\|^2$$

$$\widehat{\mathrm{Var}}_m[f - \mathcal{S}[u_\theta]] = \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - \mathcal{S}[u_\theta] - \theta_0)^2$$

  Stochastic gradient descent (SGD): Given some $\{\alpha_t\}_{t \in \mathbb{N}_+}$, we:
  - Initialise $\theta_0 \in \Theta$.
  - For $t \in \mathbb{N}_+$, $\theta_{t+1} = \theta_t - \alpha_t \nabla_\theta \hat{J}_m(\theta)$

Advantages of This Approach

- Stochastic optimisation can significantly reduce computational cost.
- We can do early stopping at any iteration. The CV corresponding to θ_t is always a valid CV (in the sense that $\int \mathcal{S}[u_{\theta_t}](x)\mathbb{P}(dx) = 0 \; \forall t$)
- Let $(\mathcal{S}_1, \mathcal{U}_1), \ldots, (\mathcal{S}_q, \mathcal{U}_q)$ be pairs of Stein operators/classes for $\mathbb{P}$. We can create very flexible families of CV as follows:

$$g = c_1 \mathcal{S}_1[u_1] + \ldots + c_1 \mathcal{S}_q[u_q]$$

satisfies $\Pi[g] = 0 \; \forall u_1 \in \mathcal{U}_1, \ldots, u_q \in \mathcal{U}_q$ and $c_1, \ldots, c_q \in \mathbb{R}$.

- The problem is convex whenever $\mathcal{V}_\Theta$ is linear in the parameters (e.g. polynomials and kernels). We can hence guarantee convergence. However, we can also use non-linear models such as neural networks:

Wan, R., Zhong, M., Xiong, H. and Zhu, Z. (2019). Neural control variates for Monte Carlo variance reduction. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 533-547.

Advantages of This Approach

- Stochastic optimisation can significantly reduce computational cost.
- We can do early stopping at any iteration. The CV corresponding to θ_t is always a valid CV (in the sense that $\int \mathcal{S}[u_{\theta_t}](x)\mathbb{P}(dx) = 0 \; \forall t$)
- Let $(\mathcal{S}_1, \mathcal{U}_1), \ldots, (\mathcal{S}_q, \mathcal{U}_q)$ be pairs of Stein operators/classes for $\mathbb{P}$. We can create very flexible families of CV as follows:

$$g = c_1 \mathcal{S}_1[u_1] + \ldots + c_1 \mathcal{S}_q[u_q]$$

satisfies $\Pi[g] = 0 \; \forall u_1 \in \mathcal{U}_1, \ldots, u_q \in \mathcal{U}_q$ and $c_1, \ldots, c_q \in \mathbb{R}$.

- The problem is convex whenever $\mathcal{V}_\Theta$ is linear in the parameters (e.g. polynomials and kernels). We can hence guarantee convergence. However, we can also use non-linear models such as neural networks:

Wan, R., Zhong, M., Xiong, H. and Zhu, Z. (2019). Neural control variates for Monte Carlo variance reduction. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 533-547.

Advantages of This Approach

- Stochastic optimisation can significantly reduce computational cost.
- We can do early stopping at any iteration. The CV corresponding to θ_t is always a valid CV (in the sense that $\int \mathcal{S}[u_{\theta_t}](x)\mathbb{P}(dx) = 0 \; \forall t$)
- Let $(\mathcal{S}_1, \mathcal{U}_1), \ldots, (\mathcal{S}_q, \mathcal{U}_q)$ be pairs of Stein operators/classes for $\mathbb{P}$. We can create very flexible families of CV as follows:

$$g = c_1 \mathcal{S}_1[u_1] + \ldots + c_1 \mathcal{S}_q[u_q]$$

satisfies $\Pi[g] = 0 \; \forall u_1 \in \mathcal{U}_1, \ldots, u_q \in \mathcal{U}_q$ and $c_1, \ldots, c_q \in \mathbb{R}$.

- The problem is convex whenever $\mathcal{V}_\Theta$ is linear in the parameters (e.g. polynomials and kernels). We can hence guarantee convergence. However, we can also use non-linear models such as neural networks:

Wan, R., Zhong, M., Xiong, H. and Zhu, Z. (2019). Neural control variates for Monte Carlo variance reduction. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 533-547.

Advantages of This Approach

- Stochastic optimisation can significantly reduce computational cost.
- We can do early stopping at any iteration. The CV corresponding to θ_t is always a valid CV (in the sense that $\int \mathcal{S}[u_{\theta_t}](x)\mathbb{P}(dx) = 0 \; \forall t$)
- Let $(\mathcal{S}_1, \mathcal{U}_1), \ldots, (\mathcal{S}_q, \mathcal{U}_q)$ be pairs of Stein operators/classes for $\mathbb{P}$. We can create very flexible families of CV as follows:

$$g = c_1 \mathcal{S}_1[u_1] + \ldots + c_1 \mathcal{S}_q[u_q]$$

satisfies $\Pi[g] = 0 \; \forall u_1 \in \mathcal{U}_1, \ldots, u_q \in \mathcal{U}_q$ and $c_1, \ldots, c_q \in \mathbb{R}$.

- The problem is convex whenever $\mathcal{V}_\Theta$ is linear in the parameters (e.g. polynomials and kernels). We can hence guarantee convergence. However, we can also use non-linear models such as neural networks:

Wan, R., Zhong, M., Xiong, H. and Zhu, Z. (2019). Neural control variates for Monte Carlo variance reduction. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 533-547.

Advantages of This Approach

- Stochastic optimisation can significantly reduce computational cost.
- We can do early stopping at any iteration. The CV corresponding to θ_t is always a valid CV (in the sense that $\int \mathcal{S}[u_{\theta_t}](x)\mathbb{P}(dx) = 0 \; \forall t$)
- Let $(\mathcal{S}_1, \mathcal{U}_1), \ldots, (\mathcal{S}_q, \mathcal{U}_q)$ be pairs of Stein operators/classes for $\mathbb{P}$. We can create very flexible families of CV as follows:

$$g = c_1 \mathcal{S}_1[u_1] + \ldots + c_1 \mathcal{S}_q[u_q]$$

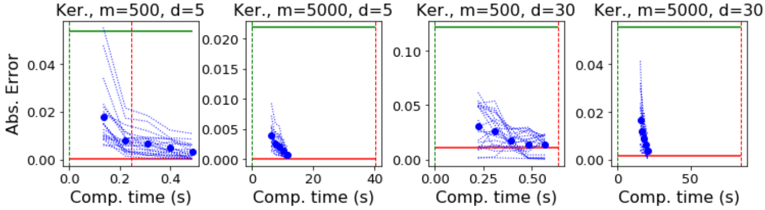
satisfies $\Pi[g] = 0 \; \forall u_1 \in \mathcal{U}_1, \ldots, u_q \in \mathcal{U}_q$ and $c_1, \ldots, c_q \in \mathbb{R}$.

- The problem is convex whenever $\mathcal{V}_\Theta$ is linear in the parameters (e.g. polynomials and kernels). We can hence guarantee convergence. However, we can also use non-linear models such as neural networks:

Wan, R., Zhong, M., Xiong, H. and Zhu, Z. (2019). Neural control variates for Monte Carlo variance reduction. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 533-547.

Computational Cost

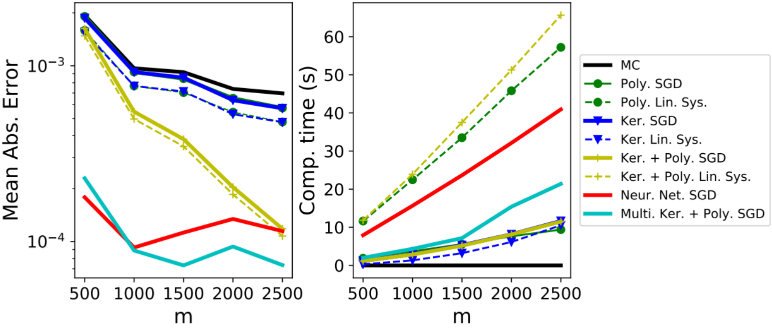
Problem: $\int_{\mathcal{X}} f(x)\mathbb{P}(dx)$ where $f(x) = \sum_{i=1}^{d} x_i$ and $\mathbb{P} = N(0,1)$.



Cost: linear system: $O(m^3 + m^2 d)$, Ours: $O(mdbt)$.

m is sample size, d is dimension, t is SGD steps, b is mini-batch size.

Bayesian Logistic Regression in $d = 61$



- Sonar dataset from UCI repository. Integral is over posterior distribution on coefficients to obtain the predictive distribution.
- The ensemble of kernel and polynomial CVs outperforms CVs based on neural nets. It is also easier to use since the objective is convex.

Other Applications

Some Other Applications of Stein's Method (#1)

Stein's Method has the potential of impacting Bayesian statistics

1) Diagnostic tools for MCMC:

Gorham, J., & Mackey, L. (2017). Measuring sample quality with kernels. International Conference on Machine Learning (pp. 1292-1301).

Gorham, J., Duncan, A., Mackey, L., & Vollmer, S. (2019). Measuring sample quality with diffusions. Annals of Applied Probability, 29(5), 2884-2928.

→ *Jackson Gorham won the Savage Award in 2017.*

2) Variational inference:

Liu, Q., and Wang, D. (2016). Stein variational gradient descent: A general purpose Bayesian inference algorithm. Neural Information Processing Systems (pp. 2378-2386).

Ranganath, R., Altosaar, J., Tran, D., & Blei, D. M. (2016). Operator variational inference. In Advances in Neural Information Processing Systems (pp. 496-504).

→ *Rajesh Ranganath won the Savage Award in 2018.*

Some Other Applications of Stein's Method (#1)

Stein's Method has the potential of impacting Bayesian statistics

1) Diagnostic tools for MCMC:

   Gorham, J., & Mackey, L. (2017). Measuring sample quality with kernels. International Conference on Machine Learning (pp. 1292-1301).

   Gorham, J., Duncan, A., Mackey, L., & Vollmer, S. (2019). Measuring sample quality with diffusions. Annals of Applied Probability, 29(5), 2884-2928.

   $\rightarrow$ *Jackson Gorham won the Savage Award in 2017.*

2) Variational inference:

   Liu, Q., and Wang, D. (2016). Stein variational gradient descent: A general purpose Bayesian inference algorithm. Neural Information Processing Systems (pp. 2378-2386).

   Ranganath, R., Altosaar, J., Tran, D., & Blei, D. M. (2016). Operator variational inference. In Advances in Neural Information Processing Systems (pp. 496-504).

   $\rightarrow$ *Rajesh Ranganath won the Savage Award in 2018.*

Some Other Applications of Stein's Method (#2)

3) Importance sampling:

Liu, Q., & Lee, J. D. (2017). Black-box importance sampling. In Proceedings of the International Conference on Artificial Intelligence and Statistics (pp. 952-961).

4) Thinning of MCMC chains:

Riabiz, M., Chen, W., Cockayne, J., Swietach, P., Niederer, S. A., Mackey, L., & Oates, C. J. (2020). Optimal thinning of MCMC output. arXiv:2005.03952.

5) Goodness-of-fit testing:

Chwialkowski, K., Strathmann, H., & Gretton, A. (2016). A kernel test of goodness of fit. International Conference on Machine Learning, 48, 2606-2615.

Liu, Q., Lee, J., & Jordan, M. (2016). A kernelized Stein discrepancy for goodness-of-fit tests and model evaluation. International Conference on Machine Learning (pp. 276-284).

Many others on the website of Yvik Swan:
https://sites.google.com/site/steinsmethod

Some Other Applications of Stein's Method (#2)

3) Importance sampling:

Liu, Q., & Lee, J. D. (2017). Black-box importance sampling. In Proceedings of the International Conference on Artificial Intelligence and Statistics (pp. 952-961).

4) Thinning of MCMC chains:

Riabiz, M., Chen, W., Cockayne, J., Swietach, P., Niederer, S. A., Mackey, L., & Oates, C. J. (2020). Optimal thinning of MCMC output. arXiv:2005.03952.

5) Goodness-of-fit testing:

Chwialkowski, K., Strathmann, H., & Gretton, A. (2016). A kernel test of goodness of fit. International Conference on Machine Learning, 48, 2606-2615.

Liu, Q., Lee, J., & Jordan, M. (2016). A kernelized Stein discrepancy for goodness-of-fit tests and model evaluation. International Conference on Machine Learning (pp. 276-284).

Many others on the website of Yvik Swan:
https://sites.google.com/site/steinsmethod

Some Other Applications of Stein's Method (#2)

3) Importance sampling:

Liu, Q., & Lee, J. D. (2017). Black-box importance sampling. In Proceedings of the International Conference on Artificial Intelligence and Statistics (pp. 952-961).

4) Thinning of MCMC chains:

Riabiz, M., Chen, W., Cockayne, J., Swietach, P., Niederer, S. A., Mackey, L., & Oates, C. J. (2020). Optimal thinning of MCMC output. arXiv:2005.03952.

5) Goodness-of-fit testing:

Chwialkowski, K., Strathmann, H., & Gretton, A. (2016). A kernel test of goodness of fit. International Conference on Machine Learning, 48, 2606-2615.

Liu, Q., Lee, J., & Jordan, M. (2016). A kernelized Stein discrepancy for goodness-of-fit tests and model evaluation. International Conference on Machine Learning (pp. 276-284).

Many others on the website of Yvik Swan:
https://sites.google.com/site/steinsmethod

Some Other Applications of Stein's Method (#2)

3) Importance sampling:

Liu, Q., & Lee, J. D. (2017). Black-box importance sampling. In Proceedings of the International Conference on Artificial Intelligence and Statistics (pp. 952-961).

4) Thinning of MCMC chains:

Riabiz, M., Chen, W., Cockayne, J., Swietach, P., Niederer, S. A., Mackey, L., & Oates, C. J. (2020). Optimal thinning of MCMC output. arXiv:2005.03952.

5) Goodness-of-fit testing:

Chwialkowski, K., Strathmann, H., & Gretton, A. (2016). A kernel test of goodness of fit. International Conference on Machine Learning, 48, 2606-2615.

Liu, Q., Lee, J., & Jordan, M. (2016). A kernelized Stein discrepancy for goodness-of-fit tests and model evaluation. International Conference on Machine Learning (pp. 276-284).

Many others on the website of Yvik Swan:
https://sites.google.com/site/steinsmethod

Conclusion

Take-Aways

- Intractable integrals pop up everywhere in Bayesian statistics and Stein's identity is a very useful trick to bypass them!
- Stein operators for $\mathbb{P}$ can be created from infinitesimal generators of Markov processes, many of which only require access to $\nabla_x \log p$. In particular, this means we do not need normalisation constants.
- I have highlighted applications for the approximation of posterior distributions, inference for unnormalised models, and control variates for MCMC. But there are many others...

Take-Aways

- Intractable integrals pop up everywhere in Bayesian statistics and Stein's identity is a very useful trick to bypass them!
- Stein operators for $\mathbb{P}$ can be created from infinitesimal generators of Markov processes, many of which only require access to $\nabla_x \log p$. In particular, this means we do not need normalisation constants.
- I have highlighted applications for the approximation of posterior distributions, inference for unnormalised models, and control variates for MCMC. But there are many others...

Take-Aways

- Intractable integrals pop up everywhere in Bayesian statistics and Stein's identity is a very useful trick to bypass them!
- Stein operators for $\mathbb{P}$ can be created from infinitesimal generators of Markov processes, many of which only require access to $\nabla_x \log p$. In particular, this means we do not need normalisation constants.
- I have highlighted applications for the approximation of posterior distributions, inference for unnormalised models, and control variates for MCMC. But there are many others...