# Control Functionals for Monte Carlo Integration

Oates, Girolami, Chopin

Arthur Gretton's notes

April 7, 2016

# What the paper is about

- A method for reducing variance of Monte Carlo estimates of the mean of a function $f(x)$ under $x \sim \pi$, where $x \in \mathbb{R}^d$.
- Standard mean estimate:

$$\hat{\mu}(f) := \frac{1}{n} \sum_{i=1}^{n} f(x_i)$$

converges to population expectation

$$\mu(f) := \int f(x)\pi(x)dx$$

with rate $O_P(n^{-1/2})$.

- Given a spare training sample, can we do better? Yes, if $f$ is smooth, $\pi$ satisfies certain conditions.

# Setting and main claim

- Split the data: $\mathcal{D}_0 := \{x_i\}_{i=1}^m$ , $\mathcal{D}_1 := \{x_i\}_{i=m+1}^n$. Ratio is $m = O(n^\gamma)$, optimal choice is $\gamma = 1$
- Learn a modified function

$$f_{\mathcal{D}_0} := f(x) - \hat{f}_{\mathcal{D}_0}(x) + \mu(\hat{f}_{\mathcal{D}_0}), \qquad \mu(f_{\mathcal{D}_0}) = \mu(f).$$

  where $\mu(\hat{f}_{\mathcal{D}_0})$ must be analytically computable.
- $\hat{f}_{\mathcal{D}_0}(x) - \mu(\hat{f}_{\mathcal{D}_0})$ is control functional (with zero expectation under $\pi$)
- Our estimate of $\mu(f)$ is:

$$\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1, f) := \frac{1}{n-m} \sum_{i=m+1}^{n} f_{\mathcal{D}_0}(x_i)$$

# Setting and main claim

Given we can learn $\hat{f}_{\mathcal{D}_0}$ with error

$$\mathbb{E}_{\mathcal{D}_0}\left[\sigma^2(f - \hat{f}_{\mathcal{D}_0})\right] = O(m^{-\delta}). \tag{1}$$

(need smoothness asumption on $f$). Then

$$\mathbb{E}_{\mathcal{D}_0}\mathbb{E}_{\mathcal{D}_1}\left[\left(\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1, f) - \mu(f)\right)^2\right] = O(n^{-1-\delta}).$$

# Setting and main claim

Given we can learn $\hat{f}_{\mathcal{D}_0}$ with error

$$\mathbb{E}_{\mathcal{D}_0}\left[\sigma^2(f - \hat{f}_{\mathcal{D}_0})\right] = O(m^{-\delta}). \tag{1}$$

(need smoothness asumption on $f$). Then

$$\mathbb{E}_{\mathcal{D}_0}\mathbb{E}_{\mathcal{D}_1}\left[(\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1, f) - \mu(f))^2\right] = O(n^{-1-\delta}).$$

Proof: By construction, $\mathbb{E}\left(\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1, f)\right) = \mu(f)$ so

$$\mathbb{E}_{\mathcal{D}_1}\left[(\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1, f) - \mu(f))^2\right] = \frac{1}{n-m}\sigma^2(f - \hat{f}_{\mathcal{D}_0}).$$

Thus

$$\mathbb{E}_{\mathcal{D}_0}\mathbb{E}_{\mathcal{D}_1}\left[|\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1, f) - \mu(f)|^2\right] = \frac{1}{n-m}\mathbb{E}_{\mathcal{D}_0}\left[\sigma^2(f - \hat{f}_{\mathcal{D}_0})\right].$$

- Then use (1) and $(n-m)^{-1} = O(n^{-1})$.

# The Stein way

How to define a function class for $\hat{f}_{\mathcal{D}_o}$?

# The Stein way

How to define a function class for $\hat{f}_{\mathcal{D}_o}$?
Define:

$$u(x) := \nabla_x \log \pi(x) \qquad \nabla_x := \left[ \begin{array}{ccc} \partial/\partial x_1 & \ldots & \partial/\partial x_d \end{array} \right]^\top.$$

Given a vector-valued function $\phi(x) : \mathbb{R}^d \to \mathbb{R}^d$, define

$$\begin{aligned}\psi(x) &:= \sum_{i=1}^d \frac{\partial}{\partial x_i}\phi_i(x) + \sum_{i=1}^d \phi_i(x)\frac{\partial}{\partial x_i}\log \pi(x)\\ &= \nabla_x^\top \phi(x) + \phi(x)^\top u(x).\end{aligned}$$

Assume boundary condition: given $n(x) \in \mathbb{R}^d$ normal to the boundary,

$$\oint_{\partial \mathcal{X}} \pi(x) \left[\phi(x)^\top n(x)\right] dS(x) = 0$$

# The Stein way (cont'd)

Then

$$\int \psi(x)\pi(x)dx = 0,$$

exactly the property we want for $\hat{f}_{\mathcal{D}_o} - \mu(\hat{f}_{\mathcal{D}_0})$.

# The Stein way (cont'd)

Then

$$\int \psi(x)\pi(x)dx = 0,$$

exactly the property we want for $\hat{f}_{\mathcal{D}_o} - \mu(\hat{f}_{\mathcal{D}_0})$.
Proof: from prev. slide,

$$\psi(x) = \nabla_x^\top \phi(x) + \phi(x)^\top \left(\nabla_x \log \pi(x)\right).$$

Using divergence theorem in (b),

$$\int \psi(x)\pi(x)dx \overset{(a)}{=} \int \nabla_x^\top \left[\phi(x)\pi(x)\right] dx \overset{(b)}{=} \oint_{\partial\mathcal{X}} \pi(x) \left[\phi(x)^\top n(x)\right] dS(x) = 0$$

using in (a) that

$$\frac{\partial}{\partial x_i} \log \pi(x) = \frac{1}{\pi(x)} \frac{\partial \pi(x)}{\partial x_i}.$$

# Stein-modified function class, kernel version

What is a good function class for entries of $\phi(x) : \mathbb{R}^d \to \mathbb{R}^d$?

# Stein-modified function class, kernel version

What is a good function class for entries of $\phi(x)\ :\ \mathbb{R}^d \to \mathbb{R}^d$?
Consider $\phi(x) \in \mathcal{H}^d$ with inner product

$$\left\langle \phi(x), \phi(x') \right\rangle_{\mathcal{H}^d} = \sum_{i=1}^{d} \left\langle \phi_i(x), \phi_i(x') \right\rangle .$$

Then $\psi(x) \in \mathcal{H}_0$, a new RKHS with kernel

$$k_0(x,x') = \sum_{i=1}^{d} \frac{\partial k(x,x')}{\partial x_i \partial x'_i} + u_i(x)\frac{\partial k(x,x')}{\partial x'_i} + u_i(x')\frac{\partial k(x,x')}{\partial x_i} + u_i(x)u_i(x')k(x,x')$$

Proof: write as $k(x,\cdot)$ as the feature map of $\mathcal{H}$, so

$$k(x,x') = \left\langle k(x,\cdot), k(x',\cdot) \right\rangle_{\mathcal{H}} .$$

# Stein-modified function class, kernel version (cont'd)

Then

$$\begin{aligned}\psi(x) &= \sum_{i=1}^{d} \frac{\partial}{\partial x_i}\phi_i(x) + \sum_{i=1}^{d} \phi_i(x)\frac{\partial}{\partial x_i}\log\pi(x) \\ &= \sum_{i=1}^{d} \left\langle \phi_i, \frac{\partial}{\partial x_i}k(x,\cdot) + k(x,\cdot)\underbrace{\frac{\partial}{\partial x_i}\log\pi(x)}_{u_i(x)} \right\rangle,\end{aligned}$$

Thus

$$k_0(x,x') = \sum_{i=1}^{d} \left\langle \frac{\partial}{\partial x_i}k(x,\cdot) + k(x,\cdot)u_i(x), \frac{\partial}{\partial x_i'}k(x',\cdot) + k(x',\cdot)u_i(x') \right\rangle_{\mathcal{H}}$$

Only need $\pi$ up to normalizing constant to compute kernel.

# Stein-modified function class, kernel version (cont'd)

Under boundary conditions

$$
\begin{aligned}
0_d &= \oint_{\partial\mathcal{X}} k(x,x')\pi(x')n(x')dS(x') \\
0 &= \oint_{\partial\mathcal{X}} \nabla_x k(x,x')^\top n(x')\pi(x')dS(x')
\end{aligned}
$$

we have

$$\int_{\mathcal{X}} k_0(x,x')\pi(x')dx' = 0.$$

Recall for all RHKS functions in $\mathcal{H}_o$,

$$\psi(x) \in \overline{\left\{ \sum_{i=1}^{\ell} \alpha_i k_0(x,x_i) \ : \ \ell \in \mathbb{N} \right\}}$$

so as required, $\mu(\psi(x)) = 0$.

# Regression problem for $\hat{f}_{\mathcal{D}_0}$

Define $\mathcal{H}_+ := C \oplus \mathcal{H}_0$, where $\mathcal{C}$ are constant functions. I.e. $f \in \mathcal{H}_+$ when $f = \psi + c$ and $\psi \in \mathcal{H}_0$.

$$\|f\|_{\mathcal{H}} := \|\psi\|_{\mathcal{H}_0} + \|c\|_{\mathcal{C}} = \|\psi\|_{\mathcal{H}_0} + |c| \,.$$

Then regression problem is

$$\hat{f}_{\mathcal{D}_0} := \operatorname*{argmin}_{g \in \mathcal{H}_+} \left\{ \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - g(x_i))^2 + \lambda \|g\|^2_{\mathcal{H}_+} \right\}$$

# Regression problem for $\hat{f}_{\mathcal{D}_0}$

Define $\mathcal{H}_+ := C \oplus \mathcal{H}_0$, where $\mathcal{C}$ are constant functions. I.e. $f \in \mathcal{H}_+$ when $f = \psi + c$ and $\psi \in \mathcal{H}_0$.

$$\|f\|_{\mathcal{H}} := \|\psi\|_{\mathcal{H}_0} + \|c\|_{\mathcal{C}} = \|\psi\|_{\mathcal{H}_0} + |c| \,.$$

Then regression problem is

$$\hat{f}_{\mathcal{D}_0} := \operatorname*{argmin}_{g \in \mathcal{H}_+} \left\{ \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - g(x_i))^2 + \lambda \|g\|^2_{\mathcal{H}_+} \right\}$$

From Sun and Wu (2009),

$$\mathbb{E}_{\mathcal{D}_0} \left[ \sigma^2 (f - \hat{f}_{\mathcal{D}_0}) \right] = O(m^{-1/6})$$

- if $f \in \mathcal{H}_+$ (i.e. true $f$ is smooth)
- $\sup_{x \in \mathcal{X}} k_+(x, x) < \infty$, $\lambda = O(m^{-1/2})$

Thus

$$\mathbb{E}_{\mathcal{D}_0} \mathbb{E}_{\mathcal{D}_1} \left[ |\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1, f) - \mu(f)|^2 \right] = O(n^{-7/6}).$$

# An application: GP regression, marginalized hyperparameters

GP regression:

$$\hat{Y}^* := \mathbb{E}([Y^*|\mathbf{y}, \mathbf{x}, x^*]) = \int \underbrace{\mathbb{E}([Y^*|\mathbf{y}, \mathbf{x}, x^*, \theta])}_{f(\theta)} \pi(\theta) d\theta,$$

- Integral over $\pi$ unavailable in closed form.
- Each evaluation of $\mathbb{E}([Y^*|\mathbf{y}, \mathbf{x}, x^*, \theta])$ is expensive,

$$\mathbb{E}([Y^*|\mathbf{y}, \mathbf{x}, x^*, \theta]) = C_{*,N} \left(C_N + \sigma^2 I\right)^{-1} \mathbf{y},$$

$(C_N)_{ij} = \mathfrak{K}(x_i, x_j)$ and $(C_{*N})_i = k(x^*, x_i)$.

# An application: GP regression, marginalized hyperparameters