Approximate Shannon Sampling in Importance Sampling: Nearly Consistent Finite Particle Estimates

Amrit Singh Bedi[†], **Alec Koppel**[†], Victor Elvira[§], Brian M. Sadler[†]

[†]U.S. Army Research Laboratory
[§] University of Edinburgh

Advances in Bayesian Machine Learning
Asilomar Conference on Signals, Systems, and Computers

Nov 4, 2019

Bayesian Methods

Supervised learning, map features to targets $\mathbf{y} \mapsto \hat{x} = f(\mathbf{y})$

$\Rightarrow$ found by minimizing loss $\ell(\hat{x}, x)$ averaged over data $(\mathbf{y}, x)$

$\rightarrow$ Bayesian methods ask: given $\{(\mathbf{y}_u, x_u)\})_{u<t}$, observe $\mathbf{y}_t$

$\Rightarrow$ how to form posterior distribution $\mathbb{P}(x_t \,\big|\, \{\mathbf{y}_u, x_u\}_{u<t} \cup \{\mathbf{y}_t\})$

$\rightarrow$ Needed for computing confidence intervals, quantiles, etc.

$\Rightarrow$ robustness/safety gaurentees, uncertainty-aware planning

$\Rightarrow$ foundation of climate forecasting, SLAM, robust MPC

Bayesian Methods

Can easily predict mean when dynamics are linear with AWGN

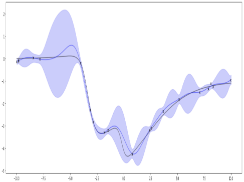
$\Rightarrow$ Kalman filter

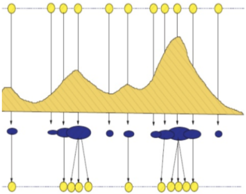
$\rightarrow$ In many modern applications, dynamics inherently nonlinear

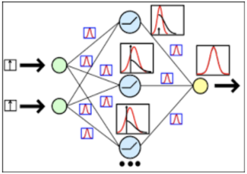
$\Rightarrow$ legged robotics, indoor localization, meterology

$\rightarrow$ How to estimate arbitrary posterior $\mathbb{P}(x_t \,\big|\, \{\mathbf{y}_u, x_u\}_{u<t} \cup \{\mathbf{y}_t\})$?

$\Rightarrow$ GPs, Monte Carlo, "Bayesian deep networks"







Bayesian Methods

Can easily predict mean when dynamics are linear with AWGN

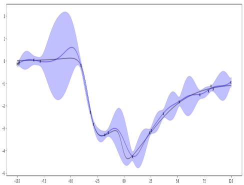
$\Rightarrow$ Kalman filter

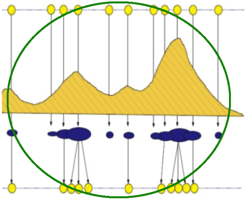
$\rightarrow$ In many modern applications, dynamics inherently nonlinear

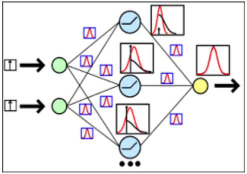
$\Rightarrow$ legged robotics, indoor localization, meterology

$\rightarrow$ How to estimate arbitrary posterior $\mathbb{P}(y \,\big|\, \{\mathbf{x}_u, y_u\}_{u \leq t} \cup \{\mathbf{x}_t\})$?

$\Rightarrow$ GPs, Monte Carlo, "Bayesian deep networks"







Monte Carlo

Bayesian inference $\Rightarrow$ Compute integral via samples $\{\mathbf{y}_k\}_{k\leq K}$

$$I(\phi) = \mathbb{E}_{\mathbf{x}}[\phi(\mathbf{x}) \,|\, \mathbf{y}] = \int_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x})\mathbb{P}(\mathbf{x}|\mathbf{y})d\mathbf{x}$$

$\rightarrow$ $\phi : \mathbb{R}^p \rightarrow \mathbb{R}$ is arbitrary, $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^p$ is a random variable

$\Rightarrow$ $\phi(\mathbf{x}) = \mathbf{x}$ for posterior mean, $\phi(\mathbf{x}) = \mathbf{x}^p$ for p-th moment.

$\rightarrow$ To compute integral, require posterior distribution

$$\mathbb{P}\left(\mathbf{x} \,\middle|\, \{\mathbf{y}_k\}_{k\leq K}\right) = \frac{\mathbb{P}\left(\{\mathbf{y}_k\}_{k\leq K} \,\middle|\, \mathbf{x}\right)\mathbb{P}\left(\mathbf{x}\right)}{\mathbb{P}\left(\{\mathbf{y}_k\}_{k\leq K}\right)}.$$

$\rightarrow$ When $\mathbb{P}(\mathbf{x} \,|\, \mathbf{y})$ is unknown, integral $I(\phi)$ cannot be evaluated

$\Rightarrow$ must resort to numerical integration, aka Monte Carlo

Curse of Dimensionality

Monte Carlo methods have complexity issues

$\Rightarrow$ consistency requires no. of particles $\rightarrow \infty$

$\Rightarrow$ posterior keeps past particles $\Rightarrow$ complexity $\approx$ no. particles

$\rightarrow$ Adaptive proposal to reduce bias [Bugallo et al ‘17]

$\rightarrow$ No. samples ensure specific bias [Agapiou et al, ‘17]

$\Rightarrow$ many other works along these lines

$\rightarrow$ statistics to diagnose estimate quality [Kong ‘92, Elvira ‘18].

$\rightarrow$ **Main drawback**: costly form for empirical measures

$\Rightarrow$ each sample from proposal into particle representation

Approximation Strategy

Emp. estimate for the cond. dist. is $\mu_n = \sum_{u=1}^{n} \bar{w}^{(u)} \delta_{\mathbf{x}^{(u)}}$

$\rightarrow$ Deltas have no "volume," $\Rightarrow$ no finite cover of $\mathcal{X}^{\text{compact}}$

$\rightarrow$ Kernel smoothing to replace deltas by kernels $\kappa : \mathcal{X} \rightarrow \mathbb{R}$

$$\hat{\mu}_n \approx \sum_{u=1}^{n} \bar{w}^{(u)} \kappa_{\mathbf{x}^{(u)}}(\mathbf{x})$$

$\rightarrow$ Once we have KDEs, propose sequential projection scheme

$\rightarrow$ Allows us to keep track of active set of particles

$\Rightarrow$ no. of particles grows/shrinks w.r.t. role in estimation error

$\Rightarrow$ trade off statistical bias ϵ w/ number of required particles

Controlling Bias

A geometric view

$\rightarrow$ Learning update rule

$$\hat{\mu}_n = \tilde{\mu}_{n-1} + g(\mathbf{x}^{(n)})\kappa_{\mathbf{x}^{(n)}}(\mathbf{x})$$

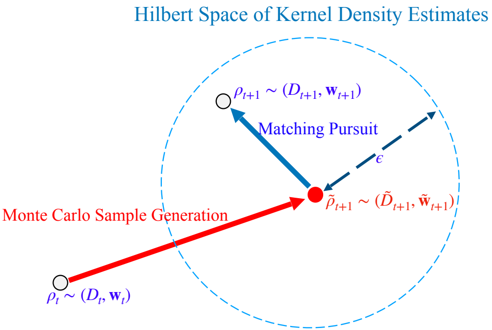
$\rightarrow$ Compress w.r.t. metric

$\Rightarrow$ causing ϵ error

$\Rightarrow$ add latest pt: $\tilde{\mathbf{D}}_n = [\mathbf{D}_{n-1}; \mathbf{x}^{(n)}]$

$\rightarrow$ Compressed $\tilde{\mu}_n$ such that

$$\|\tilde{\mu}_n - \hat{\mu}_n\|_{\mathcal{H}} \leq \epsilon$$



Importance Sampling Basics

Define posterior $q(\mathbf{x}|\mathbf{y}) := q(\mathbf{x}) = \mathbb{P}\left(\mathbf{x} \,\middle|\, \mathbf{y}\right)$

$\Rightarrow$ un-normalized $\tilde{q}(\mathbf{x}) := \tilde{q}(\mathbf{x} \mid \mathbf{y}) = \mathbb{P}\left(\{\mathbf{y}_k\}_{k \leq K} \,\middle|\, \mathbf{x}\right) \mathbb{P}\left(\mathbf{x}\right)$

$\Rightarrow$ normalizing constant $Z := \mathbb{P}\left(\{\mathbf{y}_k\}_{k \leq K}\right)$.

$\rightarrow$ Typically hypothesize likelihood model $\mathbb{P}(\{\mathbf{y}_k\}_{k \leq K} \mid \mathbf{x})$

$\Rightarrow$ for observations $\{\mathbf{y}_k\}$ drawn from a static dist. $\mathbb{P}(\mathbf{y} \mid \mathbf{x})$

$\Rightarrow$ prior for $\mathbb{P}(\mathbf{x})$.

$\rightarrow$ Example priors/likelihoods: Gaussian, Student's t, Uniform.

Importance Sampling (IS)

Def. *importance dist.* $\pi(\mathbf{x})$ w/ support of true density $q(\mathbf{x})$

$\Rightarrow$ Multiply and divide by $\pi(\mathbf{x})$ inside the integral

$$\int_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x})q(\mathbf{x})d\mathbf{x} = \int_{\mathbf{x}\in\mathcal{X}} \frac{\phi(\mathbf{x})q(\mathbf{x})}{\pi(\mathbf{x})}\pi(\mathbf{x})d\mathbf{x},$$

$\Rightarrow$ $q(\mathbf{x})/\pi(\mathbf{x})$ unnormalized density of target q w.r.t. proposal π

$\rightarrow$ Instead of requiring samples from true posterior $\mathbf{x}^{(n)} \sim q(\mathbf{x})$

$\Rightarrow$ only sample from importance dist. $\mathbf{x}^{(n)} \sim \pi(\mathbf{x})$, $n = 1, ..., N$,

$$\widehat{I}_N(\phi) := \frac{1}{N}\sum_{n=1}^{N} \frac{q(\mathbf{x}^{(n)})}{\pi(\mathbf{x}^{(n)})}\phi(\mathbf{x}^{(n)}) = \frac{1}{NZ}\sum_{n=1}^{N} g(\mathbf{x}^{(n)})\phi(\mathbf{x}^{(n)}),$$

$\Rightarrow$ where $g(\mathbf{x}^{(n)}) := \frac{q(\mathbf{x}^{(n)})}{\pi(\mathbf{x}^{(n)})}$ are the importance weights.

Importance Sampling (IS)

Def. *importance dist.* $\pi(\mathbf{x})$ w/ support of true density $q(\mathbf{x})$

$\rightarrow$ Instead of requiring samples from true posterior $\mathbf{x}^{(n)} \sim q(\mathbf{x})$

$\Rightarrow$ only sample from importance dist. $\mathbf{x}^{(n)} \sim \pi(\mathbf{x}), \; n = 1, \ldots, N,$

$$\widehat{I}_N(\phi) := \frac{1}{N} \sum_{n=1}^{N} \frac{q(\mathbf{x}^{(n)})}{\pi(\mathbf{x}^{(n)})} \phi(\mathbf{x}^{(n)}) = \frac{1}{NZ} \sum_{n=1}^{N} g(\mathbf{x}^{(n)}) \phi(\mathbf{x}^{(n)}),$$

$\Rightarrow$ where $g(\mathbf{x}^{(n)}) := \frac{q(\mathbf{x}^{(n)})}{\pi(\mathbf{x}^{(n)})}$ are the importance weights.

$\Rightarrow$ In practice, don't know $q(\mathbf{x}^{(n)})$ $\Rightarrow$ calculate via Bayes rule

$$q(\mathbf{x}^{(n)}) = \frac{\mathbb{P}\left(\{\mathbf{y}_k\}_{k \le K} \,\middle|\, \mathbf{x}^{(n)}\right) \mathbb{P}\left(\mathbf{x}^{(n)}\right)}{\int \mathbb{P}\left(\{\mathbf{y}_k\}_{k \le K} \,\middle|\, \mathbf{x}\right) \mathbb{P}\left(\mathbf{x}\right) d\mathbf{x}}.$$

$\rightarrow$ importance weights $g(\mathbf{x}^{(n)}) := \mathbb{P}(\{\mathbf{y}_k\}_{k \le K} | \mathbf{x}^{(n)}) \mathbb{P}(\mathbf{x}^{(n)}) / \pi(\mathbf{x}^{(n)})$.

$\rightarrow$ Estimator for normalizing constant $\hat{Z} := \frac{1}{N} \sum_{n=1}^{N} g(\mathbf{x}^{(n)})$.

Self-Normalized IS

Require Model $\mathbb{P}(\mathbf{y}\,|\,\mathbf{x})$, prior $\mathbb{P}(\mathbf{x})$, imp. dist. $\pi(\mathbf{x})$, obs. $\{\mathbf{y}_k\}_{k=1}^K$

$\Rightarrow$ **For** $n = 0, 1, 2, \ldots$

$\Rightarrow$ Simulate one sample from importance dist. $\mathbf{x}^{(n)} \sim \pi(\mathbf{x})$

$\Rightarrow$ Compute weight $g(\mathbf{x}^{(n)}) := \mathbb{P}(\{\mathbf{y}_k\}_{k\leq K} \big| \mathbf{x}^{(n)})\mathbb{P}(\mathbf{x}^{(n)})/\pi(\mathbf{x}^{(n)})$.

$\Rightarrow$ Normalized weights $\bar{w}^{(n)}$ by dividing by normalizing factor

$$\bar{w}^{(n)} = \frac{g(\mathbf{x}^{(n)})}{\sum_{u=1}^{n} g(\mathbf{x}^{(u)})} \quad \text{for all } n.$$

$\Rightarrow$ Form self-normalized IS estimate $I_n(\phi)$, posterior est. μ_n

$$I_n(\phi) = \sum_{u=1}^{n} \bar{w}^{(u)} \phi(\mathbf{x}^{(u)}) \ , \mu_n = \sum_{u=1}^{n} \bar{w}^{(u)} \delta_{\mathbf{x}^{(u)}}$$

Particle Selection Scheme

→ SNIS weight and dictionary update

$$\tilde{\mathbf{g}}_n = [\mathbf{g}_{n-1}; g(\mathbf{x}^n)] \;, \quad \tilde{\mathbf{w}}_n = z_n \tilde{\mathbf{g}}_n \;, \qquad \tilde{\mathbf{D}}_n = [\mathbf{D}_{n-1}; \mathbf{x}^{(n)}]$$

→ Unnormalized posterior density $\tilde{\mu}_n$ we can write

$$\begin{aligned} \tilde{\mu}_n &= \underset{y \in \mathcal{H}}{\operatorname{argmin}} \left\| y - \left(\tilde{\mu}_{n-1} + g(\mathbf{x}^{(n)})\delta_{\mathbf{x}^{(n)}}\right) \right\|^2 \\ &= \underset{y \in \mathcal{H}_{\mathbf{x}_n}}{\operatorname{argmin}} \left\| y - \left(\tilde{\mu}_{n-1} + g(\mathbf{x}^{(n)})\delta_{\mathbf{x}^{(n)}}\right) \right\|^2 \end{aligned}$$

Particle Selection Scheme

→ SNIS weight and dictionary update

$$\tilde{\mathbf{g}}_n = [\mathbf{g}_{n-1}; g(\mathbf{x}^n)] , \quad \tilde{\mathbf{w}}_n = z_n \tilde{\mathbf{g}}_n , \qquad \tilde{\mathbf{D}}_n = [\mathbf{D}_{n-1}; \mathbf{x}^{(n)}]$$

→ Two sources of approximation:

⇒ (1) Replace deltas by kernels (kernel smoothing)

⇒ (2) Subspace projection step

$$\hat{\mu}_n = \underset{y \in \mathcal{H}_{\mathbf{D}_n}}{\operatorname{argmin}} \left\| y - \left(\tilde{\mu}_{n-1} + g(\mathbf{x}^{(n)}) \delta_{\mathbf{x}^{(n)}}\right) \right\|^2$$

$$:= \mathcal{P}_{\mathcal{H}_{\mathbf{D}_n}} \left[\tilde{\mu}_{n-1} + g(\mathbf{x}^{(n)}) \phi_{\mathbf{x}}(\mathbf{x}^{(n)}) \right]$$

⇒ But how is the subspace of points $\mathcal{H}_{\mathbf{D}_n}$ chosen??

Compressed Kernelized Importance Sampling (CKIS)

A geometric view

→ Learning update rule

$$\hat{\mu}_n = \tilde{\mu}_{n-1} + g(\mathbf{x}^{(n)})\kappa_{\mathbf{x}^{(n)}}(\mathbf{x})$$

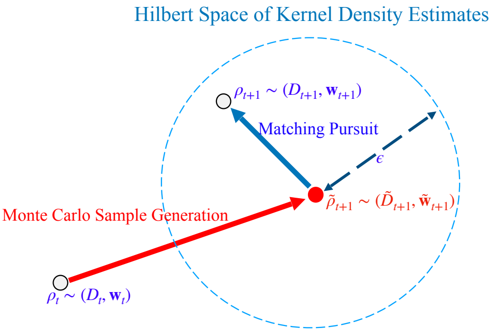
→ Compress w.r.t. RKHS norm

⇒ causing ϵ error

⇒ add latest pt: $\tilde{\mathbf{D}}_n = [\mathbf{D}_{n-1}; \mathbf{x}^{(n)}]$

→ KOMP output $\tilde{\mu}_n$ such that

$$\|\tilde{\mu}_n - \hat{\mu}_n\|_{\mathcal{H}} \leq \epsilon$$



Compressed Kernelized Importance Sampling (CKIS)

Require Model $\mathbb{P}(\mathbf{y} \,|\, \mathbf{x})$, prior $\mathbb{P}(\mathbf{x})$, imp. dist. $\pi(\mathbf{x})$, obs. $\{\mathbf{y}_k\}_{k=1}^K$

$\Rightarrow$ **For** $n = 0, 1, 2, \ldots$

$\Rightarrow$ Simulate one sample from importance dist. $\mathbf{x}^{(n)} \sim \pi(\mathbf{x})$

$\Rightarrow$ Compute weight $g(\mathbf{x}^{(n)}) := \mathbb{P}(\{\mathbf{y}_k\}_{k \le K} \,|\, \mathbf{x}^{(n)}) \mathbb{P}(\mathbf{x}^{(n)}) / \pi(\mathbf{x}^{(n)})$.

$\Rightarrow$ Normalized weights $\bar{w}^{(n)} = \frac{g(\mathbf{x}^{(n)})}{\sum_{u=1}^{n} g(\mathbf{x}^{(u)})}$ for all n

$\Rightarrow$ Form self-normalized IS estimate $I_n(\phi)$, posterior est. μ_n

$\Rightarrow$ Update kernel density via last sample & weight

$$\hat{\mu}_n = \tilde{\mu}_{n-1} + g(\mathbf{x}^{(n)}) \kappa_{\mathbf{x}^{(n)}}(\mathbf{x})$$

$\Rightarrow$ Revise $\tilde{\mathbf{D}}_n = [\mathbf{D}_{n-1}; \mathbf{x}^{(n)}]$ and $\tilde{\mathbf{g}}_n = [\mathbf{g}_{n-1}; g(\mathbf{x}^{(n)})]$

$\Rightarrow$ Compress kernel density estimate sequence as

$$(\tilde{\mu}_n, \mathbf{D}_n, \mathbf{g}_n) = \textbf{KOMP}(\hat{\mu}_n, \tilde{\mathbf{D}}_n, \tilde{\mathbf{g}}_n, \epsilon_n)$$

$\Rightarrow$ Normalized weights to ensure valid prob. measure $\tilde{\mathbf{w}}_n$

$\Rightarrow$ Estimate the expectation as

$$\hat{I}_n = \sum_{u=1}^{|\mathbf{D}_n|} \bar{w}^{(u)} \phi(\mathbf{x}^{(u)})$$

Theorem

The integral estimate iterates of CKIS exhibits posterior contraction as

$$\left| \sup_{|\phi| \le 1} \left(\mathbb{E}[\hat{I}_N(\phi) - I(\phi)] \right) \right| \le \mathcal{O}\left(\epsilon + \sigma_\kappa^2 h^2 + \frac{1}{\sqrt{Nh}} + \frac{1}{\sqrt{N}} + h^3 \right)$$

Algorithm is consistent when $\epsilon \to 0$, $h \to 0$ *as* $N \to \infty$

$\Rightarrow$ for constant compression budget, converge to ϵ bias

$\Rightarrow$ additional bias due to kernel smoothing parameter h

$\Rightarrow$ subsampling error $\approx 1/\sqrt{N}$ $\Rightarrow$ law of large numbers rate

Balancing Consistency and Memory

Theorem

Denote M_n as model order generated after n particles generated from importance density $\pi(\mathbf{x})$. For compact feature space $\mathcal{X}$ and bounded importance weights $g(\mathbf{x}^{(n)})$, $M_n < \infty$ for all n.

Merit of constant compression budget: provable finite memory

$\Rightarrow$ characterizing tradeoff of memory/consistency is difficult

$\Rightarrow$ depends on kernel hyperparameters, feature space radius

$\rightarrow$ Remaining open problem: how to establish this dependence

Experiments

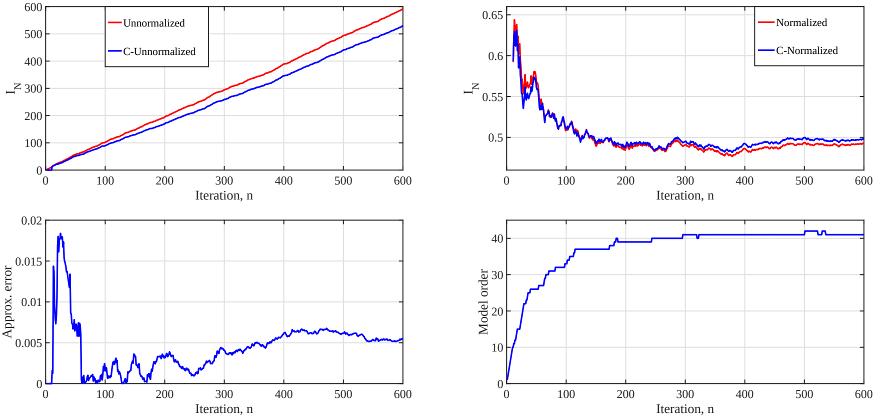
$\rightarrow$ Estimate the expected value of function $\phi(x)$

$\Rightarrow$ target $q(x)$ and the proposal $\pi(x)$ as

$$\phi(x) = 2\sin\left(\frac{\pi}{(1.5x)}\right)$$
$$q(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-1)^2}{2}\right)$$
$$\pi(x) = \frac{1}{\sqrt{4\pi}}\exp\left(-\frac{(x-1)^2}{4}\right)$$

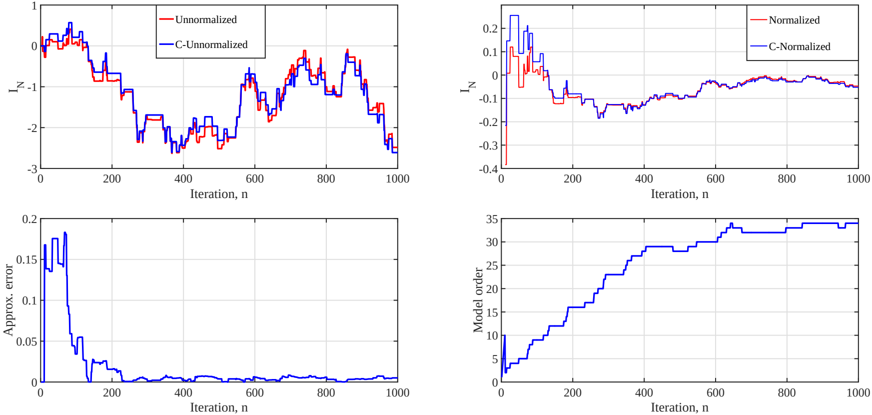
$\rightarrow$ Gaussian kernel ($h = 0.01$) and comp. budget $\epsilon = 3.5$

Experiments: Direct IS



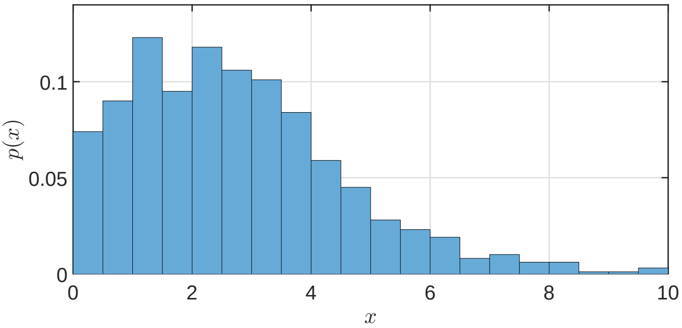
- $\Rightarrow$ $q(x)$ is known
- $\Rightarrow$ Gaussian kernel ($h = 0.01$) and comp. budget $\epsilon = 3.5$
- $\Rightarrow$ Performance is similar with only 7% samples

Experiments: Indirect IS



- ⇒ Estimate $q(\mathbf{x})$ via Bayes' rule
- ⇒ Gaussian kernel ($h = 0.01$) and comp. budget $\epsilon = 3.5$
- ⇒ Performance is similar with only 6% samples

Histogram of Particle Distribution



Histogram of resulting distribution

$\Rightarrow$ efficient rep. of arbitrary function of Gaussian distribution

Conclusion

Monte Carlo methods $\Rightarrow$ often used in autonomy/robotics

$\Rightarrow$ curse of dimensionality: complexity $\approx$ number of particles

$\Rightarrow$ a challenge common to nonparametric/Bayesian methods

$\rightarrow$ Precludes use in streaming settings

$\rightarrow$ Existing statistical tests, require inner-loop sub-sampling

$\Rightarrow$ Inefficient, missing bias characterization

$\rightarrow$ CKIS trades off consistency and memory

$\rightarrow$ Experiments $\Rightarrow$ CKIS and full Monte Carlo are comparable

$\rightarrow$ Future directions: employ CKIS in ML applications

$\Rightarrow$ off-policy evaluation in RL, weight batches of stoch. grads.

References

A. Koppel, A.B. Singh, K. Rajawat, and B.M. Sadler, "Approximate Shannon Sampling in Importance Sampling: Nearly Consistent Finite Particle Estimates," in Statistics and Computing (submitted), 2019.

Assumptions and Technicalities

Assumption

Denote the integral of test function $\phi : \mathcal{X} \to \mathbb{R}$ as $q(\phi)$.

- *Assume that ϕ is absolutely integrable, i.e., ϕ $q(|\phi|) < \infty$, and has absolute value at most unit $|\phi| \leq 1$.*
- *The test function has absolutely continuous second derivative, and $\int_{\mathbf{x}\in\mathcal{X}} \phi'''(\mathbf{x})d\mathbf{x} < \infty$.*

Assumption

Kernel is chosen such that $\int_{\mathbf{x}\in\mathcal{X}} \kappa_{\mathbf{x}^{(n)}}(\mathbf{x}) = 1$, $\int_{\mathbf{x}\in\mathcal{X}} \mathbf{x}\kappa_{\mathbf{x}^{(n)}}(\mathbf{x}) = 0$, and $\sigma_\kappa^2 = \int_{\mathbf{x}\in\mathcal{X}} \mathbf{x}^2\kappa_{\mathbf{x}^{(n)}}(\mathbf{x}) > 0$.

Assumption

The RKHS norm between full distributions lower-bounds the distance between their mean embeddings: $\|\hat{\rho} - \tilde{\rho}\|_{\mathcal{H}} \leq \|\hat{m} - \tilde{m}\|_{\mathcal{H}}$, which are related by a multiplicative factor $\|\hat{\rho} - \tilde{\rho}\|_{\mathcal{H}} = K\|\hat{m} - \tilde{m}\|_{\mathcal{H}}$.