2nd Symposium on Advances in Approximate Bayesian Inference, 2019 1–7

# A Projection Operator to Balance Consistency and Complexity in Importance Sampling

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

## Abstract

Importance sampling (IS) is a standard Monte Carlo (MC) tool to compute information about random variables such as moments or quantiles with unknown distributions. IS is asymptotically consistent as the number of MC samples, and hence deltas (particles) that parameterize the density estimate, go to infinity. However, retaining infinitely many particles is intractable. We propose a scheme for only keeping a *finite representative subset* of particles and their augmented importance weights that is *nearly consistent*. To do so in an online manner, we approximate importance sampling in two ways. First, we replace the deltas by kernels, yielding kernel density estimates (KDEs). Second, we sequentially project KDEs onto nearby lower-dimensional subspaces. We characterize the asymptotic bias of this scheme as determined by a compression parameter and kernel bandwidth, which yields a tunable tradeoff between consistency and memory. In experiments, we observe a favorable tradeoff between memory and accuracy, providing for the first time near-consistent compressions of arbitrary posterior distributions.

## 1. Introduction

Importance sampling is a MC method that addresses Bayesian inference in cases where the distribution that relates observations to the hidden state is *time-invariant* (Tokdar and Kass, 2010). More specifically, based upon independent samples from a proposal distribution, MC methods approximately compute expectations of arbitrary functions of the unknown parameter via weighted samples generated from the proposal. Recently, use of importance distributions to weight updates, e.g., coordinate descent (Allen-Zhu et al., 2016; Csiba et al., 2015) or stochastic gradient descent (Borsos et al., 2018), have been developed. Doing so yields faster deep network training (Johnson and Guestrin, 2018; Katharopoulos and Fleuret, 2018) by weighting mini-batches (Hanzely and Richtárik, 2018). Furthermore, in reinforcement learning (RL), an agent chooses actions according to a policy and then updates the policy via rewards observed (Watkins and Dayan, 1992); however, this theoretically requires an inordinate amount of random actions to be chosen before reasonable performance is learned (Tsitsiklis, 1994; Sutton et al., 2000), an issue known as the explore-exploit tradeoff. To lessen its deleterious effect, exploratory actions may be chosen via an importance distribution (Schaul et al., 2015) or policy updates may be chosen from previous experience known to be safe (Precup et al., 2000).

**Contributions.** We propose a compression scheme that operates within importance sampling, sequentially deciding which particles are statistically significant for the integral estimation. To do so, we draw connections between proximal methods in optimization (Rockafellar, 1976) and importance distribution updates: we view the empirical measure

[†] U.S. Army Research Laboratory, Adelphi, MD 20783 [*denotes equal contribution].
[†] [†]School of Mathematics, University of Edinburgh (United Kingdom)

defined by importance sampling as carrying out a sequence of projections of un-normalized empirical distributions onto subspaces of growing dimension. Then, we augment the subspace selection by replacing it by one that is nearby (according to some metric) but with lower memory. These lower-memory subspaces are selected based on greedy compression with a fixed budget parameter via matching pursuit (Pati et al., 1993). We combine this idea with kernel smoothing of the empirical measure in order to exploit the fact that compact spaces have finite covering numbers. Consequently, we have characterized the asymptotic bias of this method as a tunable constant depending on the kernel bandwidth parameter and a compression parameter. Experiments demonstrate that this approach yields an effective tradeoff of consistency and memory for MC methods.

## 2. Elements of Importance Sampling

In Bayesian inference (Särkkä, 2013)[Ch. 7], we are interested in computing expectations

$$I(\phi) = \mathbb{E}_{\mathbf{x}}[\phi(\mathbf{x}) \,|\, \mathbf{y}] = \int_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x}) p(\mathbf{x}|\mathbf{y}) d\mathbf{x} \tag{1}$$

on the basis of a set of available observations $\{\mathbf{y}_k\}_{k\leq K}$, where $\phi : \mathbb{R}^p \to \mathbb{R}$ is an arbitrary function, $\mathbf{x}$ is a random variable taking values in $\mathcal{X} \subset \mathbb{R}^p$ which is typically interpreted as the hidden parameter, and $\mathbf{y}$ is some observation process whose realizations $\mathbf{y}_k$ are assumed to be informative about parameter $\mathbf{x}$. For example, $\phi(\mathbf{x}) = \mathbf{x}$ yields the computation of the posterior mean, and $\phi(\mathbf{x}) = \mathbf{x}^p$ denotes the $p$-th moment. In particular, define the posterior distribution

$$p\left(\mathbf{x} \,\middle|\, \{\mathbf{y}_k\}_{k\leq K}\right) = \frac{p\left(\{\mathbf{y}_k\}_{k\leq K} \,\middle|\, \mathbf{x}\right) p\left(\mathbf{x}\right)}{p\left(\{\mathbf{y}_k\}_{k\leq K}\right)}. \tag{2}$$

We seek to infer the posterior (2) with $K$ data points $\{\mathbf{y}_k\}_{k\leq K}$ available at the outset. Even for this setting, estimating (2) has unbounded complexity (Li et al., 2005; Tokdar and Kass, 2010) when the posterior is unknown. Thus, we prioritize efficient estimates of (2) from an online stream of samples from an *importance distribution* to be subsequently defined. Begin by defining posterior $q(\mathbf{x})$ and un-normalized posterior $\tilde{q}(\mathbf{x})$: $q(\mathbf{x}) = \tilde{q}(\mathbf{x})/Z$, and $\tilde{q}(\mathbf{x}) := \tilde{q}(\mathbf{x} \,|\, \mathbf{y}) = p\left(\{\mathbf{y}_k\}_{k\leq K} \,\middle|\, \mathbf{x}\right) p\left(\mathbf{x}\right)$ is a non-negative function proportional to posterior $q(\mathbf{x}|\mathbf{y}) := q(\mathbf{x}) = p\left(\mathbf{x} \,|\, \mathbf{y}\right)$[1] that integrates to normalizing constant $Z := p\left(\{\mathbf{y}_k\}_{k\leq K}\right)$. In Monte Carlo, we approximate (1) by sampling. Hypothetically, we could draw $N$ (not necessarily equal to $K$) samples $\mathbf{x}^{(n)} \sim q(\mathbf{x})$ and estimate the expectation in (1) by the sample average $\mathbb{E}_{q(\mathbf{x})}[\phi(\mathbf{x})] \approx \frac{1}{N}\sum_{n=1}^{N} \phi(\mathbf{x}^{(n)})$, but typically it is difficult to obtain samples $\mathbf{x}^{(n)}$ from posterior $q(\mathbf{x})$ of the hidden state. To circumvent this issue, define the *importance distribution* $\pi(\mathbf{x})$[2] with the same (or larger) support as true density $q(\mathbf{x})$, and multiply and divide by $\pi(\mathbf{x})$ inside the integral (1):

$$\int_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x}) q(\mathbf{x}) d\mathbf{x} = \int_{\mathbf{x}\in\mathcal{X}} \frac{\phi(\mathbf{x}) q(\mathbf{x})}{\pi(\mathbf{x})} \pi(\mathbf{x}) d\mathbf{x}, \tag{3}$$

1. Note that $q(\mathbf{x})$ and $\tilde{q}(\mathbf{x})$ depend on the data $\{\mathbf{y}_k\}_{k\leq K}$, although we drop the dependence to ease notation.
2. In general, the importance distribution could be defined over any observation process $\pi(\mathbf{x} \,|\, \{\mathbf{y}_k\})$, not necessarily associated with time indices $k = 1, \ldots, K$. We define it this way for simplicity.

where the ratio $q(\mathbf{x})/\pi(\mathbf{x})$ is the Radon-Nikodym derivative, or unnormalized density, of the target $q$ with respect to the proposal $\pi$. Then, rather than requiring samples from true posterior $\mathbf{x}^{(n)} \sim q(\mathbf{x})$, one may sample from importance distribution $\mathbf{x}^{(n)} \sim \pi(\mathbf{x})$, $n = 1, ..., N$, and approximate (1) as

$$\widehat{I}_N(\phi) := \frac{1}{N}\sum_{n=1}^{N}\frac{q(\mathbf{x}^{(n)})}{\pi(\mathbf{x}^{(n)})}\phi(\mathbf{x}^{(n)}) = \frac{1}{NZ}\sum_{n=1}^{N} g(\mathbf{x}^n)\phi(\mathbf{x}^{(n)}), \quad \text{where } g(\mathbf{x}^{(n)}) := \frac{\tilde{q}(\mathbf{x}^{(n)})}{\pi(\mathbf{x}^{(n)})} \tag{4}$$

are the importance weights. We note that in practice, we cannot calculate $q(\mathbf{x}^{(n)})$ since the target distribution $q(\mathbf{x})$ is unknown and hence we calculate it using Bayes rule as follows:

$$q(\mathbf{x}^{(n)}) = \frac{p\left(\{\mathbf{y}_k\}_{k\leq K} \,\middle|\, \mathbf{x}^{(n)}\right) p\left(\mathbf{x}^{(n)}\right)}{\int p\left(\{\mathbf{y}_k\}_{k\leq K} \,\middle|\, \mathbf{x}\right) p\left(\mathbf{x}\right) d\mathbf{x}}. \tag{5}$$

Substituting (5) into (3), we obtain $g(\mathbf{x}^{(n)}) := \frac{p\left(\{\mathbf{y}_k\}_{k\leq K} \,\middle|\, \mathbf{x}^{(n)}\right)p\left(\mathbf{x}^{(n)}\right)}{\pi(\mathbf{x}^{(n)})}$. Note that (4) is unbiased, i.e., $\mathbb{E}_{\pi(\mathbf{x})}[I_N(\phi)] = \mathbb{E}_{q(\mathbf{x})}[\phi(\mathbf{x})]$ and consistent with $N$. Moreover, its variance depends on how well the importance density $\pi(\mathbf{x})$ approximates the posterior (Elvira et al., 2019). Example priors and measurement models include Gaussian, Student's t, and Uniform. Which one is appropriate depends on the context (Särkkä, 2013). The normalizing constant $Z$ can be also estimated with IS as $\hat{Z} := \frac{1}{N}\sum_{n=1}^{N} g(\mathbf{x}^{(n)})$. Hence, we can replace $Z$ in Eq. (4) by $\hat{Z}$. Then, the new estimator is given by

$$I_N(\phi) := \frac{1}{N\hat{Z}}\sum_{n=1}^{N} g(\mathbf{x}^{(n)})\phi(\mathbf{x}^{(n)}) = \frac{1}{\sum_{j=1}^{N} g(\mathbf{x}^{(j)})}\sum_{n=1}^{N} g(\mathbf{x}^n)\phi(\mathbf{x}^{(n)}) = \sum_{n=1}^{N} w^n\phi(\mathbf{x}^{(n)}), \tag{6}$$

where the "self-normalized" $\bar{w}^{(n)}$ weights are defined $\bar{w}^{(n)} := \frac{g(\mathbf{x}^{(n)})}{\sum_{u=1}^{n} g(\mathbf{x}^{(u)})}$. The estimator $I_N(\phi)$ is the self-normalized importance sampling (SNIS) estimator. It is important to note that the estimator $I_N(\phi)$ can be viewed as integrating a function $\phi$ with respect to distribution $\mu_N$ defined as $\mu_N(\mathbf{x}) := \sum_{n=1}^{N} \bar{w}^{(n)}\delta_{\mathbf{x}^{(n)}}$, which is called the particle approximation of $q$ where $\delta_{\mathbf{x}^{(n)}}$ denotes the discrete Dirac delta (indicator) which is 1 if $\mathbf{x} = \mathbf{x}^{(n)}$ and null otherwise. This delta expansion is one reason importance sampling is also referred to as histogram filters, as they quantify weighted counts of samples across the space.

As stated in (Agapiou et al., 2017), for consistent estimates of (1), we require that $N$, the number of samples $\mathbf{x}^n$ generated from the importance distribution, and hence the parameterization of the importance distribution, grows unbounded as it accumulates every particle previously generated, as $N \to \infty$. We are interested in allowing $N$, the number of particles, to become large (possibly infinite), while the importance distribution's complexity is moderate, thus overcoming an instance of the curse of dimensionality in Monte Carlo methods. Next, we proposed a compressed kernelized importance sampling algorithm summarized in Algorithm 1.

**Algorithm 1** Compressed Kernelized IS (CKIS)

**Require:** Observation model $p(\mathbf{y} \,|\, \mathbf{x})$ and prior $p(\mathbf{x})$ or target distribution $q(\mathbf{x})$ (if known), importance distribution $\pi(\mathbf{x})$, Observation collection $\{\mathbf{y}_k\}_{k=1}^{K}$

**for** $n = 0, 1, 2, \ldots, N$ **do**

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

Compute the importance weight $g(\mathbf{x}^{(n)}) := \frac{\tilde{q}(\mathbf{x}^{(n)})}{\pi(\mathbf{x}^{(n)})}$

Normalize weights $w^{(n)}$ as follows:

$$\bar{w}^{(j)} := \frac{w^{(j)}}{\sum_{u=1}^{n} w^{(u)}} z_n w^{(j)}, j = 1, ..., n. \,, \text{ where } z_n = \sum_{u=1}^{n} w^{(u)}$$

Update kernel density via last sample & weight

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

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

Compress kernel density estimate sequence as

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

Normalized weights to ensure valid probability measure $\tilde{\mathbf{w}}_n$

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

**end for**

## 3. Balancing Consistency and Memory

In this section, we establish conditions under which the asymptotic bias is proportional to the kernel bandwidth and the compression parameter using posterior distributions given by Algorithm 1. The results permits characterizing the bias of Algorithm 1 given next.

THEOREM 1:
Under Assumptions 1-3 in (Koppel et al., 2019), the estimator of Algorithm 1 exhibits posterior contraction:

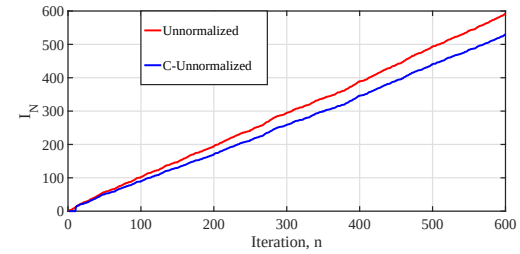
$$\left| \sup_{|\phi| \leq 1} \left( \mathbb{E}[\hat{I}_N(\phi) - I(\phi)] \right) \right| \leq \mathcal{O}\left( \epsilon + \sigma_\kappa^2 h^2 + \frac{1}{\sqrt{Nh}} + \mathcal{O}\left( \frac{1}{\sqrt{N}} \right) + h^3 \right) + \frac{24}{N} \rho, \tag{7}$$

and hence, as $N \to \infty$, is consistent when compression budget and bandwidth go to null $\epsilon, h \to 0$.

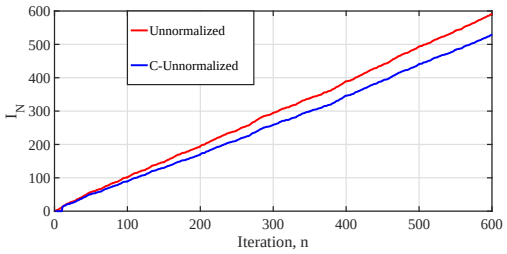
Theorem 1 (proof in (Koppel et al., 2019)) establishes that the compressed kernelized importance sampling scheme proposed in Algorithm 1 is *nearly* asymptotically consistent, and can be made arbitrarily close to exact consistency by sending the bandwidth $h$ and compression budget $\epsilon$ to null. However, when these parameters are fixed positive constants, they provide a tunable tradeoff between bias and memory.

THEOREM 2:
Denote as $\hat{\mu}_n$ the empirical distribution defined by Algorithm 1 whose model order is $M_n$ after $n$ particles generated from importance density $\pi(\mathbf{x})$. Under some Assumptions (detailed in (Koppel et al., 2019)), for compact feature space $\mathcal{X}$ and bounded importance weights $g(\mathbf{x}^{(n)})$, $M_n < \infty$ for all $n$.

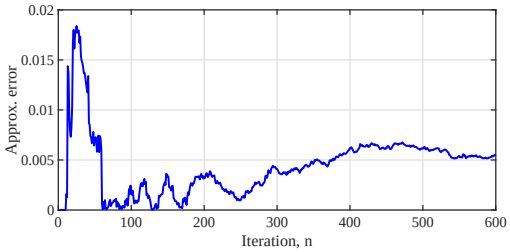
## 4. Experiments



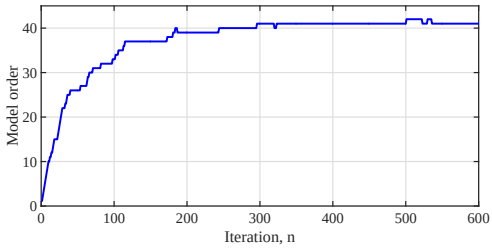

$(a)$ Unnormalized intergral estimate




$(b)$ Unnormalized intergral estimate




$(c)$ Integral Error between Algorithms (with and without compression)




$(d)$ Model Complexity of Algorithm 1

Figure 1: Simulation results for Algorithm 1 run with Gaussian kernel ($h = 0.01$) and compression budget $\epsilon = 3.5$ for the problem (8). The memory-reduction scheme nearly preserves statistical consistency, while yielding reasonable complexity.

In this section, we conduct a simple numerical experiment to demonstrate the efficacy of the proposed algorithm in terms of balancing model parsimony and statistical consistency. We consider the problem of estimating the expected value of function $\phi(\mathbf{x})$ with the target $q(\mathbf{x})$ and the proposal $\pi(\mathbf{x})$ given by

$$\phi(x) = 2\sin\left(\frac{2\pi}{3x}\right) \;,\; q(\mathbf{x}) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-1)^2}{2}\right) \;,\; \pi(\mathbf{x}) = \frac{1}{\sqrt{4\pi}}\exp\left(-\frac{(x-1)^2}{4}\right), \tag{8}$$

to demonstrate that generic Monte Carlo integration allows one to track generic quantities of random variables that are difficult to compute under usual probabilistic hypotheses. Fig. 1a shows the un-normalized integral approximation error for Algorithms with and without compression, which are close, and the magnitude of the difference depends on the choice of compression budget. This trend is corroborated in the evolution of (normalized) integral estimates in Fig. 1b: very little error is incurred by kernel smoothing and memory-reduction. The actual magnitude of the error relative to the number of particles generated is displayed in Fig. 1c: observe that the error settles on the order of $10^{-3}$. In Fig. 1d, we display the number of particles retained by Algorithm 1, which stabilizes to around 56, whereas the complexity of the empirical measure without compression grows linearly with sample index $n$, which noticeably grows *unbounded*.

## References


S Agapiou, Omiros Papaspiliopoulos, D Sanz-Alonso, AM Stuart, et al. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, 32(3):405–431, 2017.

Zeyuan Allen-Zhu, Zheng Qu, Peter Richtárik, and Yang Yuan. Even faster accelerated coordinate descent using non-uniform sampling. In *International Conference on Machine Learning*, pages 1110–1119, 2016.

Zalán Borsos, Andreas Krause, and Kfir Y Levy. Online variance reduction for stochastic optimization. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pages 324–357. PMLR, 2018.

Dominik Csiba, Zheng Qu, and Peter Richtárik. Stochastic dual coordinate ascent with adaptive probabilities. In *International Conference on Machine Learning*, pages 674–683, 2015.

Vctor Elvira, Luca Martino, David Luengo, and Mnica F. Bugallo. Generalized multiple importance sampling. *Statist. Sci.*, 34(1):129–155, 02 2019. doi: 10.1214/18-STS668. URL https://doi.org/10.1214/18-STS668.

Filip Hanzely and Peter Richtárik. Accelerated coordinate descent with arbitrary sampling and best rates for minibatches. *arXiv preprint arXiv:1809.09354*, 2018.

Tyler B Johnson and Carlos Guestrin. Training deep models faster with robust, approximate importance sampling. In *Advances in Neural Information Processing Systems*, pages 7265–7275, 2018.

Angelos Katharopoulos and Francois Fleuret. Not all samples are created equal: Deep learning with importance sampling. In *International Conference on Machine Learning*, pages 2530–2539, 2018.

Alec Koppel, Amrit Singh Bedi, Victor Elvira, and Brian M Sadler. Approximate shannon sampling in importance sampling: Nearly consistent finite particle estimates. *arXiv preprint arXiv:1909.10279*, 2019.

Bo Li, Thomas Bengtsson, and Peter Bickel. Curse-of-dimensionality revisited: Collapse of importance sampling in very large scale systems. *Rapport technique*, 85, 2005.

Y. Pati, R. Rezaiifar, and P.S. Krishnaprasad. Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition. In *Asilomar Conference*, 1993.

Doina Precup, Richard S Sutton, and Satinder P Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 759–766, 2000.

R. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. Control Opt.*, 14(5):877–898, 1976. doi: 10.1137/0314056. URL https://doi.org/10.1137/0314056.

Simo Särkkä. *Bayesian filtering and smoothing*, volume 3. Cambridge, 2013.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

Surya T Tokdar and Robert E Kass. Importance sampling: a review. *Wiley Interdisciplinary Reviews: Comp. Stat.*, 2(1):54–60, 2010.

John N Tsitsiklis. Asynchronous stochastic approximation and q-learning. *Machine learning*, 16(3):185–202, 1994.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.