Late-Breaking Developments in the Field of Artificial Intelligence
Papers Presented at the Twenty-Seventh AAAI Conference on Artificial Intelligence

Learning When to Reject an Importance Sample

Jeremy C. Weiss
University of Wisconsin-Madison
jcweiss@cs.wisc.edu

Sriraam Natarajan
Wake Forest University
snataraj@wakehealth.edu

C. David Page
University of Wisconsin-Madison
page@biostat.wisc.edu

Abstract

When observations are incomplete or data are missing, approximate inference methods based on importance sampling are often used. Unfortunately, when the target and proposal distributions are dissimilar, the sampling procedure leads to biased estimates or requires a prohibitive number of samples. Our method approximates a multivariate target distribution by sampling from an existing, sequential importance sampler and accepting or rejecting the proposals. We develop the rejection-sampler framework and show we can learn the acceptance probabilities from local samples. In a continuous-time domain, we show our method improves upon previous importance samplers by transforming a sequential importance sampling problem into a machine learning one.

Introduction

Importance sampling is a method of generating samples, by drawing from a surrogate distribution and weighting them, to approximate sampling from a target distribution. It provides the basis for many distribution approximations, including those used in particle filters and temporal models (Doucet, Godsill, and Andrieu 2000; Fan, Xu, and Shelton 2010) with applications for example in robotic environment mapping and speech recognition (Montemerlo et al. 2003; Wolfel and Faubel 2007). The characteristic shortcoming of importance sampling is that if the ratio between the probability of generating a sample under the target and surrogate distributions is large, many samples from the surrogate distribution will be required to obtain a close approximation to the target distribution. As large ratios result in large variance in weights, they result in slow convergence of the approximation to the target distribution and slow convergence of the sample bias to 0.

We focus on identifying an improved surrogate distribution for sequential importance sampling. This is the main idea behind the field of adaptive importance sampling, see *e.g.*, (Cornebise, Moulines, and Olsson 2008; Yuan and Druzdzel 2003; 2007), where domain-specific knowledge is used to design and adapt the surrogate distribution.

In this paper, we propose a general-purpose rejection-sampling technique. Given a target distribution f and a surrogate distribution g, we propose a second surrogate distribution h based on learning the acceptance probability a of a rejection sampler. That is, we sample a transition from g and decide to accept or resample. We show we can learn a binary classifier ϕ to effectively make the decision.

We apply our analysis to approximate a target distribution over continuous-time sequences given a continuous-time Bayesian network (CTBN) and evidence (Nodelman, Shelton, and Koller 2002). CTBNs model timelines using a set of discrete random variables, with applications for example in medicine (Weiss, Natarajan, and Page 2012). The CTBN importance sampler g (Fan, Xu, and Shelton 2010) uses a combination of exponential and truncated exponential distributions to select interval transitions that agree with evidence. Using g, each evidence point causes a stochastic downweighting in a fraction of the samples, resulting in an increase in variance of the importance weights. Because a sequence weight corresponds to the product of its interval weights, the stochastic downweighting of each interval approaching nonmatching evidence produces a distribution of sequence weights with potentially high variance. High variance weights indicate that the surrogate distribution g is far from the target distribution f and practically leads to either biased results or requires a prohibitive number of samples.

While techniques such as sequential Monte Carlo (SMC), *i.e.*, particle filtering, can mitigate some of this effect, they can lead to particle degeneracy, especially when many resampling iterations are required (Doucet, Godsill, and Andrieu 2000; Fan, Xu, and Shelton 2010). Particle smoothing combats particle degeneracy, but the exponentially large state spaces used in CTBNs limit its ability to find alternative, probable sample histories. Our work proposes a new method for improving existing importance sampling schemes and can similarly be used in the SMC framework.

Rejection-based importance sampling

Let $f(z)$ be the target distribution we wish to approximate for $z \in Z$ in domain $\mathcal{Z}$, $Z \subseteq \mathcal{Z}$. Let $g(z)$ be a surrogate distribution from which we can sample such that if $f(z) > 0$ then $g(z) > 0$. Then, we can approximate f with weighted samples z_i from g:

$$\int_{z \in Z} f(z)dz = \int_{Z} g(z)\frac{f(z)}{g(z)}dz \approx \frac{1}{n}\sum_{i=1}^{n}\frac{f(z_i)}{g(z_i)} = \frac{1}{n}\sum_{i=1}^{n} w_i$$

with $z_i \in Z$. We design a second surrogate $h(z)$ with the density corresponding to accepting a transition from g:

$$h(z) = g(z)a(z)\frac{1}{1 - \int_{\mathcal{Z}}(1 - a(\zeta))g(\zeta)d\zeta}$$

where $a(z) = \min(1, f(z)(1-r)/g(z))$ is the acceptance probability of the sample from g, and r is a constant in $[0, 1]$ (the familiar rejection sampler "envelope" is $g(z)/(1-r)$). In other words, we approximate f with h by (re-)sampling from g and accepting with probability a. This procedure again approximates f, with weights $w_i = w_{i;gh}w_{i;fg}$:

$$\int_Z f(z)dz = \int_Z h(z)\frac{g(z)}{h(z)}\frac{f(z)}{g(z)}dz \approx \frac{1}{n}\sum_{i=1}^{n} w_{i;gh}w_{i;fg}$$

The quality of the importance sampler is best when all sample weights are equal. The effective sample size (ESS) is an indicator of the quality of the samples (larger is better): $\text{ESS} = 1/(\sum_{i=1}^{n} W_i^2)$, where $W_i = w_i / \sum_{j=1}^{n} w_j$.

Unfortunately, the calculation of the target distribution is often intractable or impossible and thus we cannot directly recover $a(z)$. Sequential importance sampling (SIS) is used when estimating the distribution f over a sequence z of elements $z^1, \ldots, z^k$, typically given some evidence e. In time-series models, z^i is a time step; in continuous-time models, z^i is an interval. Our key idea is that we can calculate the target, element density $f(z^i = z'|e)$:

$$f(z'|e) = g(z'|e)\frac{E_g[w|\mathbb{1}_a(z'), e]}{E_g[w|e]} \quad (1)$$

where e denotes evidence, $\mathbb{1}_a(z')$ denotes acceptance of z', and w is the weight of the sample completion under g. We omit the derivation due to limited space.

The approximation of Equation 1 corresponds to sampling many sequence completions given $z^i = z'$ to recover $E_g[w|\mathbb{1}_a(z'), e]$ and z^i unconditioned to recover $E_g[w|e]$. While possible, this is inefficient because (1) it requires weight estimations for every z' of every z^i, and (2) the approximation of the expected weights relies on importance sampling, which is intractable in the first place.

However, $E_g[w|\mathbb{1}_a(z'), e]/E_g[w|e]$ is simply the expected weight ratio given acceptance and rejection of z'. We recognize that similar situations, in terms of state, evidence, model and proposal, result in similar distributions of a and h. Thus, we can learn a classifier $\phi(z'|e)$ for {acceptance, rejection} as a function of the situation.

Subject to choice of r, with $\alpha = 1/(1-r)$, such that $f/\alpha g \leq 1$ for all z', the optimal acceptance $a(z'|e)$ is:

$$a(z'|e) = \frac{f(z'|e)}{\alpha g(z'|e)} = \frac{E[w|\mathbb{1}_a(z'), e]}{\alpha E[w|e]} \approx \frac{\phi(z'|e)}{\alpha(1 - \phi(z'|e))}$$

With this estimate $a_\phi(z'|e)$, we obtain the proposal $h_\phi(z'|e)$.

Experiments

We compare our rejection method (setting $r = 0.5$) with the original CTBN sampler (Fan, Xu, and Shelton 2010). For ease of implementation, we learn a logistic regression (LR) model using online, stochastic gradient descent for each CTBN state. An LR data example is $(y, \mathbf{x}, w)$: $y = \{\text{accept, reject}\}$, $\mathbf{x}$, a set of indicator features encoding each state and features for the time to next (nonmatching) evidences, mapped to intervals $[0, 1]$, and w, an importance weight. The examples are generated using element samples from g and rejecting r percent of the time. Because of the high variance in weights for full sequences, we instead choose a local weight: the weight of the sequence through the next m evidence times. Note that this biases the learner to have low variance weights within the local window, but it does not bias the proposal h_ϕ. We test our model on the strong-cycle model(n) for $n = \{1, 2, 3\}$ (models with a single, $2n$-state cycle with rates of 0.1, and rates of 0.01 otherwise) and the drug model developed in the original CTBN paper (Nodelman, Shelton, and Koller 2002).

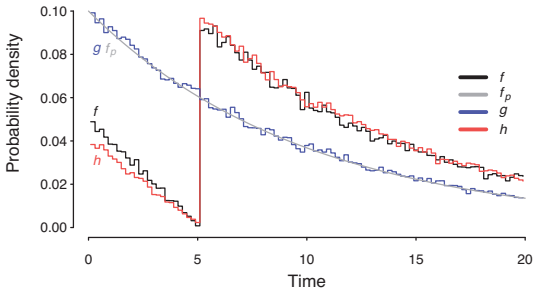

Figure 1: Approximate transition densities of f (target), f_p (target without evidence), g (surrogate), and h (learned rejection surrogate) in a one-variable, binary-state CTBN with transition rates of 0.1 and matching evidence at t=5. Our learned density h closely mimics f, the target density with evidence, while g mimics f_p (exactly, in this situation).

Figure 1 illustrates the ability of h to mimic f, the target distribution, in a one-node binary-state CTBN with matching evidence at t=5. The Fan et al. proposal g is chosen to match the target density in the absence of evidence f_p. However, when approaching evidence (at t=5), the probability of a transition given evidence goes to 0 as the next transition must also occur before t=5 to be a viable sequence. Only f and h exhibit this behavior. Table 1 shows that the learned, rejection-based proposal h outperforms g across all 4 models, resulting in an ESS an average of 2 to 10 times larger.

In sum, we present the framework of rejection learning to improve importance sampling and show that our instantiation–logistic regression models applied to CTBNs–improves the existing CTBN proposal distribution.

Model	Fan et al. (g)	Rejection learner (h)
Strong cycle, n=1	690	6400
Strong cycle, n=2	19000	35000
Strong cycle, n=3	960	5800
Drug	29	170

Table 1: Geometric mean of effective sample size (ESS) over 100 sequences, each with 100 observations; ESS is per 100k samples. The proposal h was learned with 1000 sequences.

References

Cornebise, J.; Moulines, E.; and Olsson, J. 2008. Adaptive methods for sequential importance sampling with application to state space models. *Statistics and Computing* 18(4):461–480.

Doucet, A.; Godsill, S.; and Andrieu, C. 2000. On sequential monte carlo sampling methods for Bayesian filtering. *Statistics and computing* 10(3):197–208.

Fan, Y.; Xu, J.; and Shelton, C. R. 2010. Importance sampling for continuous time Bayesian networks. *The Journal of Machine Learning Research* 11:2115–2140.

Montemerlo, M.; Thrun, S.; Koller, D.; and Wegbreit, B. 2003. Fastslam 2.0: An improved particle filtering algorithm for simultaneous localization and mapping that provably converges. In *International Joint Conference on Artificial Intelligence*, volume 18, 1151–1156. Lawrence Erlbaum Associates LTD.

Nodelman, U.; Shelton, C.; and Koller, D. 2002. Continuous time Bayesian networks. In *Uncertainty in artificial intelligence*, 378–387. Morgan Kaufmann Publishers Inc.

Weiss, J.; Natarajan, S.; and Page, D. 2012. Multiplicative forests for continuous-time processes. In *Advances in Neural Information Processing Systems 25*, 467–475.

Wolfel, M., and Faubel, F. 2007. Considering uncertainty by particle filter enhanced speech features in large vocabulary continuous speech recognition. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, IV–1049. IEEE.

Yuan, C., and Druzdzel, M. J. 2003. An importance sampling algorithm based on evidence pre-propagation. In *Uncertainty in Artificial Intelligence*, 624–631. Morgan Kaufmann Publishers Inc.

Yuan, C., and Druzdzel, M. J. 2007. Importance sampling for general hybrid Bayesian networks. In *Artificial intelligence and statistics*.