Parametric Adaptive Importance Sampling based Approach for Learning the Shape of Target Function

Abstract

This article discusses how to construct an Importance Sampling Function (ISF) \( p(x) \) to raise the efficiency of Monte Carlo estimators. In order to achieve the subject, it is needed to approximate a certain target distribution, which is known only up to the normalizing constant, by ISF. Adaptive Importance Sampling (AIS) is an algorithm for learning its shape recursively using current information. Here, information corresponds to what is represented by a set of importance weights. Our proposal for the learning is Parametric AIS (PAIS) on the assumption that ISF belongs to a family of finite mixture distribution as an alternative to Non-parametric AIS (NAIS). PAIS algorithm explores the member to be the minimum Kullback-Leibler divergence from the target distribution. It can be implemented with less intensive computation and also improves the quality of estimates significantly. Simulation studies show that the estimators obtained by PAIS are superior in the efficiency and computing time to the estimators obtained by NAIS.

Key Words: Importance Sampling, Finite mixture distribution, EM algorithm, Bayes estimation, Kullback-Leibler divergence.

1 INTRODUCTION

Consider the problem of finding a generic \( d \)-dimensional integral of the form

\[
     I = \int_{\mathcal{F}} \psi(x) f(x) dx,
\]

where \( f : \mathcal{R}^d \to \mathcal{R} \) denotes a probability density function on \( \mathcal{F} \in \mathcal{R}^d \) and \( \psi : \mathcal{R}^d \to \mathcal{R} \) is a certain integrable function. In many areas, for instance, Bayesian statistics, statistical physics and engineering literature, intractable integrands arise from model complexity. This is a more significant problem in high dimensional models.

The latest computational developments enable easy estimation via simulation based methods that realize the approximation of \( f(x) \) using an empirical distribution of a sample set from a certain distribution. Such of methods, i.e. conventional Monte Carlo, Markov Chain Monte Carlo, enable flexible modeling in a broader range of applications, i.e.
Bayesian statistics (Kass and Raftery 1995; Liu 2001; Pauler, Wakefield and Kass; 1999),
time series analysis (Higuchi 1997; Kitagawa 1996; Pitt and Shephard 1999; Stavropoulos
and Titterington 2002), engineering (Srinivasan 2002). Further progress of computational
capabilities will accelerate applicability to other fields.

The method of importance sampling is a variance reduction technique that improves
the efficiency of conventional Monte Carlo estimates (Gelman, Carlin, Stern and Rubin
2000; Liu 2001). This method focuses on the region of importance to save computational
resources. It proceeds by

**Step1.** Draw a iid sample set \( \{X_i\}_{i=1}^n \) from Importance Sampling Function (ISF) (or envelope) \( p(x) \),

**Step2.** Calculate the importance weights given by \( \omega(X_i) = \psi(X_i) f(X_i)/p(X_i) \) \( (i = 1, \ldots, n) \),

**Step3.** Estimate \( I \) by \( \hat{I}_{is} = \frac{1}{n} \sum_{i=1}^n \omega(X_i) / n \).

Obviously, the quantity \( \hat{I}_{is} \) is an unbiased estimator of \( I \) and the mean square error is
expressed by \( V(\hat{I}_{is}; p) = var_p(\omega(x))/n \). Thus, an appropriate choice of envelope \( p(x) \) leads
to an immense efficiency gain of Monte Carlo estimation with importance sampling.

Under certain conditions, there exists an optimal envelope, denoted by \( g^*(x) \), that
minimizes \( var_p(\omega(x)) \) with respect to \( p(x) \). It follows from applying Jensen’s inequality
to the expectation in \( var_p(\omega(x)) \) that (Zhang (1996), Srinivasan (2002))

\[
g^*(x) := |\psi(x)| f(x)/E_f(|\psi(x)|).
\] (2)

Substituting \( p(x) = g^*(x) \) in \( V(\hat{I}_{is}; p) \) yields the lower boundary of estimation error as,

\[
V(\hat{I}_{is}; g^*) = \frac{1}{n} \left[ E_f(|\psi(x)|)^2 - I^2 \right].
\]

In this study, we refer to \( g^*(x) \) as optimal distribution, or target distribution. As a special
case, if \( \psi(x) \) has one sign in the region \( F \), it holds that \( V(\hat{I}_{is}; g^*) = 0 \). That is, the optimal
choice of ISF gives a perfect resolution of \( I \). Unfortunately, direct sampling from \( g^*(x) \) is
usually impossible since it includes the unknown quantity \( E_f(|\psi(x)|) \). Our study focuses
on constructing a reasonable approximation of target distribution \( g^*(x) \propto g(x) \), which is
known only up to the normalizing constant. Here, the known part \( |\psi(x)| f(x) \) corresponds
to \( g(x) \). A great deal of research has investigated this problem.

Adaptive Importance Sampling (AIS) is an algorithm for learning the shape of \( g^*(x) \)
through a weighted sample set \( \{\omega(X_i), X_i\}_{i=1}^n \) where \( \omega(X_i) \) is importance weight. The
method proceeds by using the obtained information \( \{\omega(X_i), X_i\}_{i=1}^n \) to refine the next ISF
via a specific learning method. If the refinement is well implemented, a weighted sample
set from new ISF would include a lot of information on the shape of the target distribution.
Thus, additional learning under the informative weighted sample set, currently obtained,
yields more sophisticated ISF.

A well-designed learning method is the most important factor in controlling the performance
of the AIS algorithm. Many researchers have investigated non-parametric methods of estimating
the target distribution (West 1993; Givens and Raftery 1996; Zhang 1996;
Stavropoulos and Titterington 2001). However, as will be shown in a later section, the estimators produced by these methods give rise to intractable computation and even poor quality estimations. An alternative is the parametric approach on the assumption that ISF belongs to a family of parametric distribution \( \mathcal{M} = \{p(x; \theta), \theta \in \Theta\} \) where \( \theta \) denotes the parameter vector defined over set \( \Theta \). The learning algorithm explores a member to minimize the arbitrary cost function, which measures the distance from the target distribution.

Oh and Berger (1993) proposed specifying the family of finite mixture distribution with multivariate \( t \) components as an envelope. While the difficulty of the parametric approach is to specify a suitable family in which to include \( g^*(x) \) as a member, such specification makes it possible to comprise the broader class of distributions, at least the good approximation of \( g^*(x) \), by increasing the number of components. In fact, as is shown in McLachlan and Peel (2000), it can represent various shapes of distributions despite of fewer components. However, in their algorithm, a computational difficulty arises from the procedure for minimizing the sample estimate of mean square error \( V(\hat{L}; \theta) \) with respect to its parameters.

Our study also follows a parametric approach under the assumption of finite mixture distribution with various components to ISF. The learning algorithm detects the member to be the minimum Kullback-Leibler divergence from the target distribution. It is implemented with less intensive computation and significantly improves the quality of an estimate. Our simulation studies show that the estimators produced by PAIS are more efficient and require less computing time than estimators given by NAIS.

The article is organized as follows. Section 2 presents a general framework of the AIS algorithm and Non-parametric AIS (NAIS) as a special case of it. Section 3 describes the PAIS algorithm with finite mixture distribution. It is constructed by adopting the principle of minimum Kullback-Leibler divergence to target. Particularity we discuss Gaussian component cases as an illustration and some related topics. Section 4 presents a simulation study to compare computing times and the precision of obtained estimators between PAIS and NAIS. Section 5 describes the dimensional effect of the PAIS algorithm. Application to a high-dimensional target is limited due to well known phenomena, the curse of dimensionality. We test the performance of PAIS and will present a benchmark test for applying PAIS to different problems.

2 GENERAL FRAMEWORK OF ADAPTIVE IMPORTANCE SAMPLING

AIS is a special sampling scheme for learning about the shape of target distribution \( g^*(x) \propto g(x) \) defined over the subspace of \( \mathcal{R}^d \). The method proceeds by using a weighted sample set \( \{\omega(X^{[l-1]}_i), X^{[l-1]}_i\}_{i=1}^{n_{l-1}} \) from \( p^{(l-1)}(x) \) which is obtained in the previous stage \( l - 1 \) to refine the next envelope \( p^{(l)}(x) \). Moreover, a sample set \( \{\omega(X^{[l]}_i), X^{[l]}_i\}_{i=1}^{n_l} \) from the current \( p^{(l)}(x) \) also serves as further improvement of the sampling function. These
steps are repeated if the current envelope is still poor.

![Schematic expression of AIS](image)

Figure 1: Schematic expression of AIS

The general AIS procedure is as follows:

**AIS algorithm:**

**Step 0.** Let the initial envelop be \( p^0(x) \) at step \( l = 0 \) and draw a weighted sample set \( \{\omega(X_i^{(0)}), X_i^{(0)}\}_{i=1}^{n_0} \) where \( \omega(x) = \psi(x)f(x)/p^0(x) \) and \( n_0 \) denotes the sample size.

**Step 1.** (Learning Stage) Increase the step index, \( l = l + 1 \), and update the envelope \( p^l(x) \) by a certain density estimation technique for \( g^*(x) \) under the current information \( \{\omega(X_i^{(l-1)}), X_i^{(l-1)}\}_{i=1}^{n_{l-1}} \).

**Step 2.** (Sampling Stage) Generate a sample set \( \{\omega(X_i^{(l)}), X_i^{(l)}\}_{i=1}^{n_l} \) from \( p^l(x) \) and implement importance sampling estimate by using the set if test statistics \( T\left(\{\omega(X_i^{(l)}), X_i^{(l)}\}_{i=1}^{n_l}\right) \), which measures the quality of obtained envelope, exceeds a threshold.

**Step 3.** go to step 1.

Step 0 generates an initial reference sample for the density estimation routine at stage \( l = 1 \). AIS is composed of a learning stage and a sampling stage as shown in Figure 1. Step 1 corresponds to the learning stage and updates ISF by using a previously obtained weighted sample set. Step 2 implements importance sampling estimation using \( p^l(x) \) if the obtained envelope is a good approximation of the target. Here, \( T\left(\{\omega(X_i^{(l)}), X_i^{(l)}\}_{i=1}^{n_l}\right) \) is a test statistics for assessing the quality of weighted samples. The degree of efficiency the method achieves depends upon the selected density estimation algorithm in the learning stage.

Non-parametric AIS (NAIS) algorithm is what the weighted kernel density estimator is interpolated in the learning stage. Let \( \tilde{\omega}(X_i^{(l-1)}) = |\omega(X_i^{(l-1)})|/\sum_{i=1}^{n_{l-1}} |\omega(X_i^{(l-1)})| \),
In this article, we refer to it as normalizing absolute weight. The refinement of ISF is implemented by
\[
p^{(l)}(x) = \sum_{i=1}^{n_{t-1}} \omega(X_i^{[l-1]}) K(x; X_i^{[l-1]}, h_i^{[l-1]} V_i^{[l-1]}),
\]
where \(K(x|a, B)\) denotes an elliptically symmetric kernel defined over \(\mathcal{R}^d\) centered at mode \(a\) and scaled by some positive definite matrix \(B\), and \(h_i^{[l-1]}\) is a smoothing parameter. Givens and Raftrey (1996) named this method Local Adaptive Importance Sampling (LAIS), which is a generalization of the study by West (1993). Their algorithm particularly takes the form as \(V_i^{[l-1]} = \Sigma_i^{[l-1]}(\lambda)\) which denotes the covariance estimate for the \(\lambda\)-nearest neighbors of \(i\) th sample and the nearness is measured by Mahalanobis distance. Global AIS (GAIS) proposed by West (1993) is a special case of this as \(\lambda = n_{t-1}\). Moreover, Zhang (1996) studied the asymptotic property of the estimator obtained by NAIS. Section 4 compares the estimator accuracy and required computing time of our proposed algorithm and GAIS.

3 PARAMETRIC ADAPTIVE IMPORTANCE SAMPLING

3.1 Design of Learning Stage

In the previous section, the optimal envelope for importance sampling is defined by \(g^*(x)\). Our strategy for obtaining \(p(x)\) to approximate the target is to specify a parametric distribution family to the envelope. Here, consider a family of \(\beta\)-component finite mixture distribution in which each member takes the form,
\[
p(x; \theta) = \sum_{k=1}^{\beta} \psi_k f_k(x; \eta_k),
\]
where \(\psi_k\) is the mixing proportion, \(f_k(x; \eta_k)\) denotes \(k\)th component density with the parameter vector \(\eta_k\). The vector \(\theta\) contains all unknown parameters, namely \(\theta = (\psi_1, \cdots, \psi_{\beta-1}, \eta^T)\). The mixing rates obey the restrictions such that \(\psi_k \geq 0\) for \(k = 1, \cdots, \beta\) and \(\sum_{k=1}^{\beta-1} \psi_k = 1 - \sum_{k=1}^{\beta-1} \psi_k\).

Let \(\mathcal{M} = \{p(x; \theta), \theta \in \Theta\}\) be the family of envelope. The objective is to get optimal one \(p(x; \theta^*)\) to be the minimum distance from \(g^*(x)\) among \(\mathcal{M}\) with respect to a certain distance measure. In this article, we consider Kullback-Leibler divergence as a distance and it is represented by
\[
\mathcal{D}(g^*; p) = \int_G g^*(x) \log \left( \frac{g^*(x)}{p(x; \theta)} \right) dx = C - \mathcal{L}(\theta),
\]
where $C = \int g^*(x) \log g^*(x) \, dx$ and $\mathcal{L}(\theta) = \int g^*(x) \log p(x; \theta) \, dx$. Support of $g^*(x)$, $\mathcal{G}$ must be consistent with the support of $p(x; \theta)$ in order to assure this criteria to be a distance. A minimization of $\mathcal{D}(g^*; p)$ with respect to $\theta$ is equivalent to maximize $\mathcal{L}(\theta)$ and it holds that $\mathcal{D}(g^*; p) = 0$ if and only if $g^*(x) = p(x; \theta)$ for any $x$.

In our PAIS algorithm, the term $\mathcal{L}(\theta)$ is estimated by Monte Carlo estimation using the sample set $\{\omega(X_i), X_i\}_{i=1}^n$ which is generated from the previous stage envelope $p^{l-1}(x)$. One can be written by

$$\hat{\mathcal{L}}(\theta) = \sum_{i=1}^n \hat{\omega}(X_i) \log p(X_i; \theta), \quad (3)$$

where, again, $\hat{\omega}(X_i)$ is the normalizing absolute weight, that is, $|\omega(X_i)|/\sum_{i=1}^n |\omega(X_i)|$ and $\omega(X_i) = \psi(x) f(x)/\bar{p}^{l-1}(x)$. For notational convenience, the subscript for the step index is omitted.

By applying idea of Liu (2001), we can confirm that the estimator (3) is consistent to $\mathcal{L}(\theta)$ as $n$ goes to infinity with a fixed $\theta$ and the mean square error is approximated by

$$V(\hat{\mathcal{L}}; p^{l-1}) \approx \frac{1}{n} \left\{ \text{var}_{p^{l-1}}[\omega(x) \log p(x; \theta)] + \mathcal{L}(\theta)^2 \text{var}_{p^{l-1}}[\omega(x)] \right\} / \left[ \int g^*(x) \, dx \right]^2.$$ 

Clearly, equation (3) is a weighted log-likelihood function for finite mixture distribution and optimization is performed by applying EM algorithm (Dempster, Laird and Rubin 1977; Titterington, Smith and Markov 1985; McLachlan and Peel 2000). By letting $Z_{ik}$ be unknown class label to take 1 if $X_i$ is generated from $k$th component, otherwise 0, the complete data weighted log-likelihood takes the following form:

$$\hat{\mathcal{L}}_c(\theta) = \sum_{i=1}^n \sum_{k=1}^\beta \hat{\omega}(X_i) Z_{ik} \{ \log f_k(X_i; \eta_k) + \log \psi_k \}.$$ 

Denoting the iteration index by $j$ and the current guess on parameter by $\theta^{(j-1)}$, the membership probability that sample $X_i$ belongs to $k$th component is assessed by,

$$\lambda(k|i; \theta^{(j-1)}) = \frac{\psi_k^{(j-1)} f_k(X_i; \eta_k^{(j-1)})}{p(X_i; \theta^{(j-1)})},$$

for $i = 1, \ldots, n$ and $k = 1, \ldots, \beta$. Taking expectation of $\hat{\mathcal{L}}_c(\theta)$ with respect to this membership probability gives

$$E_{\theta^{(j-1)}} \{ \mathcal{L}_c(\theta) \} = \sum_{i=1}^n \sum_{k=1}^\beta \hat{\omega}(X_i) \lambda(k|i; \theta^{(j-1)}) \{ \log f_k(X_i; \eta) + \log \psi_k \}. \quad (4)$$

Updating estimates $\theta^{(j)}$ in order to maximize (4) with respect to $\theta$ yields a non-decreasing sequence of $\hat{\mathcal{L}}(\theta^{(j)})$ and thus proceeding to this recursion produces the local maxima of (3) under certain conditions. Updated estimates of the mixing proportions are calculated independently of component parameters $\eta$ by

$$\psi_k^{(j)} = \frac{n}{\sum_{i=1}^n \hat{\omega}(X_i) \lambda(k|i; \theta^{(j-1)})}, \quad (5)$$

6
Learning Stage of PAIS:

Step L-0. Set a family of envelopes as \( \mathcal{M}_\beta \), initial parameter \( \theta^{(0)} \) and step index as \( j = 0 \).

Step L-1. Increase the step index as \( j = j + 1 \) and calculate the membership probability \( \lambda(k|i; \theta^{(j-1)}) \) and the convoluting weights by,
\[
\xi(k|i; \theta^{(j-1)}) = \tilde{\omega}(X_i)\lambda(k|i; \theta^{(j-1)})
\]
for \( i = 1, \ldots, n \) and \( k = 1, \ldots, \beta \).

Step L-2. Update estimate \( \theta^{(j)} \) by maximizing \( E_{\theta^{(j-1)}}[\mathcal{L}_e(\theta)] \) with respect to \( \theta \).

Step L-3. If the sequence of \( \theta^{(j)} \) converges or refinement of \( \mathcal{L}(\theta^{(j)}) \) is stopped, then return the envelope \( p^{(j)}(x) = p(x; \theta^{(j)}) \) to Step 3 (1st Sampling Stage) of AIS, otherwise go to Step L-1.

As is noted by McLachlan and Peel (2000), some specifications of component distribution, for example, exponential family, make it trivial to implement the maximization of expected complete log-likelihood (4) in Step L-2. It is possible to choose various models for problems.

3.2 Gaussian component distribution

In this article, we focus on a finite mixture with normal component, which takes the form
\[
f_k(x; \mu_k, \Sigma_k) = (2\pi)^{-d/2} |\Sigma_k|^{-1/2} \exp\left\{ -\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \right\},
\]
where \( \mu_k, \Sigma_k \) correspond to mean vector and covariance matrices, respectively. In this case, the maximization of \( E_{\theta^{(j-1)}}[\mathcal{L}_e(\theta)] \) at Step L-2 is achieved by
\[
\mu_k^{(j)} = \frac{\sum_{i=1}^{n} \xi(k|i; \theta^{(j-1)})X_i}{\sum_{i=1}^{n} \xi(k|i; \theta^{(j-1)})},
\]
\[
\Sigma_k^{(j)} = \frac{\sum_{i=1}^{n} \xi(k|i; \theta^{(j-1)})(X_i - \mu_k^{(j)})(X_i - \mu_k^{(j)})^T}{\sum_{i=1}^{n} \xi(k|i; \theta^{(j-1)})},
\]
and (5). Such form of ISF has extensive fields of application. However, there is a numerical problem inherent to Gaussian mixture distribution. As noted in studies on the maximum likelihood estimation of finite mixtures (Titterington, Smith and Markov 1985; McLachlan and Peel 2000), there exists multiple roots corresponding to local maxima in likelihood equations. In general, the number of roots increases as the number of components increases and the sample size becomes smaller. This phenomena becomes more serious in the case of volatile weighting for log-likelihood function in (3). Let \( \theta^* \) be the one which meets \( \text{ming}_D(\theta^*; p) \). Under some regularity condition, the asymptotic variance of our density estimator is expressed by
\[
V_\alpha(\hat{\theta}) = I^{-1}J(\tilde{\omega})I^{-1},
\]
(6)
where,

\[
J(\tilde{\omega}) = E_q \{ \tilde{\omega}(x) \partial_\theta \log p(x; \theta^*) \partial_\theta \log p(x; \theta^*) \},
\]

\[
I = -E_q \{ \partial^2_\theta \log p(x; \theta^*) \},
\]

\(\tilde{\omega} = g^*(x)/p^{(l-1)}(x)\). The asymptotic error of \(V^*_a(\theta) = I^{-1}J(1)I^{-1}\) is obtained by using the optimal envelope in the previous stage, i.e. \(g^*(x) = p^{(l-1)}(x)\). Since \(V_a(\theta) - V^*_a(\theta)\) is positive definite, then our weighted likelihood estimator is less efficient than the ordinal maximum likelihood estimator. As a result, we have to be more careful in selecting local maximizers. A practical way to deal with this problem is to restrict the family of envelopes to ones with homoscedastic covariance matrices, \(\Sigma = \Sigma_k \ (k = 1, \ldots, \beta)\). In this case, the maximum weighted log-likelihood estimator exists as the global maximizer of (3) and the updated estimate of \(\Sigma\) is given by

\[
\Sigma^{(j)} = \sum_{i=1}^n \sum_{k=1}^\beta \xi(k|i; \theta^{(j-1)}) (X_i - \mu_k^{(j)}) (X_i - \mu_k^{(j)})^T / \sum_{i=1}^n \sum_{k=1}^\beta \xi(k|i; \theta^{(j-1)}).
\]

For Gaussian mixtures with heteroscedastic covariance matrices, we have to select a local maximizer in order to exclude the singular components that are highly spiked at a sample point. Since successful application of an estimate with importance sampling requires that the envelope \(p(x)\) has a longer tail than \(g^*(x)\), thus such operation is a crucial for constructing efficient estimators. Our proposal to avoid unreasonable local maximizers is simply to restrict to each covariance matrices such that,

\[
\arg\text{max}_{k \in \beta} |\Sigma_k| / \arg\text{max}_{k \in \beta} |\Sigma_k| \geq \epsilon,
\]

where \(\epsilon\) is a positive real value. After applying the density estimation with some starting values, we choose a root that maximizes the weighted likelihood among the roots that meet this requirement.

### 3.3 Starting value

In the literature related to maximum likelihood fitting of finite mixture distributions, there are various proposals for specifying initial parameters. These proposals are basically applicable to our method by using a reference sample set from the importance sampling function that is previously obtained. Since application of these proposals is beyond the scope of this article, as a reference, we give only an explanation for a simple alternative in case of Gaussian components.

By using the weighted sample set \(\{ \omega(X_i), X_i \}_{i=1}^n\), the mean and variance of the optimal envelope can be estimated by

\[
\bar{X} = \sum_{i=1}^n \omega(X_i) X_i
\]

\[
V = \sum_{i=1}^n (X_i - \bar{X}) (X_i - \bar{X})^T,
\]
These are consistent estimators of $E_g(x)$, $\text{var}_g(x)$, respectively. Under this setting, the starting parameter values in the learning stage are constructed by

$$
\psi_k^{(0)} = 1/\beta,
\sum_k^{(0)} = \Gamma_w,
\mu_k^{(0)} \sim N(\Sigma_w, \Gamma_w),
$$

for $k = 1, \cdots, \beta$. In practice, good starting values produce rapid convergence of iterations in the density estimation algorithm. In contrast, poor specification often generates odd local maximizers. Thus, if it is possible to use certain information of target distribution, for example, location of modes, its number, we should take advantage of it.

### 3.4 Simulation Gain

In any implementation of importance sampling it is useful to quantify the efficiency gain provided by the obtained importance sampling function. A natural measure of efficiency gain is the estimate of $\text{var}_p(\omega(x))$ which is given by

$$
\tilde{\text{var}}_p(\omega(x)) = \frac{1}{n} \sum_{i=1}^{n} \omega(X_i)^2 - \hat{T}^2.
$$

Alternatively, we may use the normalization of it, that is $\tilde{\text{var}}_p(\omega(x)) / \hat{T}^2$. Let $T^{(l)}$ be the estimate of $\text{var}_p(\omega(x))$ constructed by the weighted sample set in lst stage. This sequence would be useful for making a stopping criterion for an AIS algorithm. The monitoring sequence $T^{(l)}$ whether a threshold is reached repeatedly provides the information whether to stop the refinement of the produced envelope. Another quantity to be checked relates to the appropriateness of tail approximation. In practice, extremely high importance weights often arise in the part of the tail of $g^*(x)$ due to violation of the imposed condition that the envelope $p(x)$ must be longer in the tail than $g^*(x)$. A candidate for assessing whether this condition holds is $(\max_i |\omega(X_i)| - \min_i |\omega(X_i)|) / |\hat{T}|$.

### 4 ILLUSTRATIVE EXAMPLE

West and Winkler (1991) is a study of data base error prediction and a posterior distribution is introduced by the literature through their statistical modeling. As an illustration of our algorithm, consider the posterior distribution of parameters $\bar{x} = [\bar{x}_1, \bar{x}_2]$ given a positive integer observation $y = [n, r, s]$ as

$$
p(\bar{x}|y) \propto \bar{g}(\bar{x}) = \{\bar{x}_1(1 - \bar{x}_2)\}^r \{\bar{x}_2(1 - \bar{x}_1)\}^s \{1 - \bar{x}_1(1 - \bar{x}_2) - \bar{x}_2(1 - \bar{x}_1)\}
$$

where $0 \leq \bar{x}_1, \bar{x}_2 \leq 1$ and $n \geq s + r$. 


West (1993) demonstrates an application of NAIS to the estimation of normalizing constant of this target with \( n = 45, r = 5, s = 3 \). We also adopt same experimental design by following. In order to construct the approximation by finite normal mixture, we define the target function taken form as \( g(x) = \tilde{g}(\tilde{x}) |J| \) where \( x_i = \logit(\tilde{x}_i) \) \( (i = 1, 2) \) and \( J \) denotes its Jacobian. Such transformation makes it possible to use Kullback-Leibler divergence as a distance measure. The interest is to estimate the normalizing constant \( I = \int_{R^2} g(x) dx \) and then the optimal distribution is clearly \( g^*(x) = g(x)/I \).

As is shown in Figure 2, the target function has highly spiked bimodal and this form of integrand results in giving a less efficient estimator by the relative diffuseness of ISF. The

![Figure 2](image2.png)

**Figure 2:** Bimodal target function \( p(x_1, x_2|y) \propto g(x_1, x_2) \) where \( g(x) = \tilde{g}(\tilde{x}) |J|, x_i = \logit(\tilde{x}_i) \) \( (i = 1, 2) \) and \( J \) denotes its Jacobian (West and Winkler 1991)

![Figure 3](image3.png)

**Figure 3:** Plots of weighted sample set at stage \( l = 0 \) (left) and \( l = 1 \) (right) (the vertical axis shows the importance weight, \( \omega(X_i) \), and \( x - y \) plane shows samples \( X_i \)). The left panel shows a reference sample set from initial ISF \( (l = 0) \). The right panel shows a weighted sample set from ISF produced by 1st stage learning.

The family of envelope \( \mathcal{M}_\beta \) employed in this experiment is 4-components normal mixture distribution with heteroscedastic covariance matrices and the initial envelope is constructed by the procedure described in Section 3.3. Such model specification for ISF is ad-hoc and we do not deal with such matter here. In addition, sample size is fixed by \( n = 3000 \) and the replacement of starting values for density estimation is repeated five times and smoothness parameter for covariance matrices is chosen to be 0.1.
The left panel in Figure 3 shows a weighted sample set from the initial envelope and the sample variance of importance weights, $\text{var}_p(\omega(x))/\tilde{I}^2$, is 5.3981. The left panel in Figure 4 describes the refined ISF, which is constructed by using it as a reference sample set of the learning stage. In addition, the right panel in Figure 4 is the plot of its 95% ellipsoids for each component distribution based on the produced parameters.

Figure 4: Importance sampling function obtained via 1st stage leaning (left) and plot of the 95% ellipsoids for each component distribution based on the produced parameters (right).

Figure 5: Comparison for the accuracy of estimators obtained from 1st stage PAIS (right), 2nd stage PAIS (center), NAIS (left) by using a sample set per a trial.

Good agreement between the target and produced envelope can be seen in comparing Figure 2 and the left panel of Figure 4, and produces less volatile importance weights as shown in Figure 3 on the right. The sample variance of importance weights is 0.8201 and it is concluded that our algorithm realizes drastic variance reduction of the estimator.

For a more detailed examination of the obtained estimator, we show 15 sequences of estimates using a different sample set in each trial. It follows from Figure 5 on the left that PAIS achieves the accuracy of 2 decimal place precision ($\max \tilde{I}_{si} = 0.1237 \times 10^{-2}$, $\min \tilde{I}_{si} = 0.12131 \times 10^{-2}$). Figure 6 shows CPU times with a Pentium-4 processor, 2.8 Ghz as a benchmark. The average time for a learning stage with 5 times replacement of starting values is 5.33 seconds. Here, we adopted the following stopping criteria, $|\theta_h^{(j)} - \theta_h^{(j-1)}| \leq 10^{-5}$, $\hat{L}(\theta^{(j)}) - \hat{L}(\theta^{(j-1)}) \leq 10^{-5}$ where subscript $h$ specifies the element index of the vector. Although the convergence of the density estimation algorithm is rapid due to
Figure 6: Comparison of computing time between PAIS (left) and NAIS (right) with a Pentium-4 processor-2.8 Ghz. This figure shows CPU times (seconds) of 20 trials. In PAIS, 5 times trials for the initial parameter values are made.

well separation of the target, the rate of convergence depends entirely upon the shape of target function, the number of components β and sample size n. In practical applications to various targets, we may have to pay attention to this matter.

Figure 7: Plots of Importance sampling function obtained via West’s GAIS (left) and the weighted sample (right)

Figure 7 describes a envelope and weighted sample set based on the NAIS algorithm proposed by West (1993). Relatively poor approximation results and, in fact, the sample estimate of importance weight is 3.8991. Figure 5 shows a comparison of estimators with PAIS and NAIS by using a different sample set in each trial. While 1st stage PAIS gives the one of 2 decimal places, NAIS cannot achieve the accuracy of 2 decimal places (\( max \hat{\mathcal{L}}_{is} = 0.1297 \times 10^{-2}, min \hat{\mathcal{L}}_{is} = 0.1175 \times 10^{-2} \) ) . The superiority of PAIS with regard to computing time is apparent from comparison between Figure 6 on the left and right. In spite of 5 times replacement of initial values in PAIS, the CPU times of NAIS is about 17 times the CPU times of PAIS.

Additionally, we implement 2nd stage learning which is performed using the sample set in Figure 3 on the right as reference. The middle panel of Figure 5 describes 15 times estimates from the produced ISF. From this, it can be confirmed that our algorithm achieves the accuracy to 3 decimal places (\( max \hat{\mathcal{L}}_{is} = 0.1227 \times 10^{-2}, min \hat{\mathcal{L}}_{is} = 0.1221 \times 10^{-2} \) ) . As compared with the 1st stage estimator, the dispersion of estimates is small.
This consequence clearly shows that improvement of the quality of the reference sample set yields more precise density estimation.

5 DIMENSIONAL EFFECT OF PAIS

In this section, we consider the dimensional effect of PAIS. In general, due to the curse of dimensionality, a direct application of PAIS to high dimensional target function like $d \geq 100$ is infeasible. Such intractability also occurs in applying NAIS and other related methods. Since this difficulty is unresolved, Unfortunately, we need to introduce mitigation tailored to each case.

In such cases, one of most essential elements for controlling the performance of the PAIS algorithm is to prepare a relatively less volatile weighted sample set at step $l = 0$. Under the circumstance that no information on target function is given, we have to make an initial set as a reference at random. For a high-dimensional target, it is difficult that such samples fall in the importance region. The difficulty of this problem arises in spite of about 20 dimensional target. Below, we will show only the applications to relative low dimensional problems ($d = 2, \cdots, 11$) for users of PAIS.

Consider the target distributions with 3-components $d$-dimensional normal mixture distributions $g^*(x)$ where the parameters are generated by

\[
\tilde{\psi}_k = \frac{1}{3}, \\
\tilde{\Sigma}_k = \text{diag}(2), \\
\tilde{\mu}_k \sim N_d(0, Q),
\]

where $\tilde{\psi}_k$, $\tilde{\mu}_k$ and $\tilde{\Sigma}_k$ correspond to mixing proportion, mean vector and covariance matrices of $k$th components, respectively. In particular, each mean vector is generated from $d$-dimensional normal distribution centered at 0 and scaled by diagonal matrices $Q = \text{diag}(4)$ and covariance matrices $\tilde{\Sigma}_k$ is simply $\text{diag}(2)$ for $k = 1, \cdots, 3$. We aim to estimate the quantity of $I = \int_{R^d} g^*(x)dx$, that is $I = 1$. While the target is changed 15 times at each $d$ ($d = 1, \cdots, 11$), we apply AIS algorithm with model $M_3$, sample size $n = 10000$. Here, the refinements of starting values for density estimation are iterated 5 times and the smoothing parameter $\epsilon$ is chosen to be 0.1.

The left panel in Figure 8 describes the boxplot of 15 times estimators with ISF after the 1st stage learning of PAIS for $d = 1, \cdots, 11$. This figure suggests that the dispersion of produced estimates is gradually increasing along with increasing dimension $d$. While in the case of $d = 6$ the accuracy of the estimator is 2 decimal places ($\max \tilde{I}_{is} = 1.002$, $\min \tilde{I}_{is} = 0.9968$), it is 1 decimal place for $d = 11$ ($\max \tilde{I}_{is} = 1.011$, $\min \tilde{I}_{is} = 0.9865$). Moreover, Figure 8 on the right shows the boxplot of log-scaled sample variance of importance weights for each trial and dimension and indicates linearly increase of it. Our simulation study indicates exponentially degradation of the precision along with dimensionality. However, the estimators produced by the PAIS algorithm are relatively efficient in spite of high
Figure 8: Box plots of 15 times estimates $\hat{I}_M$ (left) and corresponding log-scaled variance of importance weights (right).

Figure 9: Importance weight from initial envelope where the target function is 20 dimensional.

dimensional problems. In order to achieve higher accuracy, additional sampling effort, i.e. increasing sample size, applying some variance reduction techniques, compensates the efficiency loss caused by dimensionality.

Figure 9 illustrates a plot of initial importance weights for a 20-dimensional target. The parameters of sampling distribution are determined at random. The sample mean and variance are 0.5420 and 22.098, respectively. This underestimate suggests that the reference sample set never hits the importance region due to the broader space of target function. The resulting learning algorithm using such reference samples generates a poor envelope. This phenomena appears frequently for $d \geq 20$. Thus, direct application of PAIS to a high-dimensional target is actually infeasible.

6 CONCLUDING REMARKS

This article presents a PAIS algorithm for estimating an intractable integral by importance sampling and constructing very efficient estimators with less intensive computation. It is a sampling based approximator to a certain density function by following the principle of minimum Kullback-Leibler divergence. The fields of application extend to Sampling Importance Resampling (SIR) (Rubin 1987). While the PAIS algorithm improves the
efficiency of Monte Carlo estimators, the curse of dimensionality remains. In order to improve the durability to the dimensional effect of our algorithm, we need a relatively less volatile weighted sample at stage $l = 0$. To obtain one, some variance reduction techniques, for example, Rejection Control algorithm (Liu 2001), may be effective. In addition, the problem of choosing the number of components remains.

ACKNOWLEDGEMENTS
The author gratefully acknowledges the helpful comments of Professor Tomoyuki Higuchi. This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology, Grant-in-AID for Scientific Research (A) 14208025.

References


