Nonparametric Importance Sampling with Control Variates

ZHANG Yan (supervised by Dr. LI Wentao)

November 8, 2021

1 Introduction

In Monte Carlo (MC), a typical problem is to approximate the expectation or integral $\mu = \mathbb{E}_\pi[f(\boldsymbol{X})] = \int \pi(\boldsymbol{x})f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, where $\pi(\boldsymbol{x})$ is called a target distribution or a nominal distribution. A large amount of approaches for devising a point estimator for μ follows the following generic recipe:

Step 1. Generate n distinct samples $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ from the target $\pi(\boldsymbol{x})$ or some other sampling procedures independently or dependently, totally randomly or partially randomly or even totally deterministically (if Quasi-MC methodologies are applied);

Step 2. Allocate weights $\{w_1 \dots w_n\}$ (where we typically have $\sum_{i=1}^n w_i = 1$ or $\sum_{i=1}^n w_i \approx 1$) for each sample point to obtain the weighted sample set $\{(\boldsymbol{x}_1, w_1), \dots, (\boldsymbol{x}_n, w_n)\}$;

Step 3. Compute $\hat{\mu} = \sum_{i=1}^n w_i f(\boldsymbol{x}_i)$ as an estimator of μ.

How we accomplish Step 1 and Step 2 relies on the particular MC techniques being adopted, including the sampling procedures and the variance reduction methods. Different techniques will yield different sample points or the weights, which in turn will produce different estimators in Step 3.

For naive Monte Carlo, with samples $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ from $\pi(\boldsymbol{x})$, the equal weights $w_i = 1/n$ will be assigned to each sample point. We then compute the estimator $\hat{\mu}_{\text{NMC}} = \sum_{i=1}^n w_i f(\boldsymbol{x}_i) = \sum_{i=1}^n f(\boldsymbol{x}_i)/n$.

Importance Sampling (IS) came from the identity $\int \pi(\boldsymbol{x})f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int q(\boldsymbol{x})w(\boldsymbol{x})f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, where $q(\boldsymbol{x})$ is a density called an importance distribution or simply a proposal, and $w(\boldsymbol{x}) = \pi(\boldsymbol{x})/q(\boldsymbol{x})$ is a likelihood ratio function. With samples $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ from $q(\boldsymbol{x})$, the weighted set $\{(\boldsymbol{x}_1, w_1), \dots, (\boldsymbol{x}_n, w_n)\}$, $w_i = w(\boldsymbol{x}_i)/n$ is adopted. Thus, the corresponding unbiased estimator is $\hat{\mu}_{\text{IS}} = \sum_{i=1}^n w_i f(\boldsymbol{x}_i) = \sum_{i=1}^n w(\boldsymbol{x}_i)f(\boldsymbol{x}_i)/n$, which has the asymptotic variance

$$\sigma_q^2 = \mathbb{E}_q[(w(\boldsymbol{X})f(\boldsymbol{X}) - \mu)^2] = \int \frac{(\pi(\boldsymbol{x})f(\boldsymbol{x}) - \mu q(\boldsymbol{x}))^2}{q(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} = \int \frac{(\pi(\boldsymbol{x})f(\boldsymbol{x}))^2}{q(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} - \mu^2.$$

Straightly following Jensen's inequality, the last expression gives us the variance minimizer $q_{\text{opt}}(\boldsymbol{x}) \propto \pi(\boldsymbol{x})|f(\boldsymbol{x})|$, which we simply refer to as the optimal proposal or the best proposal.

Sometimes we can only compute an unnormalized version of $\pi(\boldsymbol{x})$, $\pi_u(\boldsymbol{x}) = Z_\pi \pi(\boldsymbol{x})$ where Z_π is unknown. The same may be true of the proposal $q_u(\boldsymbol{x}) = Z_q q(\boldsymbol{x})$. In such cases, the weights $w_i = w_u(\boldsymbol{x}_i)/\sum_{i=1}^n w_u(\boldsymbol{x}_i)$ are instead used, where $w_u(\boldsymbol{x}_i) = \pi_u(\boldsymbol{x}_i)/q_u(\boldsymbol{x}_i)$. The resulting self-normalized or ratio version IS estimator is $\hat{\mu}_{\text{IS,sn}} = \sum_{i=1}^n w_u(\boldsymbol{x}_i)f(\boldsymbol{x}_i)/\sum_{i=1}^n w_u(\boldsymbol{x}_i)$, which has the asymptotic variance

$$\sigma_{q,\text{sn}}^2 = \frac{\mathbb{E}_q[w_u(\boldsymbol{X})^2(f(\boldsymbol{X}) - \mu)^2]}{\mathbb{E}_q[w_u(\boldsymbol{X})]^2} = \mathbb{E}_q[w(\boldsymbol{X})^2(f(\boldsymbol{X}) - \mu)^2] = \int \frac{\pi(\boldsymbol{x})^2(f(\boldsymbol{x}) - \mu)^2}{q(\boldsymbol{x})}\mathrm{d}\boldsymbol{x},$$

indicating an optimal proposal $q_{\text{opt,sn}}(\boldsymbol{x}) \propto \pi(\boldsymbol{x})|f(\boldsymbol{x}) - \mu|$ different from the previous one.

The IS in most cases is viewed as a powerful variance reduction technique but sometimes it can be really tricky. Firstly, in the real simulations, it is often hard or impossible to directly sample from the optimal proposals, so loads of methodologies were developed to approximate them, parametrically or nonparametrically. Secondly, even if the optimal proposals are no longer the problem, the performance of IS is still restricted in many scenarios. In the non-ratio version, IS can potentially give us the zero variance estimator if and only if $f(\boldsymbol{x})$ is always positive or negative on its support. In the ratio version, $\sigma_{q,\text{sn}}^2$ is zero if and only if $f(\boldsymbol{x}) \equiv C$ for a constant C, while, in which case, we don't even need to do the estimation. The issue in non-ratio estimators was solved by [Owen & Zhou, 2000] with a simple technique called the positivisation. It is based on the decomposition $\mu = \mathbb{E}_\pi[f(\boldsymbol{X})] = \mathbb{E}_\pi[f_+(\boldsymbol{X})] - \mathbb{E}_\pi[f_-(\boldsymbol{X})]$, where $f_+(\boldsymbol{x})$ and $f_-(\boldsymbol{x})$ are the positive and negative parts of $f(\boldsymbol{x})$ respectively. We can estimate $\mathbb{E}_\pi[f_+(\boldsymbol{X})]$ and $\mathbb{E}_\pi[f_-(\boldsymbol{X})]$ separately to potentially achieve zero asymptotic variance, but this method leaves the new practical difficulties such as that two optimal proposals need to be considered instead of only one and that it is hard to figure out the most efficient way to allocation computation resource for these two parts of estimation. This work is trying to build four types of IS estimators with theoretically optimal properties that can tackle the two problems of Importance Sampling simultaneously and automatically.

2 Methodologies

2.1 Kernel Density Estimation

To deal with the first IS problem, a scheme called Adaptive Importance Sampling (AIS) was studied, where a variety of algorithms were made to approximate the best proposal iteratively, within which a method called Nonparametric Importance Sampling (NIS) and its adaptive version was proposed and developed practically and theoretically by West (1993), Givens and Raftery (1996) and Zhang (1996). Assuming that $p(\boldsymbol{x})$ is the optimal proposal and initial samples $\boldsymbol{x}_j, j = 1, \dots, m$ are from an initial proposal $q_0(\boldsymbol{x})$, NIS is to construct the optimal proposal with the weighted Kernel Density Estimation (KDE) $\hat{p}(\boldsymbol{x}) = \sum_{j=1}^m w_j K_j(\boldsymbol{x}), w_j = w_0(\boldsymbol{x}_j)/\sum_{j=1}^m w_0(\boldsymbol{x}_j)$, where $w_0(\boldsymbol{x}) = p(\boldsymbol{x})/q_0(\boldsymbol{x})$, and the d-dimensional kernels have the form

$$K_j(\boldsymbol{x}) = \frac{1}{|H_j|} K(H_j^{-1}(\boldsymbol{x} - \boldsymbol{x}_j)),$$

with three moment conditions

$$\int K(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 1, \int \boldsymbol{x} K(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \boldsymbol{0}, \int \boldsymbol{x}\boldsymbol{x}^T K(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = I_d.$$

The choice of shape matrices H_j is the center problem in KDE. West (1993) adopted a traditional way $H_j = h\hat{\Sigma}^{1/2}$ and the Givens and Raftery (1996) considered a local version $H_j = h\hat{\Sigma}_j(\gamma)^{1/2}, \gamma \in (0, 1]$, where $\hat{\Sigma}$ stands for the sample covariance matrix while $\hat{\Sigma}_j(\gamma)$ for covariance estimate of $\lceil \gamma m \rceil$ nearest neighbors of the jth point. Neighborhoods were measured by the Mahalanobis distance, and each point is included among its own neighbors. To avoid the performance of the local KDE to degrade due to under-smoothing, γ was suggested to be greater than 0.2, and these two strategies will grow more and more similar as γ approaches 1. Notice that $\hat{\Sigma}$ is a special case of $\hat{\Sigma}_j(\gamma)$ when $\gamma = 1$. In most cases, the traditional version is good enough, but the local shape matrices are especially useful when the optimal proposal we want to approximate has unusual structures, such as strong nonlinear relationships between variables.

In this paper, for theoretical and practical convenience, without special specifying, $K(\boldsymbol{x})$ is always assumed to be standard multivariate normal distribution, while, in practice, for theoretical or computational benefits, alternatives like the Epanechnikov kernel can be used (Silverman (1986)). With these kinds of kernels and generally small bandwidths, the KDE proposal $\hat{p}(\boldsymbol{x})$ often has the light tail thus would be unreliable in the IS estimation. So the mixture proposal $q(\boldsymbol{x}) = \alpha_0 q_0(\boldsymbol{x}) + (1 - \alpha_0)\hat{p}(\boldsymbol{x})$ is suggested to be used, where the initial proposal is utilized again as a tail protector with the tail protection rate α_0. In fact, we will always assume that a relative good initial proposal is given, which, at least, already covers the tails in the IS estimation problem. Furthermore, when working with the mixture proposal $q(\boldsymbol{x})$, instead of generating n samples directly from it, a standard technique is to draw stratified samples $\{\boldsymbol{x}_{j1}, \dots, \boldsymbol{x}_{jn_j}\}$ with deterministic sample size $n_j = \alpha_j n$ from the jth kernel $K_j(\boldsymbol{x}), j = 0, \dots, m$, where $K_0(\boldsymbol{x}) = q_0(\boldsymbol{x})$ and α_j is the corresponding proportion in the mixture. So, as one example, the stratified version of the non-ratio IS is $\hat{\mu}_{\text{IS,stra}} = \sum_{j=0}^m \sum_{i=1}^{n_j} w(\boldsymbol{x}_{ji}) f(\boldsymbol{x}_{ji})/n$.

As for the theoretical aspect of the kernel method, according to Givens (1995), if

$$\lim_{m\to\infty} m(\min_j |H_j|) \to \infty \text{ a.s.}, \lim_{m\to\infty} (\max_j ||H_j||_\infty) \to 0 \text{ a.s.}, \tag{1}$$

where $|H_j|$ is the absolute value of the determinant and $||H_j||_\infty$ is the infinity norm, there would be

$$\lim_{m\to\infty}\int|\hat{p}(\boldsymbol{x})-p(\boldsymbol{x})|\mathrm{d}\boldsymbol{x}\to 0 \text{ a.s.},$$

which guarantees the $\mathcal{L}_1$ strong consistency of the non-local or local KDE when h decreases at an appropriate slow rate as m goes to infinity. And with the non-local KDE and a bandwidth $h\propto m^{-1/(d+4)}$, if $n=Cm$ samples $(C>0)$ are drawn from $\hat{q}(\boldsymbol{x})$ for a non-ratio IS estimation where a zero variance estimator exists, Zhang (1996) proved that its MSE is of the order $\mathcal{O}(m^{-(d+8)/(d+4)})$. As we have seen, KDE provided a promising path to tackle the first IS problem, and through the Control Variates technique, we will solve the second one.

2.2 Control Variates

Control Variates (CV) combined with regression provides a very powerful way in variance reduction, especially useful when we have a mixture density (Owen and Zhou (2000)). CV can exploit the known to tackle the unknown. In the construction of the mixture proposal by weighted KDE, what is known is that $\int\boldsymbol{g}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=\boldsymbol{0},\boldsymbol{g}(\boldsymbol{x})=(K_1(\boldsymbol{x})-q(\boldsymbol{x}),\ldots,K_m(\boldsymbol{x})-q(\boldsymbol{x}))^T$, and with a vector $\boldsymbol{\beta}$, the following unbiased estimator can be established

$$\hat{\mu}_{\text{CV}}(\boldsymbol{\beta})=\frac{1}{n}\sum_{i=1}^{n}\frac{\pi(\boldsymbol{x}_i)f(\boldsymbol{x}_i)-\boldsymbol{\beta}^T\boldsymbol{g}(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)},\boldsymbol{x}_i\sim q(\boldsymbol{x}),$$

which has the asymptotic variance

$$\sigma_q^2(\boldsymbol{\beta})=\mathbb{E}_q[(\frac{\pi(\boldsymbol{X})f(\boldsymbol{X})-\boldsymbol{\beta}^T\boldsymbol{g}(\boldsymbol{X})}{q(\boldsymbol{X})}-\mu)^2]=\int\frac{(\pi(\boldsymbol{x})f(\boldsymbol{x})-\mu q(\boldsymbol{x})-\boldsymbol{\beta}^T\boldsymbol{g}(\boldsymbol{x}))^2}{q(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}.$$

To minimize the variance with respect to $\boldsymbol{\beta}$, easy to see that it is equivalent to use all the proposal components $K_j(\boldsymbol{x}),j=0,\ldots,m$ to linearly approximate $t_{\text{Reg}}(\boldsymbol{x})=\pi(\boldsymbol{x})f(\boldsymbol{x})$, which is termed the regression target in this work.

Besides, the ratio version estimator with control variates is

$$\hat{\mu}_{\text{CV,sn}}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)=\sum_{i=1}^{n}\frac{\pi_u(\boldsymbol{x}_i)f(\boldsymbol{x}_i)-\boldsymbol{\beta}_1^T\boldsymbol{g}(\boldsymbol{x}_i)}{q_u(\boldsymbol{x}_i)}/\sum_{i=1}^{n}\frac{\pi_u(\boldsymbol{x}_i)-\boldsymbol{\beta}_2^T\boldsymbol{g}(\boldsymbol{x}_i)}{q_u(\boldsymbol{x}_i)},$$

with

$$\sigma_{q,\text{sn}}^2(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)=\mathbb{E}_q[(\frac{\pi(\boldsymbol{X})(f(\boldsymbol{X})-\mu)-\boldsymbol{\beta}^T\boldsymbol{g}(\boldsymbol{X})}{q(\boldsymbol{X})})^2]=\int\frac{(\pi(\boldsymbol{x})(f(\boldsymbol{x})-\mu)-\boldsymbol{\beta}^T\boldsymbol{g}(\boldsymbol{x}))^2}{q(\boldsymbol{x})}\mathrm{d}\boldsymbol{x},$$

where $\boldsymbol{\beta}=\boldsymbol{\beta}_1-\mu\boldsymbol{\beta}_2$. This time, a different regression target $t_{\text{Reg,sn}}(\boldsymbol{x})=\pi(\boldsymbol{x})(f(\boldsymbol{x})-\mu)$ should be approached by the proposal components.

A particularly interesting fact is that, to better approximate the above two regression targets, a very natural way is to build kernels centered at samples draw from distributions whose densities are proportional to $|t_{\text{Reg}}(\boldsymbol{x})|$ and $|t_{\text{Reg,sn}}(\boldsymbol{x})|$ respectively, which are exactly the corresponding optimal proposals. Thus, with a little adjustment that will be mentioned later, the KDE-based proposal construction methodology fits quite well with the regression-based variance reduction methodology. By regression, it means that the optimal $\boldsymbol{\beta}$ can be typically estimated via the least square estimation

$$\hat{\boldsymbol{\beta}}_{\text{ls}}=\widehat{\text{Var}}_q[\frac{\boldsymbol{g}(\boldsymbol{X})}{q(\boldsymbol{X})}]^{-1}\widehat{\text{Cov}}_q[\frac{\boldsymbol{g}(\boldsymbol{X})}{q(\boldsymbol{X})},\frac{\pi(\boldsymbol{X})f(\boldsymbol{X})}{q(\boldsymbol{X})}],$$

and the non-ratio form of the regression estimator is $\hat{\mu}_{\text{Reg}}=\hat{\mu}_{\text{CV}}(\hat{\boldsymbol{\beta}}_{\text{ls}})$. Note that the same superscript of $\hat{\mu}_{\text{CV}}(\boldsymbol{\beta})$ and $\hat{\boldsymbol{\beta}}_{\text{ls}}$ means that that these estimators are actually based on the same set of samples, and thus will result in mildly annoying bias, which can be get rid of by a Cross Validation-like approach from Avramidis and Wilson (1993). But we don't talk about it in this work.

The generic recipe mentioned before can also explain the least square-based estimator. Firstly, we have the samples $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ from $q(\boldsymbol{x})$, with which estimate

$$\hat{\boldsymbol{\zeta}}_{\text{ls}} = \widehat{\text{Var}}_q[\frac{\boldsymbol{g}(\boldsymbol{X})}{q(\boldsymbol{X})}]^{-1}\widehat{\mathbb{E}}_q[\frac{\boldsymbol{g}(\boldsymbol{X})}{q(\boldsymbol{X})}].$$

And then, the adjusted weight for the sample point $\boldsymbol{x}_i$ is

$$w_i = w(\boldsymbol{x}_i)(1 - \hat{\zeta}_{\text{ls}}^T(\frac{\boldsymbol{g}(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)} - \widehat{\mathbb{E}}_q[\frac{\boldsymbol{g}(\boldsymbol{X})}{q(\boldsymbol{X})}]))/n,$$

which, in Step 3, will give us the exact same estimation value of $\hat{\mu}_{\text{Reg}}$. This version of regression estimator is particular useful when $\hat{\boldsymbol{\beta}}_{\text{ls}}$ should be calculated multiple times. For examples, we need to estimate expectations for different $\pi(\boldsymbol{x})$ or $f(\boldsymbol{x})$, or we are doing quantile estimation (Hesterberg (1998)), or we need to estimate $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ separately in the self-normalized estimator $\hat{\mu}_{\text{Reg,sn}} = \hat{\mu}_{\text{CV}}(\hat{\boldsymbol{\beta}}_{1,\text{ls}}, \hat{\boldsymbol{\beta}}_{2,\text{ls}})$, where

$$\hat{\boldsymbol{\beta}}_{1,\text{ls}} = \widehat{\text{Var}}_q[\frac{\boldsymbol{g}(\boldsymbol{X})}{q_u(\boldsymbol{X})}]^{-1}\widehat{\text{Cov}}_q[\frac{\boldsymbol{g}(\boldsymbol{X})}{q_u(\boldsymbol{X})}, \frac{\pi_u(\boldsymbol{X})f(\boldsymbol{X})}{q_u(\boldsymbol{X})}], \hat{\boldsymbol{\beta}}_{2,\text{ls}} = \widehat{\text{Var}}_q[\frac{\boldsymbol{g}(\boldsymbol{X})}{q_u(\boldsymbol{X})}]^{-1}\widehat{\text{Cov}}_q[\frac{\boldsymbol{g}(\boldsymbol{X})}{q_u(\boldsymbol{X})}, \frac{\pi_u(\boldsymbol{X})}{q_u(\boldsymbol{X})}].$$

The corresponding sample weight would be

$$w_i = w_u(\boldsymbol{x}_i)(1 - \hat{\boldsymbol{\zeta}}_{\text{ls}}^T(\frac{\boldsymbol{g}(\boldsymbol{x}_i)}{q_u(\boldsymbol{x}_i)} - \widehat{\mathbb{E}}_q[\frac{\boldsymbol{g}(\boldsymbol{X})}{q_u(\boldsymbol{X})}]))/\sum_{i=1}^{n} w_u(\boldsymbol{x}_i)(1 - \hat{\boldsymbol{\zeta}}_{\text{ls}}^T(\frac{\boldsymbol{g}(\boldsymbol{x}_i)}{q_u(\boldsymbol{x}_i)} - \widehat{\mathbb{E}}_q[\frac{\boldsymbol{g}(\boldsymbol{X})}{q_u(\boldsymbol{X})}])),$$

where

$$\hat{\boldsymbol{\zeta}}_{\text{ls}} = \widehat{\text{Var}}_q[\frac{\boldsymbol{g}(\boldsymbol{X})}{q_u(\boldsymbol{X})}]^{-1}\widehat{\mathbb{E}}_q[\frac{\boldsymbol{g}(\boldsymbol{X})}{q_u(\boldsymbol{X})}],$$

and it only need to be calculated once to have $\hat{\mu}_{\text{Reg,sn}} \equiv \sum_{i=1}^{n} w_i f(\boldsymbol{x}_i)$.

So, as long as the regression targets can be linearly expressed or approximated by the initial proposal and the kernels, the regression estimators discussed above would exactly or asymptotically have zeros asymptotic variance. The advantage of these estimator is also justified by a theoretically optimal and asymptotically equivalent approach, the Likelihood Approach.

2.3 Likelihood Approach

When talking about weights allocation in Step 2, a theoretical optimal approach was proposed by Tan (2004) under the framework of Kong et al. (2003). Left the theoretical aspect behind, an intuitive introduction will be made here.

With the stratified samples given before, the core of the likelihood approach is to solve the nonparametric maximum likelihood problem

$$\hat{\boldsymbol{\zeta}} = \underset{\boldsymbol{\zeta}}{\text{argmax}}\frac{1}{n}\sum_{j=0}^{m}\sum_{i=1}^{n_j}\log(q(\boldsymbol{x}_{ji}) + \boldsymbol{\zeta}^T\boldsymbol{g}(\boldsymbol{x}_{ji})),$$

to derive a discrete measure $\hat{\nu} = \frac{1}{n}\mathbf{1}_n(\boldsymbol{x})/(q(\boldsymbol{x}) + \hat{\boldsymbol{\zeta}}^T\boldsymbol{g}(\boldsymbol{x}))$ supported by the n sample points, which result in the adjusted weights $w_i = \frac{1}{n}\pi(\boldsymbol{x}_{ji})/(q(\boldsymbol{x}_{ji}) + \hat{\boldsymbol{\zeta}}^T\boldsymbol{g}(\boldsymbol{x}_{ji}))$. A close examination can reveal the facts that $\hat{\boldsymbol{\zeta}}$ asymptotically equals $\hat{\boldsymbol{\zeta}}_{\text{ls}}$ and these MLE weights are asymptotically equivalent to the regression weights.

With the Newton–Raphson algorithm, an equivalent problem is to solve the equations

$$\frac{1}{n}\sum_{j=0}^{m}\sum_{i=1}^{n_j}\frac{\boldsymbol{g}(\boldsymbol{x}_{ji})}{q(\boldsymbol{x}_{ji}) + \hat{\boldsymbol{\zeta}}^T\boldsymbol{g}(\boldsymbol{x}_{ji})} = \mathbf{0},$$

which are to utilize the information that $\int \boldsymbol{g}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \mathbf{0}$. And because $\int q(\boldsymbol{x})\mathrm{d}\hat{\nu} = 1 - \hat{\boldsymbol{\zeta}}^T \int \boldsymbol{g}(\boldsymbol{x})\mathrm{d}\hat{\nu} = 1$, any $\mu = \int \pi(\boldsymbol{x})f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ where $\pi(\boldsymbol{x})f(\boldsymbol{x})$ can be linearly expressed by $K_j(\boldsymbol{x}), j = 0, \ldots, m$ can be estimated by

$\hat{\mu}_{\text{MLE}} = \int \pi(\boldsymbol{x})f(\boldsymbol{x})\mathrm{d}\hat{\nu}$ with zero variance. And it is also intuitively true that if $\pi(\boldsymbol{x})f(\boldsymbol{x})$ can be linearly approximated by the proposal components, its integral will be estimated by the MLE estimator with little variance.

As the regression approach and the likelihood approach are based on the exploitation of the same information, they can achieve the same asymptotic efficiency, the highest asymptotic efficiency among the large set of estimators in the form of

$$\sum_{j=0}^{m} \frac{1}{n_j} \sum_{i=1}^{n_j} \omega_j(\boldsymbol{x}_{ji}) \frac{\pi(\boldsymbol{x}_{ji})f(\boldsymbol{x}_{ji}) - \boldsymbol{\beta}_j(\boldsymbol{x}_{ji})^T \boldsymbol{g}(\boldsymbol{x}_{ji})}{q_j(\boldsymbol{x}_{ji})},$$

where $\text{supp}(\omega_j(\boldsymbol{x})) \subset \text{supp}(q_j(\boldsymbol{x}))$, $\sum_{j=0}^{m} \omega_j(\boldsymbol{x}) \equiv 1$ and $\sum_{j=0}^{m} \omega_j(\boldsymbol{x})\boldsymbol{\beta}_j(\boldsymbol{x}) \equiv \boldsymbol{b}$ for some constant vector $\boldsymbol{b}$.

To sum up, as $\hat{\mu}_{\text{Reg}}$ and $\hat{\mu}_{\text{MLE}}$ dominate the other estimators, they will be mainly considered in this paper as well as their ratio versions, $\hat{\mu}_{\text{Reg,sn}}$ and $\hat{\mu}_{\text{MLE,sn}}$. Combining them with KDE, an asymptotically zeros-variance IS estimation scheme is proposed.

2.4 Algorithm

Assume we already have a relative good initial proposal $q_0(\boldsymbol{x})$, here is the algorithm:

Step 1: Draw $n_0 = r_{\text{SIR}} m_0$ initial samples from $q_0(\boldsymbol{x})$, and do Sampling Importance Resampling (SIR) against the optimal proposal $p(\boldsymbol{x}) = q_{\text{opt}}(\boldsymbol{x})$ or $p(\boldsymbol{x}) = q_{\text{opt,sn}}(\boldsymbol{x})$ to obtain the weighted samples: $\{(\boldsymbol{x}_j, w_j);\ j = 1, \ldots, m\}$ $(\sum_{j=1}^{m} w_j = 1)$.

Step 2: Based on prior knowledge and cluster analysis (hierarchical methods, k-means), divide samples into different groups: $k : \{1, \ldots, m\} \to \{1, \ldots, K_{\text{Group}}\}$.

Step 3: Calculate all groups' Scott factors $r_{k(j)} = m_{k(j)}^{-1/(d+4)}$, where $m_{k(j)}$ means the effective sample size for the sample j's group $k(j)$. Build KDE $\hat{p}(\boldsymbol{x})$ by constructing $H_j = h r_{k(j)} \gamma^{-1/d} \hat{\Sigma}_{k(j)}(\gamma)^{1/2}$ (or $H_j = h_{k(j)} r_{k(j)} \gamma_{k(j)}^{-1/d} \hat{\Sigma}_{k(j)}(\gamma_{k(j)})^{1/2}$) and choice the smoothing parameter(s) h $(h_{k(j)})$ and sensitive parameter(s) γ $(\gamma_{k(j)})$ manually or automatically.

Step 4: Draw Monte Carlo samples $\{\boldsymbol{x}_i^*;\ i = 1, \ldots, n\}$ from $q(\boldsymbol{x}) = \alpha_0 q_0(\boldsymbol{x}) + (1 - \alpha_0)\hat{p}(\boldsymbol{x})$, and do estimation based on regression or likelihood estimators.

[r_{SIR}: $n_0/\text{ESS} \sim n_0/\text{RSS}$ (ESS: $(\sum_{i=1}^{n_0} w_{0i})^2 / \sum_{i=1}^{n_0} w_{0i}^2$; RSS: $\sum_{i=1}^{n_0} w_{0i} / \max(w_{0i})$)]

3 Theoretical Part

4 Experiments

4.1 Normalization Constant for Normal Distribution

Based on an experiment of normal kernels against a standard normal target, here are the conclusions:

Conclusion 1.1: When $\gamma = 1$, the rule-of-thumb works perfectly for ISE, but the optimal bandwidth for the KL-divergence is a little larger, which can, however, be used as a reliable bandwidth for the KDE based or mixture proposal based importance sampling.

Conclusion 1.2: The most noteworthy phenomenon is that the best bandwidth for regression estimator is typically totally different from those for others, always larger than yet becomes closer and closer to those in this example.

Conclusion 1.3: The performance of LAIS is always worse than that of GAIS. The performance of regression estimator is dramatically great in lower dimension but quickly becomes bad as dimension increases.

Conclusion 1.4: At the optimal bandwidth of regression, the mixture proposal usually performs worse than that of the KDE proposal, which would becomes better when the bandwidth is too large or too small.

Conclusion 2.1: The lower reference for resampling ratio increase relatively slow while the upper reference increase very fast. The calculated upper reference becomes notable less than the theoretical value as the dimension increases.

Cal(100000)/The	1D	2D	3D	4D	5D	6D	7D	8D	9D	10D
n_0/ESS	2/2	2/2	3/3	5/5	8/8	12/12	18/18	26/27	39/41	60/62
n_0/RSS	2/2	4/4	8/8	16/16	31/32	57/64	106/128	166/256	222/512	474/1024

Conclusion 2.2: The resampling ratio don't influence a lot the performance of KDE based or mixture proposal based importance sampling at the corresponding optimal bandwidths, but can influence a lot the regression estimation.

Conclusion 2.3: The position of the best bandwidth is invariant against the resampling ratio. As the ratio increase, the asymptotic variance of the regression estimation decreases quickly until 60 which is just between the corresponding references 18/18 - 106/128.

Conclusion 3.1: As the resampling size increases, the performance of the regression can be improved much faster than that of the previous estimators, which may indicates a faster convergence rate of the estimation MSE than that of simply NIS.

Conclusion 3.2: The optimal bandwidths for the normal IS estimators move left as the resampling size increases, while move right for the regression IS estimator.

Conclusion 4.1: The performance of the regression is quite invariant against the tail protection rate when it is small enough. The position of the optimal bandwidth is also invariant against this rate. But the performance of regression from mixture to regression is greater when the rate is bigger.

Conclusion 5.1: The performance of the regression is quite invariant against the estimation size when it is large enough. The position of the optimal bandwidth is also invariant against this size.

Conclusion 6.1: Regression and MLE estimators can achieve the least nMSE, amount which MLE estimator can give a slightly smaller nMSE. Ridge, Lasso and unbiased regression estimators can't give better results than ordinary least square regression estimation.

	IS	NIS	MIS	RIS(O)	RIS(R)	RIS(L)
$\widehat{\text{aVar}}$	17.2302	1.0224	1.2106	0.0251	0.0254	0.0547
aVar	15.3185	0.8401	0.5457	0.0220	0.0224	0.0334
nMSE	16.3809	0.8458	0.5471	0.0222	0.0226	0.0338
	RIS(O,u)	RIS(R,u)	RIS(L,u)	RIS(T)	MLE(T)	MLE(O)
$\widehat{\text{aVar}}$	0.0253	0.0256	0.0548			
aVar	0.0272	0.0271	0.0415	0.0220	0.0236	0.0221
nMSE	0.0286	0.0286	0.0474	0.0222	0.1422	0.0221

Conclusion 7.1: The type of the kernel shape do play a role in importance sampling. It can influence the position of optimal bandwidth, especially for the regression-based estimation. The t kernel KDE with a small degree of freedom require a larger bandwidth because the mode is narrower. Interestingly, the best improvement made by regression don't change a lot with the change of kernel.

Conclusion 0.0: Parameter tuning:

(1) $\alpha_0 = 0.1$;

(2) $r_{\text{SIR}} \in [n_0/\text{ESS}, n_0/\text{RSS}]$, the larger the better, computation consideration in initial sampling;

(3) k, prior knowledge and cluster analysis;

(4) m_0, the larger the better, computation consideration in regression;

(5) γ, the specific structure of the optimal proposal;

(6) h, theoretical analysis or manual search;

(7) n, the bigger the better, computation consideration in regression.

4.2 Probability Estimation for Normal Distribution

4.3 Quantile Estimation for Normal Distribution

4.4 Quantile Estimation for GARCH(1,1)

5 Conclusion

References

[Owen & Zhou, 2000] Owen, A. & Zhou, Y. (2000). Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449), 135–143.