A Quadrature Rule combining Control Variates and Adaptive Importance Sampling

Rémi Leluc[1], François Portier[2], Johan Segers[3], Aigerim Zhuman[3]
[1] LTCI, Télécom Paris, [2] CREST, ENSAI, [3] LIDAM, ISBA, UCLouvain

TELECOM Paris · IP PARIS · ENSAI École nationale de la statistique et de l'analyse de l'information · LIDAM Louvain Institute of Data Analysis and Modeling in economics and statistics

Sequential Framework

• **GOAL:** Given **integrand** $g:\mathbb{R}^d\to\mathbb{R}$ and target **density function** f: $\boxed{I=\mathbb{E}_f[g]=\int_{\mathbb{R}^d} g(x)f(x)\,\mathrm{d}x}$
(f is posterior distribution in Bayesian)

• $(q_i)_{i\geq 0}$ is the **policy** of the algorithm: a sequence of densities which evolves adaptively depending on previous outcomes.

• **particles** $(X_i)_{i\geq 1}$ are generated sequentially according to policy $X_i\sim q_{i-1}$ with **importance weights** $w_i=f(X_i)/q_{i-1}(X_i)$.

• The integral is estimated by the normalized sum

$$\boxed{I_n^{(\mathrm{ais})}=\left(\sum_{i=1}^n w_i g(X_i)\right)\Big/\left(\sum_{i=1}^n w_i\right)}$$

Control Variates and OLS

• $h=(h_1,\dots,h_m)^\top$ vector of **control variates** *i.e.* functions such that integral $\int h_k f\,\mathrm{d}\lambda$ is known. w.l.o.g. $\mathbb{E}_f[h]=0$. For any $\beta\in\mathbb{R}^m$, $\mathbb{E}_f[g-\beta^\top h]=\mathbb{E}_f[g]$ yielding unbiased estimator

$$\boxed{I_n^{(\mathrm{cv})}(g,\beta)=\frac{1}{n}\sum_{i=1}^n\left(g(X_i)-\beta^\top h(X_i)\right)}$$

• Provided matrix $G=\mathbb{E}_f[hh^\top]$ is invertible, there is a unique $\beta^*\in\mathbb{R}^m$ for which the variance of $I_n^{(\mathrm{cv})}(g)$ is minimal: $\beta^*=\left(\mathbb{E}_f[hh^\top]\right)^{-1}\mathbb{E}_f[hg]$.

• Casting the problem in an **Ordinary Least Squares** framework leads to the control variate estimate

$$I_n^{(\mathrm{cv})}(g)=I_n^{(\mathrm{cv})}(g,\hat\beta_n^{(\mathrm{cv})})=\hat\alpha_n^{(\mathrm{cv})}\quad\text{where}$$

$$(\hat\alpha_n^{(\mathrm{cv})},\hat\beta_n^{(\mathrm{cv})})\in\operatorname*{arg\,min}_{(a,b)\in\mathbb{R}\times\mathbb{R}^m}\frac{1}{n}\sum_{i=1}^n\{g(X_i)-a-b^\top h(X_i)\}^2$$

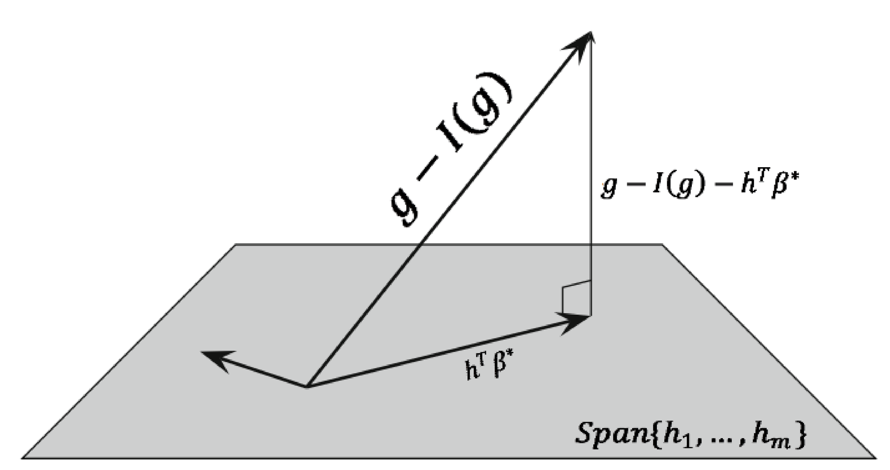

Figure 1: L^2 projection of the integrand g onto the space of control variates $Span\{h_1,\dots,h_m\}$.

AISCV Estimator and wOLS

• AISCV estimate is the first coordinate of the solution to the **weighted least squares** problem

$$\boxed{(\hat\alpha_n,\hat\beta_n)=\operatorname*{arg\,min}_{a\in\mathbb{R},b\in\mathbb{R}^m}\sum_{i=1}^n w_i\left[g(X_i)-a-b^\top h(X_i)\right]^2}$$

• (a) (Exact integration) whenever g is of the form $\alpha+\beta^\top h$ for some $\alpha\in\mathbb{R}$ and $\beta\in\mathbb{R}^m$, **the error is zero**, i.e., $\hat\alpha_n=\alpha=\int gf\,\mathrm{d}\lambda$.

• (b) (Quadrature Rule) the estimate takes the form of a quadrature rule $\hat\alpha_n=\sum_{i=1}^n v_{n,i}g(X_i)$, for **quadrature weights** $v_{n,i}$ **that do not depend on the function** g and that can be computed by a single weighted least squares procedure.

• (c) (Bayesian) it can be computed even when f **is known only up to a multiplicative constant**.

• (d) (post-hoc scheme) CV can be brought into play in a ***post-hoc* scheme**, after generation of the particles and importance weights, and **this for any AIS algorithm**

AISCV Algorithm

Require: integrand g, target density f, stages $T\in\mathbb{N}^*$, allocation policy $(n_t)_{t=1}^T$, initial density q_0, update rule for the sampling policy

1: **for** $t=1,\dots,T$ **do**
2: Generate $X_{t,1},\dots,X_{t,n_t}\sim q_{t-1}$
3: Compute the vector of weights $(w_{t,i})_{i=1}^{n_t}$ where
4: $w_{t,i}=f(X_{t,i})/q_{t-1}(X_{t,i})$
5: Build CV matrix $H_t=\left(h_j(X_{t,i})\right)_{i=1,\dots,n_t}^{j=1,\dots,m}$
6: Evaluate integrand on particles: $(g(X_{t,i}))_{i=1}^{n_t}$
7: Update the sampler q_t based on all previous particles $(X_{s,i}:s=1,\dots,t;i=1,\dots,n_s)$
8: **end for**
9: $(\hat\alpha_T,\hat\beta_T)=\operatorname*{arg\,min}_{(a,b)\in\mathbb{R}\times\mathbb{R}^m}\Phi(a,b)$ with
10: $\Phi(a,b)=\sum_{t=1}^T\sum_{i=1}^{n_t}w_{t,i}\left(g(X_{t,i})-a-b^\top h(X_{t,i})\right)^2$
11: **return** $I_n^{(\mathrm{aiscv})}(g)=\hat\alpha_T$.

Non-asymptotic bound

Theorem 1 (Concentration inequality). *Under assumptions, for any $\delta\in(0,1)$ and for all $n\geq C_1c^2B\log(10m/\delta)$, we have, with probability at least $1-\delta$,*

$$\left|I_n^{(\mathrm{aiscv})}(g)-I\right|\leq C_2\tau\sqrt{\frac{\log(10/\delta)}{n}}+C_3cB\tau\frac{\log(10m/\delta)}{n},$$

where C_1, C_2, C_2 are universal constants and $B=\sup_{x:f(x)>0}\|\hbar(x)\|_2^2$. with $\hbar=G^{-1/2}h$.

AISCV post-hoc scheme

Require: integrand g, $T\in\mathbb{N}^*$, allocation policy $(n_t)_{t=1}^T$, weights $(w_t)_{t=1}^T$ with $w_t=(w_{t,i})_{i=1}^{n_t}$, matrices $(H_t)_{t=1}^T$ with $H_t=\left(h_j(X_{t,i})\right)_{i=1,\dots,n_t}^{j=1,\dots,m}$, particles $(X_{t,i}:t=1,\dots,T;i=1,\dots,n_t)$

1: $\hat\beta_n(\mathbf{1}_n)=\operatorname*{arg\,min}_{b\in\mathbb{R}^m}\sum_{t=1}^T\sum_{i=1}^{n_t}w_{t,i}\left(1-b^\top h(X_{t,i})\right)^2$
2: $u_t=\mathrm{diag}(w_t)[\mathbf{1}_{n_t}-H_t\hat\beta_n(\mathbf{1}_n)]$ for $t=1,\dots,T$
3: Compute $s=\sum_{t=1}^T\sum_{i=1}^{n_t}u_{t,i}$
4: Compute $v_{t,i}=u_{t,i}/s$ for $1\leq t\leq T;1\leq i\leq n_t$
5: **return** $I_T^{(\mathrm{aiscv})}(g)=\sum_{t=1}^T\sum_{i=1}^{n_t}v_{t,i}g(X_{t,i})$

CV in practice

• **Stein control variates** [1] are built with operator $\mathcal{L}$ on functions $\varphi\in\mathcal{C}^2(\mathbb{R}^d,\mathbb{R})$ to have $\mathbb{E}_f[\mathcal{L}\varphi]=0$.
$(\mathcal{L}\varphi)(x)=\Delta_x\varphi(x)+\nabla_x\varphi(x)^\top\nabla_x\log f(x)$.

• $\nabla_x\log f(x)$ can either be directly computed (Bayesian regression) or with autodiff (Tensorflow and PyTorch).

Bayesian inference

• Given data $\mathcal{D}$ and parameter of interest $\theta\in\mathbb{R}^d$, posterior integrals take the form $\int_{\mathbb{R}^d}g(\theta)p(\theta|\mathcal{D})\,\mathrm{d}\theta$, where $p(\theta|\mathcal{D})\propto\ell(\mathcal{D}|\theta)\pi(\theta)$ is the posterior distribution, proportional to prior $\pi(\cdot)$ and a likelihood $\ell(\mathcal{D}|\cdot)$.

• (Linear regression) $\ell(X,y|\theta)$ is proportional to $(\sigma^2)^{-N/2}\exp(-(y-X\theta)^\top(y-X\theta)/(2\sigma^2))$, yielding the score function $\nabla_\theta\log\ell(X,y|\theta)=X^\top(y-X\theta)/(2\sigma^2)$.

• (Logistic regression) $\ell(X,y|\theta)=\prod_{i=1}^N\sigma(\theta^\top x_i)^{y_i}(1-\sigma(\theta^\top x_i))^{1-y_i}$. The score function is simply $\nabla_\theta\log\ell(X,y|\theta)=X^\top(y-\sigma(X\theta))$.

Numerical experiments

• sampling policy is multivariate Student t of degree ν denoted by $\{q_{\mu,\Sigma_0}:\mu\in\mathbb{R}^d\}$ with $\Sigma_0=\sigma_0 I_d(\nu-2)/\nu$ and $\nu>2,\sigma_0>0$. The mean μ_t is updated by the generalized method of moments (GMM), leading to $\mu_t=(\sum_{s=1}^t\sum_{i=1}^{n_s}w_{s,i}X_{s,i})/(\sum_{s=1}^t\sum_{i=1}^{n_s}w_{s,i})$ [2].

• The allocation policy is fixed to $n_t=1000$ and the number of stages is $T\in\{5;10;20;30;50\}$.

• $g(x)=x$, $f_\Sigma(x)=0.5\Phi_\Sigma(x-\mu)+0.5\Phi_\Sigma(x+\mu)$ where $\mu=(1,\dots,1)^\top/2\sqrt{d}$, $\Sigma=I_d/d$ and Φ_Σ is pdf $\mathcal{N}(0,\Sigma)$.

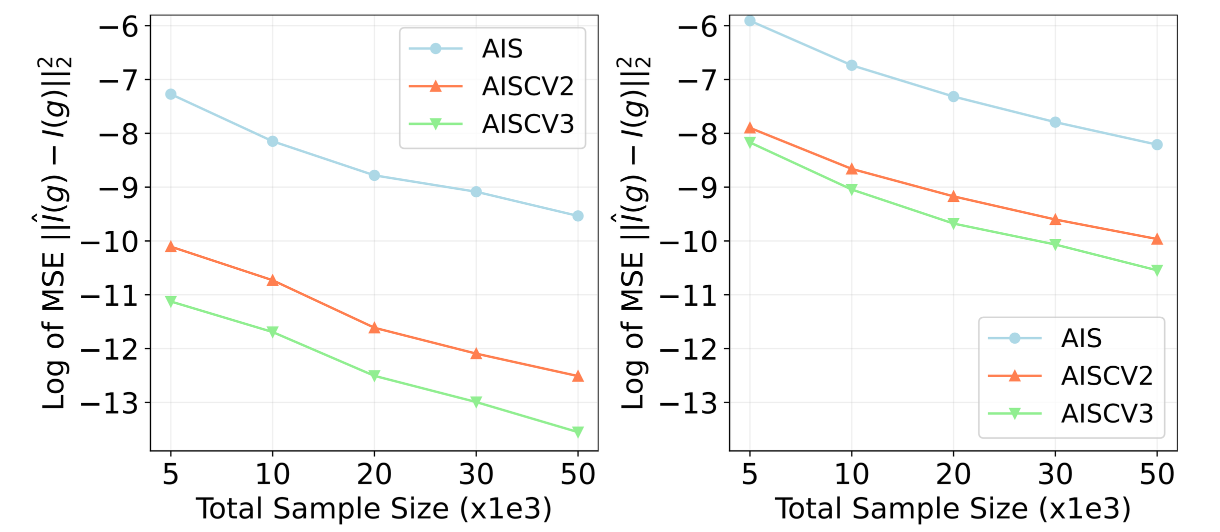

Figure 2: Gaussian mixture density: Logarithm of $\|\hat I(g)-I(g)\|_2^2$ for $g(x)=x$ with target isotropic f_Σ with $d=4$ (left), $d=8$ (right).

• **(Bayesian LR)** $g(\theta)=\sum_{i=1}^d\theta_i^2$ with Stein CV out of monomials with total degree $Q\in\{1;2\}$.

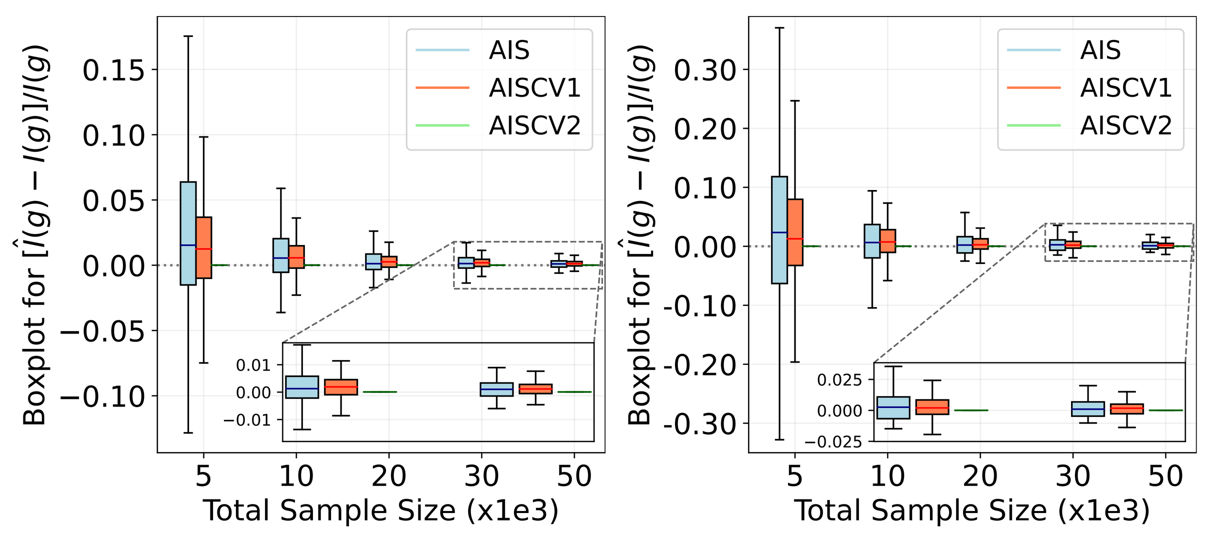

Figure 3: BLR: boxplots of $(\hat I(g)-I(g))/I(g)$ for $g(\theta)=\sum_{j=1}^d\theta_j^2$ with datasets Housing (left) and Abalone (right).

References

[1] C. J. Oates, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):695–718, 2017.

[2] F. Portier and B. Delyon. Asymptotic optimality of adaptive importance sampling. *Advances in Neural Information Processing Systems*, 31, 2018.