Proceedings of CONF-APMM 2025 Symposium: Applied Mathematics and Quantitative Modeling
DOI: 10.54254/2753-8818/2026.PU33589

A Natural Gradient Descent Algorithm of Importance Sampling

Yi Cheng

School of Mathematics, Sichuan University, Chengdu, China
3574148493@qq.com

Abstract. Importance Sampling (IS) is a fundamental variance reduction technique in Monte Carlo simulation, particularly for estimating rare event probabilities. Its core idea is to find a better sampling measure $\mathbb{Q}$ under which the target events occur more frequently, to reduce the cost of simulation of rare events. However, determining such an optimal $\mathbb{Q}$ is notoriously difficult, often leading to complicated partial differential equations (PDE) that many classical methods fail. This paper aims to introduce a novel geometric perspective of this optimization problem that we consider all proper likelihood ratios as a manifold, and use the natural gradient property to design the algorithm, which avoids PDE and outperforms the ordinary stochastic gradient descent (SGD) algorithm. Under standard regularity and Novikov conditions, we establish the almost sure convergence of the proposed algorithm utilizing the Robbins-Siegmund theorem and end up with numerical validation on option pricing and portfolio risk assessment, confirming that the geometric approach significantly enhances variance reduction efficiency.

Keywords: Importance Sampling, Monte Carlo, Riemannian Manifold, Stochastic Gradient Descent, Algorithm Analysis

1. Introduction

Importance Sampling is an often-used method to reduce the variance when estimating probabilities of rare events in Monte Carlo. We replace the original measure $\mathbb{P}$ with another one $\mathbb{Q}$, where the events are more frequent, and use the unbiased estimator $\hat{e} = \frac{1}{N}\sum h(X_i)L(X_i)$, with L defined as the Radon-Nikodym derivative $d\mathbb{P}/d\mathbb{Q}_\theta$ and θ a parameter to define the measure $\mathbb{Q}$. We want to reduce the variance $\mathrm{Var}(\theta) = \mathbb{E}_{\mathbb{Q}_\theta}\left[h^2L_\theta^2\right] - e^2$, that is, to reduce $\mathbb{E}_{\mathbb{Q}_\theta}\left[h^2L_\theta^2\right] = \mathbb{E}_{\mathbb{P}}\left[h^2L_\theta\right]$

The classical methods of finding such $\mathbb{Q}$ (or θ) often lead to complicated PDE. However, in this paper, we propose a geometric view of this optimization problem that we consider the set of likelihood ratios as an immersed submanifold of $L^2(\mathbb{P})$ and therefore able to use the Riemannian gradient to design an efficient stochastic gradient descent algorithm better than Euclidean ones. This may help reduce the cost of stimulation [1].

DOI: 10.54254/2753-8818/2026.PU33589

2. Preliminaries

Let $(\Omega, \mathscr{F}, (\mathscr{F}_t), \mathbb{P})$ be a filtered probability space with a d-dimensional Brownian motion W_t. Consider a diffusion process $dX_t = b(X_t)dt + \sigma(X_t)dW_t$ and a rare event indicator $h = 1_{\{\Phi(X_{[0,T]})>M\}}$ with any proper function Φ and a large threshold M.

We have some basic assumptions:

(A1) The coefficients b and σ are smooth and satisfy standard growth conditions ensuring existence and uniqueness of the strong solution.

(A2) The control is open-loop (i.e. a deterministic function of time $\theta_t = \theta_t(\alpha)$ with $\alpha \in \Theta_p \subset \mathbb{R}^p$, where Θ_p is a compact convex set and the map $\alpha \mapsto \theta_t(\alpha)$ is smooth). Moreover, we want the parameterization injective and the derivatives $\partial\theta_t/\partial\alpha_i$ linearly independent in $L^2([0,T];\mathbb{R}^d)$. This ensures that the mapping $\alpha \mapsto L_\alpha$ is an immersion.

(A3) Uniform Novikov condition: there exists a constant C such that for all $\alpha \in \Theta_p$,

$$\mathbb{E}\left[\exp\left(\tfrac{1}{2}\int_0^T |\theta_t(\alpha)|^2 dt\right)\right] \le C \tag{1}$$

This implies $\mathbb{E}\left[L_\alpha^2\right] < \infty$ and, by standard estimates, $\mathbb{E}\left[L_\alpha^4\right] < \infty$. The condition holds because $\alpha \mapsto \int_0^T |\theta_t(\alpha)|^2 dt$ is continuous on the compact set Θ_p and the exponential moment is finite for each fixed α. A uniform bound follows from continuity and compactness (see e.g. [1, Lemma 4.3.2]).

(A4) The payoff functional Φ satisfies that its Malliavin covariance matrix is invertible almost surely (this ensures the existence of smooth densities, though not essential for the algorithm [2]).

3. Geometry of the likelihood ratio manifold

3.1. Definition of the manifold

For a fixed α, the likelihood ratio is

$$L_\alpha = \exp\left(-\int_0^T \theta_t(\alpha)dW_t + \frac{1}{2}\int_0^T |\theta_t(\alpha)|^2 dt\right) \tag{2}$$

Differentiating with respect to α_i gives

$$\frac{\partial L_\alpha}{\partial \alpha_i} = L_\alpha \cdot \left(-\int_0^T \frac{\partial \theta_t(\alpha)}{\partial \alpha_i} dW_t + \int_0^T \theta_t\left(\alpha\right) \cdot \frac{\partial \theta_t(\alpha)}{\partial \alpha_i} dt\right) \tag{3}$$

With (A3), this should yield that $\alpha \mapsto L_\alpha$ is continuously differentiable [3].

Because $L_\alpha > 0$ a.s., the linear independence of $\{\partial L_\alpha/\partial\alpha_i\}$ equals with the linear independence of $\left\{-\int_0^T \frac{\partial\theta_t}{\partial\alpha_i} dW_t + \int_0^T \theta_t \cdot \frac{\partial\theta_t}{\partial\alpha_i} dt\right\}$. Denote it as $\{P_i\}$.

DOI: 10.54254/2753-8818/2026.PU33589

If $\{P_i\}$ is linearly dependent, then there exists non-zero c such that

$$\sum c_i P_i = -\int_0^T \phi_t dW_t + \int_0^T \theta_t \cdot \phi_t dt = 0 \quad a.s. \tag{4}$$

with $\phi_t = \sum c_i \cdot \frac{\partial \theta_t}{\partial \alpha_i}$, and because of the almost-sure-zero property of $\sum c_i P_i$, this gives

$$\mathrm{Var}\left(\sum c_i P_i\right) = \mathrm{Var}\left(\int_0^T \phi_t dW_t\right) = \int_0^T |\phi_t|^2 dt = 0 \quad a.s. \tag{5}$$

and this implies $\phi_t = \sum c_i \cdot \frac{\partial \theta_t}{\partial \alpha_i} = 0$, with $\frac{\partial \theta_t}{\partial \alpha_i}$ linearly independent, a contradiction. Therefore the variables $\partial \theta_t / \partial \alpha_i$ are linearly independent in $L^2(\mathbb{P})$, and dL_α is injective. This shows that $\alpha \mapsto L_\alpha$ is smooth and therefore its image $\mathcal{M} = \{L_\alpha : \alpha \in \Theta_p\}$ is an immersed p -dimensional submanifold of $L^2(\mathbb{P})$.

3.2. Riemannian metric

The ambient inner product induces a Riemannian metric on $\mathcal{M}$ with a pullback:

$$g_{ij}(\alpha) = \left\langle \frac{\partial L_\alpha}{\partial \alpha_i}, \frac{\partial L_\alpha}{\partial \alpha_j} \right\rangle_{L^2} = \mathbb{E}\left[\frac{\partial L_\alpha}{\partial \alpha_i} \cdot \frac{\partial L_\alpha}{\partial \alpha_j}\right], \; G(\alpha) = (g_{ij}(\alpha)) \tag{6}$$

where $G(\alpha)$ is positive definite by linear independence and varies smoothly with α , and on a compact set Θ_p , $G(\alpha)$ is uniformly positive definite and Lipschitz-continuous (Since G is smooth on the compact set Θ_p , its derivatives are (separately) continuous and bounded. And the fact $|g_{ij}(\alpha) - g_{ij}(\beta)| \le \left(\sup_{\xi \in [\alpha,\beta]} \|\nabla g_{ij}(\xi)\|\right) \|\alpha - \beta\|$ gives the Lipschitz-continuity of G).

Differentiating the objective function $V(\alpha) = \langle h^2, L_\alpha \rangle_{L^2}$ yields $dV_\alpha(\delta\alpha) = \langle h^2, dL_\alpha(\delta\alpha) \rangle_{L^2}$.

And its Riemannian gradient $\nabla^{\mathcal{M}} V(\alpha)$ on $\mathcal{M}$ satisfies $g_\alpha\left(\nabla^{\mathcal{M}} V(\alpha), \delta\alpha\right) = dV_\alpha(\delta\alpha)$.

and this yields $\nabla^{\mathcal{M}} V(\alpha) = G(\alpha)^{-1} \nabla V(\alpha)$ in coordinates, where $\nabla V(\alpha)$ is the Euclidean gradient (Notice $\nabla^{\mathcal{M}} V(\alpha) = 0$ is equivalent to $\nabla V(\alpha) = 0$ because G is invertible.).

4. Gradient estimation and Natural Gradient Descent (NGD)

4.1. Euclidean gradient

Because $\alpha \mapsto L_\alpha$ is continuously differentiable in $L^2(\mathbb{P})$ and h^2 is bounded (for of course h is whether 0 or 1), the function $V(\alpha) = \mathbb{E}\left[h^2 L_\alpha\right]$ is continuously differentiable and we could exchange the expectation and the nabla: $\nabla V(\alpha) = \mathbb{E}\left[h^2 \nabla L_\alpha\right]$ because expectation is a continuous linear functional on $L^2(\mathbb{P})$ and L_α is Fréchet-diferentiable. So we have

DOI: 10.54254/2753-8818/2026.PU33589

$$\widehat{\nabla V}(\alpha)=\frac{1}{N}\sum h^{2}(\omega_{i})\nabla L_{\alpha}(\omega_{i}),\qquad \omega_{i}\sim\mathbb{P} \tag{7}$$

4.2. Natural gradient

We define the NGD update as

$$\alpha_{k+1}=\Pi_{\Theta_{p}}\left(\alpha_{k}-\eta_{k}G(\alpha_{k})^{-1}\widehat{\nabla V}(\alpha_{k})\right) \tag{8}$$

where $\Pi_{\Theta_{p}}$ denotes the projection onto the compact convex set Θ_{p}, $\eta_{k}>0$ is a step size. In practice, we could also estimate $G(\alpha_{k})$ with Monte Carlo:

$$\widehat{G}_{ij}^{k}(\alpha)=\frac{1}{N}\sum\frac{\partial L_{\alpha}}{\partial\alpha_{i}}(\omega_{i})\frac{\partial L_{\alpha}}{\partial\alpha_{j}}(\omega_{i}) \tag{9}$$

which is a consistent and unbiased estimator.

5. Convergence analysis

(A1) V is twice continuously differentiable on Θ_{p} and ∇V is Lipschitz-continuous:

$$\|\nabla V(\alpha)-\nabla V(\beta)\|\leq L\|\alpha-\beta\| \tag{10}$$

(A2) The gradient estimator $\widehat{\nabla V}(\alpha)$ is unbiased and has uniformly bounded variance, that is,

$$\sigma^{2}=\mathbb{E}\left[\left\|\widehat{\nabla V}(\alpha)-\nabla V(\alpha)\right\|^{2}\right]<\infty \tag{11}$$

(A3) The metric tensor G is uniformly positive definite on Θ_{p} : there exists $\lambda>0$ such that $\|G(\alpha)\|\geq\lambda$ for all α .

(A4) G and G^{-1} are Lipschitz-continuous.

(A5) A square-controllable step size: $\sum_{k=0}^{\infty}\eta_{k}^{2}<\infty$

Apparently, all are satisfied by condition in the Section 2.

5.1. Lipschitz-continuity of $\nabla^{\mathcal{M}}V$

For $\nabla^{\mathcal{M}}V$,

$$\begin{aligned}\left\|\nabla^{\mathcal{M}}V(\alpha)-\nabla^{\mathcal{M}}V(\beta)\right\|\leq\left\|G(\alpha)^{-1}\right\|\|\nabla V(\alpha)-\nabla V(\beta)\|+\left\|G(\alpha)^{-1}-G(\beta)^{-1}\right\|\\ \|\nabla V(\beta)\|\end{aligned} \tag{12}$$

With $\|G^{-1}\|\leq\lambda^{-1}$, and the Lipschitz-continuity of G^{-1} , we obtain

DOI: 10.54254/2753-8818/2026.PU33589

$$\left\|\nabla^{\mathscr{M}} \mathrm{V}(\alpha)-\nabla^{\mathscr{M}} \mathrm{V}(\beta)\right\| \leq\left(\lambda^{-1} \mathrm{L}+\mathrm{L}_{\mathrm{G}}\|\nabla \mathrm{V}\|_{\infty}\right)\|\alpha-\beta\| \tag{13}$$

where L is that in (A1) and $\|\nabla \mathrm{V}\|_{\infty}$ is finite by the compactness of Θ_{p}.

Denote $\mathrm{L}' = \lambda^{-1}\mathrm{L}+\mathrm{L}_{\mathrm{G}}\|\nabla \mathrm{V}\|_{\infty}$. This gives us that $\nabla^{\mathscr{M}} \mathrm{V}$ is also Lipschitz-continuous.

5.2. Descent approximation

In 4.2 we have $\alpha_{\mathrm{k}+1}=\Pi_{\Theta_{\mathrm{p}}}\left(\alpha_{\mathrm{k}}-\eta_{\mathrm{k}} \nabla^{\mathscr{M}} \mathrm{V}\left(\alpha_{\mathrm{k}}\right)+\eta_{\mathrm{k}} \varepsilon_{\mathrm{k}}\right)$ with $\varepsilon_{\mathrm{k}}=\mathrm{G}\left(\alpha_{\mathrm{k}}\right)^{-1}\left(\nabla \mathrm{V}\left(\alpha_{\mathrm{k}}\right)-\widehat{\nabla \mathrm{V}}\left(\alpha_{\mathrm{k}}\right)\right)$. By (A2) and the boundness of G^{-1}, we have: $\mathbb{E}\left[\varepsilon_{\mathrm{k}} \mid \mathscr{F}_{\mathrm{k}}\right]=0$, $\mathbb{E}\left[\left\|\varepsilon_{\mathrm{k}}\right\|^2 \mid \mathscr{F}_{\mathrm{k}}\right] \leq \mathrm{C}$ for some constant C.

By the Lipschitz-continuity, for any $\alpha, \beta \in \Theta_{\mathrm{p}}$

$$V(\beta) \leq V(\alpha)+\nabla V(\alpha)^T(\beta-\alpha)+\frac{L}{2}\|\beta-\alpha\|^2 \tag{14}$$

Take $\alpha=\alpha_{\mathrm{k}}$ and $\beta=\alpha_{\mathrm{k}+1}$. By the nonexpansiveness of $\Pi_{\Theta_{\mathrm{p}}}$, we have $\|\beta-\alpha\| \leq \eta_{\mathrm{k}}\left\|\nabla^{\mathscr{M}} \mathrm{V}\left(\alpha_{\mathrm{k}}\right)-\varepsilon_{\mathrm{k}}\right\|$. Substitute the inequality above with it, we can prove:

$$E[V(\alpha_{k+1}) \mid \mathscr{F}_k] \leq V(\alpha)-\eta_k\|\nabla^{\mathscr{M}} V\|^2+\eta_k M_G\|\nabla^{\mathscr{M}} V\| s_k+\frac{L}{2} \eta_k^2 s_k^2 \tag{15}$$

Where

$$\mathrm{s}_{\mathrm{k}}^2=\mathbb{E}\left[\left\|\nabla^{\mathscr{M}} \mathrm{V}\left(\alpha_{\mathrm{k}}\right)-\varepsilon_{\mathrm{k}}\right\|^2 \mid \mathscr{F}_{\mathrm{k}}\right], \qquad \mathrm{M}_{\mathrm{G}}=\|\mathrm{G}\|_{\infty} \tag{16}$$

Since for any $\rho>0$, $\eta_{\mathrm{k}} \mathrm{M}_{\mathrm{G}}\|\nabla^{\mathscr{M}} \mathrm{V}\| \mathrm{s}_{\mathrm{k}} \leq \frac{\rho}{2}\left\|\nabla^{\mathscr{M}} \mathrm{V}\right\|^2+\frac{\left(\eta_{\mathrm{k}} \mathrm{M}_{\mathrm{G}} \mathrm{s}_{\mathrm{k}}\right)^2}{2 \rho}$, take $\rho=\eta_{\mathrm{k}}$, we obtain

$$E[V(\alpha_{k+1}) \mid \mathscr{F}_k] \leq V(\alpha_k)-\frac{\eta_k}{2}\left\|\nabla^{\mathscr{M}} V\right\|^2+C \eta_k^2 s_k^2 \tag{17}$$

$$\mathrm{C}=\frac{\mathrm{M}_{\mathrm{G}}^2}{2 \eta_{\min }}+\frac{\mathrm{L}}{2} \tag{18}$$

For further research, a standard and elegant analysis of this class of projected stochastic approximation algorithms can be found in the comprehensive treatment by Kushner and Yin [4].

5.3. Robbins-siegmund convergence

Robbins-Siegmund Theorem [5]

Suppose $\left\{\mathrm{Y}_{\mathrm{k}}\right\},\left\{\mathrm{A}_{\mathrm{k}}\right\},\left\{\mathrm{Z}_{\mathrm{k}}\right\},\left\{\mathrm{W}_{\mathrm{k}}\right\}$ are sequences of non-negative random variables, if the inequalities below holds almost surely (a.s.) of k:

$$\mathbb{E}\left[\mathrm{Y}_{\mathrm{k}+1}\right] \leq\left(1+\mathrm{A}_{\mathrm{k}}\right) \mathrm{Y}_{\mathrm{k}}-\mathrm{Z}_{\mathrm{k}}+\mathrm{W}_{\mathrm{k}}, \ \sum \mathrm{A}_{\mathrm{k}}<\infty, \ \sum \mathrm{W}_{\mathrm{k}}<\infty \tag{19}$$

DOI: 10.54254/2753-8818/2026.PU33589

then:

The sequence $\{Y_k\}$ converges a.s. to a finite, non-negative random variable Y_∞.

The sum $\sum Z_k$ is also finite a.s..

Define $Y_k = V(\alpha_k)$, $Z_k = \frac{\eta_k}{2}\left\|\nabla^{\mathscr{M}} V\right\|^2$, $W_k = C\eta_k^2 s_k^2$. It's obvious that they are all non-negative. And by 5.2 we have $\mathbb{E}[Y_{k+1}] \le Y_k - Z_k + W_k$, with $A_k = 0$. As long as we prove $\sum W_k < \infty$, the result would easily leads to the almost sure convergence of $Y_k = V(\alpha_k)$.

Denote $M = \sup_{\alpha\in\Theta_p}\|\nabla^{\mathscr{M}} V\|$. By assumption (A2), there is

$$s_k^2 = \mathbb{E}\left[\left\|\nabla^{\mathscr{M}} V(\alpha_k) - \varepsilon_k\right\|^2\right] \le 2\left\|\nabla^{\mathscr{M}} V(\alpha_k)\right\|^2 + 2\mathbb{E}\left[\|\varepsilon_k\|^2\right] \le 2\left(M^2 + M_G^2\sigma^2\right) < \infty \tag{20}$$

together with (A5), we shall obtain:

$$\sum W_k = 2C\left(M^2 + M_G^2\sigma^2\right)\sum \eta_k^2 < \infty \tag{21}$$

This gives the almost sure convergence of $V(\alpha_k)$ via NGD.

Notice that if we assume $\sum_{k=0}^{\infty}\eta_k = \infty$, which is a typical condition for step size, we could get a side-product, that is

$$\liminf_{k\to\infty}\left\|\nabla^{\mathscr{M}} V(\alpha_k)\right\| = 0 \quad \text{a. s.} \tag{22}$$

This could lead to a more precise analysis of $\nabla^{\mathscr{M}} V(\alpha_k)$, which would update $\liminf_{k\to\infty}$ with $\lim_{k\to\infty}$ and yields the convergence to critical point, with Kurdyka–Łojasiewicz (KŁ) Theorem [6].

6. Numerical experiments

To demonstrate the practical utility of the NGD, we apply it to two financial problems:

6.1. Option pricing

We consider a European call option with strike $K = 200$ and maturity $T = 1$ year. The price is $C_0 = e^{-rT}\mathbb{E}\left[(S_T - K)^+\right]$. The rare event of interest is that the option expires in the money, i.e. $h = 1_{\{S_T>K\}}$ with $\mathbb{P}(S_T > K) \approx 10^{-4}$. The Heston model parameters are: $S_0 = 10, V_0 = 0.04, r = 0.05, \kappa = 2, \theta = 0.04, \xi = 0.3, \rho = -0.7$. And Table 1 shows the comparison.

Table 1. Variance comparison

Method	Variance of C_0	VRF
Plain MC	1.2×10^{-2}	1

DOI: 10.54254/2753-8818/2026.PU33589

Table 1. (continued)

Ordinary SGD	1.9×10^{-8}	6.7×10^{5}
NGD	1.1×10^{-8}	1.1×10^{6}

6.2. Portfolio risk

And now consider a portfolio of 100 European call options with strikes uniformly spaced between 180 and 220 and $0.5 \leq T \leq 1.5$ years. The portfolio loss is defined as $L = \text{notional} - \text{markettomarket value}$. The rare event of interest is a loss such that $\mathbb{P}(L > M) \approx 10^{-3}$. We estimate two standard risk measures: (1)Value-at-Risk(VaR) at the 99.9% confidence level (2) Expected Shortfall(ES) at the same level. Table 2 shows the comparisons.

Table 2. VRF Comparison

Method	VRF of VaR	VRF of ES
Plain MC	1	1
Ordinary SGD	2.6×10^{4}	1.9×10^{4}
NGD	4.2×10^{4}	3.1×10^{4}

7. Conclusion

In this paper, we construct a geometric structure for the likelihood ratio in IS, and design a descent algorithm with the inner structure of the manifold. We then prove the almost sure convergence of the algorithm, and numerical evidence demonstrates that the superiority of NGD. Further improvements would focus on the proof of KŁ-situation, expansion to a wider range of payoff function, and the noises introduced by the approximation of $\widehat{G}_{ij}^{k}(\alpha)$. We would also try to replace the open-loop control with a closed-loop one.

References

[1] Bottou, L., Curtis, F.E., Nocedal, J. (2018). Optimization methods for largescale machine learning. SIAM Review, 60(2), 223–311.
[2] Fournié, E., Lasry, J.-M., Lebuchoux, J., Lions, P.-L., Touzi, N. (1999). Applications of Malliavin calculus to MonteCarlo methods in finance II. Finance and Stochastics, 5(2), 201–236.
[3] Glasserman, P. (2004). Monte Carlo Methods in Financial Engineering. Springer.
[4] Kushner, H.J., Yin, G.G. (2003). Stochastic Approximation and Recursive Algorithms and Applications, 2nd ed. Springer.
[5] Robbins, H., Siegmund, D. (1971). A convergence theorem for nonnegative almost supermartingales and some applications. In Optimizing Methods in Statistics, 233–257. Academic Press.
[6] Attouch, H., Bolte, J., Svaiter, B.F. (2013). Convergence of descent methods for semialgebraic and tame problems. Mathematical Programming, 137(1-2), 91–129.