A framework for adaptive Monte-Carlo procedures

Jérôme Lelong (with B. Lapeyre)

http://www-ljk.imag.fr/membres/Jerome.Lelong/

Journées MAS – Bordeaux

Friday 3 September 2010

Outline

Monte-Carlo framework

- Compute $\mathbb{E}(Z)$ using a Monte-Carlo method
- Suppose we know that

$$\mathbb{E}(Z) = \mathbb{E}\left[H(\theta, X)\right], \theta \in \mathbb{R}^d$$

- Aim: use the previous representation to reduce the variance.
- Find a way to minimize the variance.

$$v(\theta) = \mathrm{Var}(H(\theta, X)) = \mathbb{E}\left[H(\theta, X)^2\right] - \mathbb{E}[Z]^2$$

- $\exists \theta^\star$ s.t. $\forall \theta \, v(\theta) \geq v(\theta^\star)$.

The algorithms I

Algorithm (Non adaptive importance sampling, Fu and Su (2002), Arouna (2003))

1. *Draw* $(X_1, \ldots, X_n)$ *to compute* θ_n *an estimator of* $\theta^\star$.
2. *Draw* $(X'_1, \ldots, X'_n)$ *independent of* $(X_1, \ldots, X_n)$ *and compute*

$$\bar{\xi}_n = \frac{1}{n} \sum_{i=1}^{n} H(\theta_n, X'_i).$$

The algorithms II

Algorithm (Adaptive Importance Sampling, Arouna (2004))

Draw $(X_1, \dots, X_n)$ *and do for* $i = 1 : n$

1. *Compute an estimator* θ_i *of* $\theta^\star$ *using* $(X_1, \dots, X_i)$
2. *Update* ξ_i

$$\xi_{i+1} = \frac{i}{i+1}\xi_i + \frac{1}{i+1}H(\theta_i, X_{i+1}), \quad \textit{with } \xi_0 = 0.$$

Remarks:

- No need to store the whole sequence $(X_1, \dots, X_n)$ for computing ξ_n
- $\xi_n = \frac{1}{n}\sum_{i=1}^{n} H(\theta_{i-1}, X_i)$

Common frameworks I

- **Importance sampling framework in a Gaussian setting**
  If $G \sim \mathcal{N}(0, I_d)$. For all $\theta \in \mathbb{R}^d$,

$$\mathbb{E}\left[f(G)\right] = \mathbb{E}\left[e^{-\theta\cdot G - \frac{|\theta|^2}{2}} f(G+\theta)\right],$$

$$v(\theta) = \mathbb{E}\left[e^{-\theta\cdot G + \frac{|\theta|^2}{2}} f^2(G)\right] - \mathbb{E}[f(G)]^2.$$

  v is strongly convex if $\exists \varepsilon > 0$ s.t. $\mathbb{E}[|f(G)|^{2+\varepsilon}] > \infty$.
  Arouna (2004 and 2005), Lemaire and Pagès (2008), Jourdain and L. (2009)

Common frameworks II

- **Escher transform**
  Let X be a r.v. in $\mathbb{R}^d$ with density p and $\theta \in \mathbb{R}^d$.

$$p_\theta(x) = p(x)\,\mathrm{e}^{\theta\cdot x - \psi(\theta)} \quad \text{with } \psi(\theta) = \log \mathbb{E}\left[\,\mathrm{e}^{\theta\cdot X}\right].$$

Let $X^{(\theta)}$ have p_θ as a density, then

$$\mathbb{E}[f(X)] = \mathbb{E}\left[f(X^{(\theta)})\frac{p(X^{(\theta)})}{p_\theta(X^{(\theta)})}\right],$$
$$v(\theta) = \mathbb{E}\left[f(X)^2\frac{p(X)}{p_\theta(X)}\right] - \mathbb{E}[f(X)]^2.$$

v is strongly convex if $\exists \varepsilon > 0$ s.t. $\mathbb{E}[|f(G)|^{2+\varepsilon}] > \infty$ and $\lim_{|\theta|\to\infty} p_\theta(x) = 0$ for all x.
Kawai (2007 and 2008), Lemaire and Pagès (2008).

An adaptive strong law

Assume $\mathbb{E}[Z] = \mathbb{E}[H(\theta, X)]$ for all θ and $v(\theta) = \mathrm{Var}(H(\theta, X))$ is strongly convex. Let $(X_n, n \geq 0)$ be i.i.d $\sim X$. $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$.

Theorem 1

Let $(\theta_n)_{n\geq 0}$ be a $(\mathcal{F}_n)$−adapted sequence with values in $\mathbb{R}^d$, s. t. for all $n \geq 0$, $\theta < \infty$ p.s.

(H1) For any compact subset $K \subset \mathbb{R}^d$, $\sup_{\theta \in K} \mathbb{E}[|H(\theta, X)|^2] < \infty$

(H2) $\inf_{\theta \in \mathbb{R}^d} v(\theta) > 0$ *and* $\frac{1}{n} \sum_{i=0}^{n} v(\theta_i) < \infty$

Then,

$$\xi_n = \frac{1}{n} \sum_{i=1}^{n} H(\theta_{i-1}, X_i) \xrightarrow[n \to \infty]{a.s} \mathbb{E}(Z).$$

Remarks on the assumptions

- (H1) : No need of $\mathbb{E}\left[\sup_{\theta \in K} |H(\theta, X)|^2\right] < \infty$ by the use of locally square integrable martingales. If $\theta \longmapsto \mathbb{E}[|H(\theta, X)|^2]$ is continuous, (H1) is equivalent to $\mathbb{E}[|H(\theta, X)|^2] < \infty$ for all $\theta \in \mathbb{R}^d$.
- (H2) : is clearly true when θ_n converges to a deterministic constant θ_∞ and v is continuous at θ_∞.
- No assumption to be checked along the path $(\theta_n)_n$.

A Central Limit Theorem

Theorem 2

Let the sequence $(\theta_n, n \geq 0)$ *be adapted to* $\mathcal{F}_n$. *Assume (H1), (H2) and*

- $\theta_n \longrightarrow \theta^\star$ *p.s.*
- $\exists \eta > 0$ *s.t.* $\theta \longmapsto \mathbb{E}[|H(\theta, X)|^{2+\eta}]$ *is continuous at* $\theta^\star$ *and finite* $\forall \theta \in \mathbb{R}^d$.
- v *is continuous at* $\theta^\star$ *and* $v(\theta^\star) > 0$

Then,

$$\sqrt{n}\left(\frac{1}{n}\xi_n - \mathbb{E}[Z]\right) \xrightarrow[n\to\infty]{D} \mathcal{N}(0, v(\theta^\star)).$$

Moreover, assume that

- $\exists \eta > 0$ *s.t.* $\theta \longmapsto \mathbb{E}[|H(\theta, X)|^{4+\eta}]$ *is continuous at* $\theta^\star$ *and finite* $\forall \theta \in \mathbb{R}^d$.

Then,

$$\frac{\sqrt{n}}{\sigma_n}(\xi_n - \mathbb{E}[Z]) \xrightarrow[n\to\infty]{D} \mathcal{N}(0,1) \quad \textit{with} \quad \sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} H(\theta_{i-1}, X_i)^2 - \xi_n^2$$

Variance minimisation

- v is strongly convex
- v is infinitely differentiable and

$$\nabla_\theta v(\theta) = \nabla_\theta \mathbb{E}\left[(H(\theta,X)^2)\right] = \mathbb{E}\left[U(\theta,X)\right]$$

- Minimizing v is equivalent to finding $\theta^\star$ s.t.

$$\mathbb{E}\left[U(\theta^\star,X)\right] = 0$$

Randomly truncated procedure

- Let $(\gamma_n)_n \geq 0$ s.t. $\sum_n \gamma_n = +\infty$ and $\sum_n \gamma_n^2 < +\infty$.
  For $\theta_0 \in K_0$ and $\alpha_0 = 0$, we define

$$\begin{cases} & \theta_{n+\frac{1}{2}} = \theta_n - \gamma_{n+1} U(\theta_n, X_{n+1}), \\ \text{if } \theta_{n+\frac{1}{2}} \in K_{\alpha_n} & \theta_{n+1} = \theta_{n+\frac{1}{2}} \qquad \alpha_{n+1} = \alpha_n, \\ \text{if } \theta_{n+\frac{1}{2}} \notin K_{\alpha_n} & \theta_{n+1} = \theta_0 \qquad \alpha_{n+1} = \alpha_n + 1. \end{cases}$$

  α_n = number of truncations up to time n.

$$\theta_{n+1} = \mathcal{T}_{K_{\alpha_n}}\left(\theta_n - \gamma_{n+1} U(\theta_n, X_{n+1})\right)$$

- θ_n is $\mathcal{F}_n$−measurable and X_{n+1} is independent of $\mathcal{F}_n$.
- Introduced by Chen and Zhu (1986).

a.s convergence

Let $u(\theta) = \mathbb{E}[U(\theta, X)]$.

Theorem 3 (L., 2008)

Assume

- *(H3) u is continuous and* $\exists! \theta^\star \in \mathbb{R}^d, u(\theta^\star) = 0$ *and* $\forall \theta \in \mathbb{R}^d, \theta \neq \theta^\star, (\theta - \theta^\star | u(\theta)) > 0$.
- *For all* $q > 0$, $\sup_{|\theta| \leq q} \mathbb{E}[|U(\theta, X)|^2] < \infty$.

Then, the sequence $(\theta_n)_n$ *converges a.s. to* $\theta^\star$ *and the sequence* $(\alpha_n)_n$ *is a.s. finite.*

- Previous results from Chen and Zhu (1986), and Chen, Gao and Guo (1988).
- (H3) satisfied if U is the gradient of a strictly convex function

Moving window average

Assume $\gamma_n = \frac{\gamma}{(n+1)^\alpha}$ with $1/2 < \alpha < 1$.
For all $\tau > 0$, we set

$$\hat{\theta}_n(\tau) = \frac{\gamma_p}{\tau} \sum_{i=p}^{p+\lfloor \tau/\gamma_p \rfloor} \theta_i \quad \text{with } p = \sup\{k \geq 1 : k + \tau/\gamma_k \leq n\} \wedge n.$$

- Averaging smooths the convergence.
- Averaging from a strictly positive rank reduces the impact of the initial condition.
- If $(\theta_n)_n$ converges, so does $(\hat{\theta}_n(\tau))_n$ for all $\tau > 0$.
- True Césaro averaging : Polyak and Juditsky (1992), Pelletier (2000), Andrieu and Moulines (2006).

General problem I

- Generalised multidimensional Black-Scholes Model

$$dS_t = S_t(\mu(t, S_t)dt + \sigma(t, S_t) \cdot dW_t), \; S_0 = x$$

- Payoff $\hat{\psi}(S_t, t \leq T)$, price

$$V_0 = \mathbb{E}[e^{-rT}\hat{\psi}(S_t, t \leq T)]$$

- Can be approximated by

$$\hat{V}_0 = \mathbb{E}[e^{-rT}\hat{\psi}(S_{T_1}, \ldots, S_{T_d})]$$

With $G \sim \mathcal{N}(0, I_d)$

$$\hat{V}_0 = \mathbb{E}[\psi(G)] = \mathbb{E}\left[\psi(G + A\theta)e^{-A\theta\cdot G - \frac{|A\theta|^2}{2}}\right]$$

with $\theta \in \mathbb{R}^p$ and $A \in \mathbb{R}^{d\times p}$, $p << d$.

General Problem II

- Minimise $v(\theta) = \mathbb{E}\left[\psi(G+A\theta)^2 e^{-2A\theta\cdot G-|A\theta|^2}\right] = \mathbb{E}\left[\psi(G)^2 e^{-A\theta\cdot G+\frac{|A\theta|^2}{2}}\right]$.

$$\nabla v(\theta) = \mathbb{E}\left[A^*(A\theta - G)\psi(G)^2 e^{-A\theta\cdot G+\frac{|A\theta|^2}{2}}\right] \qquad U^1(\theta, G) = A^*(A\theta - G)\phi(G)^2 e^{-A\theta\cdot G+\frac{|A\theta|^2}{2}}$$

$$\nabla v(\theta) = \mathbb{E}\left[-A^* G\psi(G+A\theta)^2 e^{-2A\theta\cdot G+A\theta^2}\right] \qquad U^2(\theta, G) = -A^* G\phi(G+A\theta)^2 e^{-2A\theta\cdot G+|A\theta|^2}$$

- we can write $\nabla v(\theta) = \mathbb{E}[U^2(\theta, G)] = \mathbb{E}[U^1(\theta, G)]$ to construct two estimators of $\theta^\star$: $(\theta_n^1)_n$ and $(\theta_n^2)_n$

Bespoke estimators

We define

$$\theta^1_{n+1} = \mathcal{T}_{K_{\alpha_n}}\left(\theta^1_n - \gamma_{n+1} U^1(\theta^1_n, G_{n+1})\right) \quad \textit{involves } \phi(G)$$
$$\theta^2_{n+1} = \mathcal{T}_{K_{\alpha_n}}\left(\theta^2_n - \gamma_{n+1} U^2(\theta^2_n, G_{n+1})\right) \quad \textit{involves } \phi(G + A\theta^2_n)$$

and their averaging versions $(\widehat{\theta^1}_n)_n$ and $(\widehat{\theta^2}_n)_n$.
For the different estimators of $\theta^\star$, we can define as many approximations of $\mathbb{E}(\phi(G))$

$$\xi^1_n = \frac{1}{n}\sum_{i=1}^n H(\theta^1_{i-1}, G_i), \quad \xi^2_n = \frac{1}{n}\sum_{i=1}^n H(\theta^2_{i-1}, G_i),$$
$$\widehat{\xi^1}_n = \frac{1}{n}\sum_{i=1}^n H(\widehat{\theta^1}_{i-1}, G_i), \quad \widehat{\xi^2}_n = \frac{1}{n}\sum_{i=1}^n H(\widehat{\theta^2}_{i-1}, G_i),$$

with $H(\theta, G) = \phi(G + A\theta) e^{-A\theta\cdot G - \frac{|A\theta|^2}{2}}$ and where the sequence $(G_n)_n$ has already been used to build the $(\theta_n)_n$ estimators.

Complexity of the different estimators

We assume : complexity ≈ number of evaluations of ϕ.

- Non-adaptive algorithms : we need $2n$ samples to achieve a convergence rate $\sqrt{v(\theta^\star)/n}$.
  Complexity: $2n$, efficient only when $v(\theta^\star) \leq v(0)/2$.
- Adaptive algorithms : we need n samples to achieve a convergence rate $\sqrt{v(\theta^\star)/n}$.

Estimators	ξ^1	ξ^2	$\widehat{\xi^1}$	$\widehat{\xi^2}$
Complexity	$2n$	n	$2n$	$2n$

FIG.: Complexities of the different estimators

Basket Option : $(\sum_{i=1}^{d} \omega^i S_T^i - K)_+$ where $(\omega^1, \ldots, \omega^d) \in \mathbb{R}^d$

ρ	K	γ	Price	Var MC	Var ξ^2	Var $\widehat{\xi^2}$
0.1	45	1	7.21	12.24	1.59	1.10
	55	10	0.56	1.83	0.19	0.14
0.2	50	0.1	3.29	13.53	1.82	1.76
0.5	45	0.1	7.65	43.25	6.25	4.97
	55	0.1	1.90	14.74	1.91	1.4
0.9	45	0.1	8.24	69.47	10.20	7.78
	55	0.1	2.82	30.87	2.7	2.6

TAB.: Basket option in dimension $d = 40$ with $r = 0.05$, $T = 1$, $S_0^i = 50$, $\sigma^i = 0.2$, $\omega^i = \frac{1}{d}$ for all $i = 1, \ldots, d$ and $n = 100\,000$.

Estimators	MC	ξ^2	$\widehat{\xi^2}$
CPU time	0.85	0.9	1.64

TAB.: CPU times for the option of Table 1.

Barrier Basket option: $(\sum_{i=1}^{D} \omega^i S_T^i - K)_+ \mathbf{1}_{\{\forall i \le D,\, \forall j \le N,\, S_{t_j}^i \ge L^i\}}$

$$\begin{pmatrix} B_{t_1} \\ B_{t_2} \\ \vdots \\ B_{t_{N-1}} \\ B_{t_N} \end{pmatrix} = \begin{pmatrix} \sqrt{t_1} I_D & 0 & 0 & \dots & 0 \\ \sqrt{t_1} I_D & \sqrt{t_2 - t_1} I_D & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sqrt{t_{N-1} - t_{N-2}} I_D & 0 \\ \sqrt{t_1} & \sqrt{t_2 - t_1} I_D & \dots & \sqrt{t_{N-1} - t_{N-2}} I_D & \sqrt{t_N - t_{N-1}} I_D \end{pmatrix} G$$

where I_D is the identity matrix in dimension D.

$$A = \begin{pmatrix} \sqrt{t_1} I_D \\ \sqrt{t_2 - t_1} I_D \\ \vdots \\ \sqrt{t_N - t_{N-1}} I_D \end{pmatrix}$$

$G + A\theta$ corresponds to $(B_{t_1} + \theta t_1, B_{t_2} + \theta t_2, \dots, B_{t_N} + \theta t_N)^*$.
$\theta \in \mathbb{R}^5$ whereas $G \in \mathbb{R}^{120}$.

Barrier Basket option

K	γ	Price	Var MC	Var ξ^2	Var $\widehat{\xi^2}$	Var $\widehat{\theta^2}$+MC	Var ξ^2 reduced	Var $\widehat{\xi^2}$ reduced	Var $\widehat{\theta^2}$+MC reduced
45	0.5	2.37	22.46	4.92	3.52	2.59	2.64	2.62	2.60
50	1	1.18	10.97	1.51	1.30	0.79	0.80	0.80	0.79
55	1	0.52	4.85	0.39	0.38	0.19	0.24	0.23	0.19

TAB.: Down and Out Call option in dimension $I = 5$ with $\sigma = 0.2$, $S_0 = (50, 40, 60, 30, 20)$, $L = (40, 30, 45, 20, 10)$, $\rho = 0.3$, $r = 0.05$, $T = 2$, $\omega = (0.2, 0.2, 0.2, 0.2, 0.2)$ and $n = 100\,000$.

Estimators	MC	ξ^2	$\widehat{\xi^2}$	θ^2 + MC	ξ^2 reduced	$\widehat{\xi^2}$ reduced	$\widehat{\theta^2}$ + MC reduced
CPU time	1.86	1.93	3.34	4.06	1.89	2.89	3.90

TAB.: CPU times for the option of Table 3.

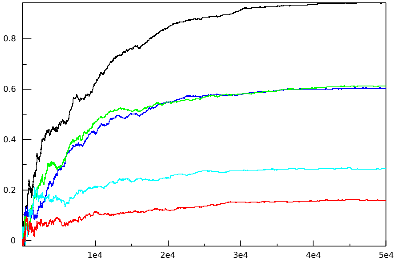

FIG.: approximation of $\theta^\star$ with averaging

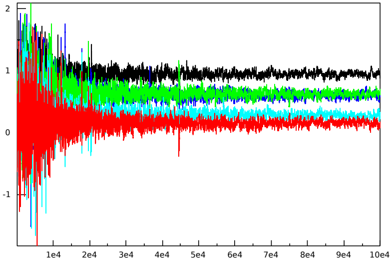

FIG.: approximation of $\theta^\star$ without averaging

Conclusion

- It always reduces the variance
- The extra computational cost can be negligible
- No regularity assumptions on the payoff
- Averaging improves the robustness of the algorithm w.r.t the step sequence but adds an extra computational cost
- To encounter the fine tuning of the algorithm, one can use sample average approximation (Jourdain and L., 2008), but it cannot be implemented in an adaptive manner which increases its computational cost