Fluctuation Analysis of Adaptive Multilevel Splitting

Arnaud Guyader (UPMC - Paris VI)

Joint work with Frédéric Cérou (INRIA Rennes)

International Conference on Monte Carlo Techniques
July 7, 2016, Paris

Framework

- $X \sim \eta$, a random vector in $\mathbb{R}^d$ (typically: $\eta(dx) \propto f(x)dx$).
- $S : \mathbb{R}^d \to \mathbb{R}$ is called the score function (black-box).
- Hence, one can only simulate the random variable $Y = S(X)$.
- The quantile q lies far out in the tail of the pdf of Y.

$\Rightarrow$ Goal: estimate $p = \mathbb{P}(S(X) > q) = \mathbb{P}(Y > q) \approx 0$.

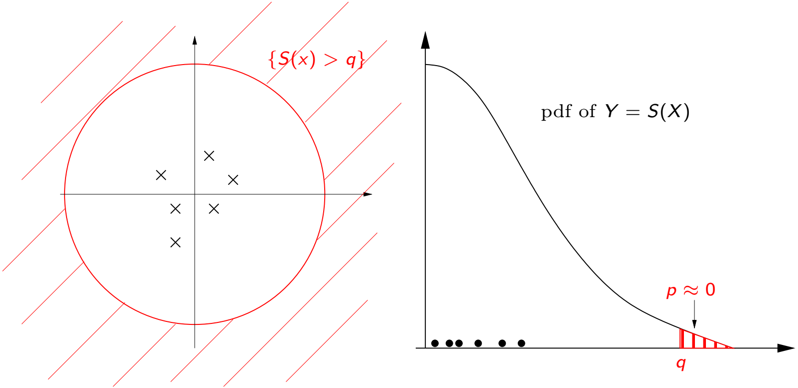
Remarks:

1. $p \approx 0 \Rightarrow$ Crude Monte Carlo is computationally intractable.

$$\frac{\mathrm{Var}(\hat{p}_{mc})}{p^2} = \frac{\mathrm{Var}(\#\{i : \ S(X_i) > q\}/N)}{p^2} = \frac{1-p}{Np} \approx \frac{1}{Np}.$$

2. Assuming S acts as a black-box $\Rightarrow$ no Importance Sampling.

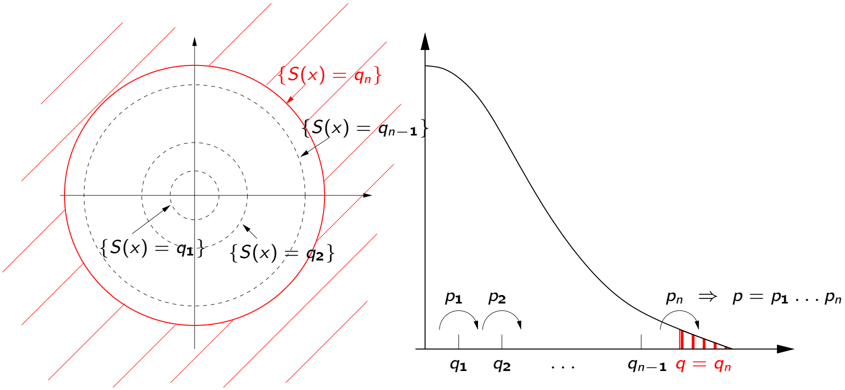
Toy Example



- Random vector: X is such that $\mathbb{E}[X] = 0$.
- Score function: $Y = S(X) = \|X\|$ is the Euclidean norm of X.
- Aim: Estimate $p = \mathbb{P}(S(X) > q)$, where $q \gg \mathbb{E}\|X\| = \mathbb{E}[Y]$.

Multilevel Splitting: Basic Idea

- Fix n and a sequence of levels $-\infty = q_0 < q_1 < \cdots < q_n = q$.
- Notation: let $p_j = \mathbb{P}(Y > q_j | Y > q_{j-1})$ s.t. each $p_j \not\approx 0$.
- Bayes formula: $p = \mathbb{P}(Y > q) = p_1 \times p_2 \times \cdots \times p_n$.
- Multilevel Splitting estimator: $\hat{p} = \hat{p}_1 \times \hat{p}_2 \times \cdots \times \hat{p}_n$.



Implementation: First step

- Ingredient: an η-symmetric transition kernel K_1, *i.e.*

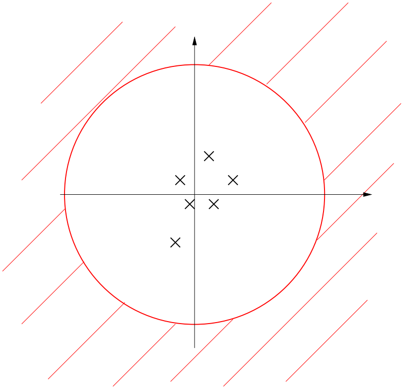
$$\eta(dx)K_1(x,dx') = \eta(dx')K_1(x',dx).$$

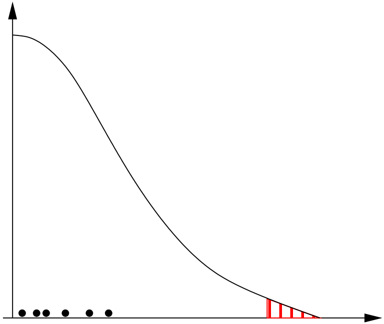
- Example: if $X \sim \mathcal{N}(0, I_d)$, then $X' = \frac{X+\sigma_1 W}{\sqrt{1+\sigma_1^2}}$ makes the job for any $\sigma_1 > 0$ as long as $W \sim \mathcal{N}(0, I_d)$ and $W \perp X$.
- Remark: if no obvious K_1, but $\eta(dx) \propto f(x)dx$, then one may apply Metropolis-Hastings algorithm.

Letting $\mathcal{A}_1 = \{x \in \mathbb{R}^d,\ S(x) > q_1\}$, apply (iterate) the kernel

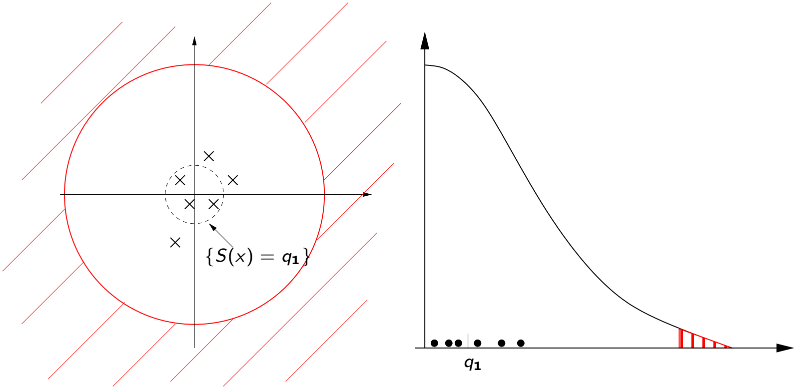
$$M_1(x,dx') = K_1(x,dx')\mathbb{1}_{\mathcal{A}_1}(x') + K_1(x,\bar{\mathcal{A}}_1)\delta_x(dx').$$

Illustration

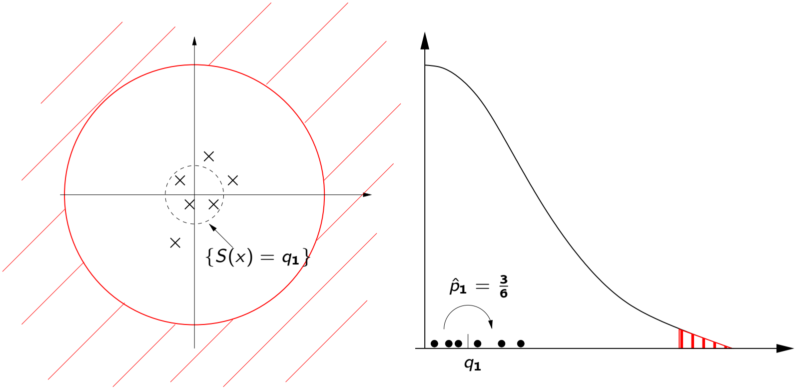




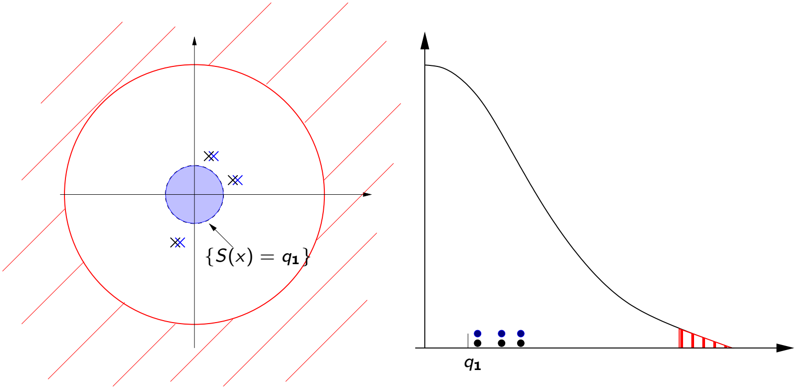
Illustration



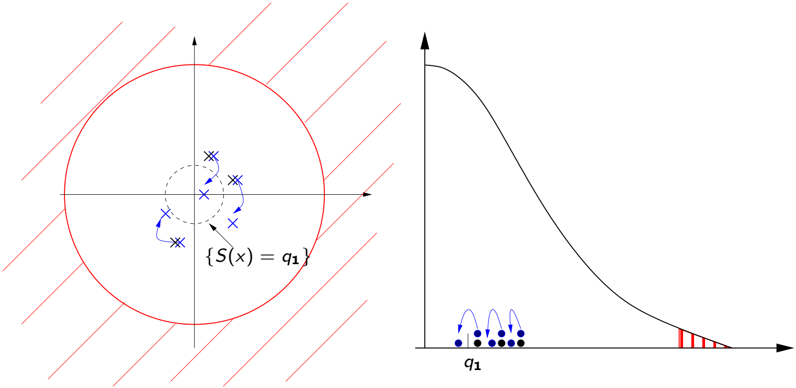
Illustration



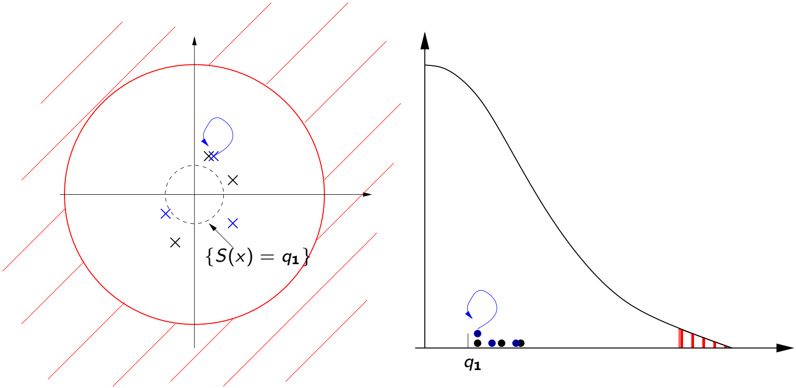
Illustration



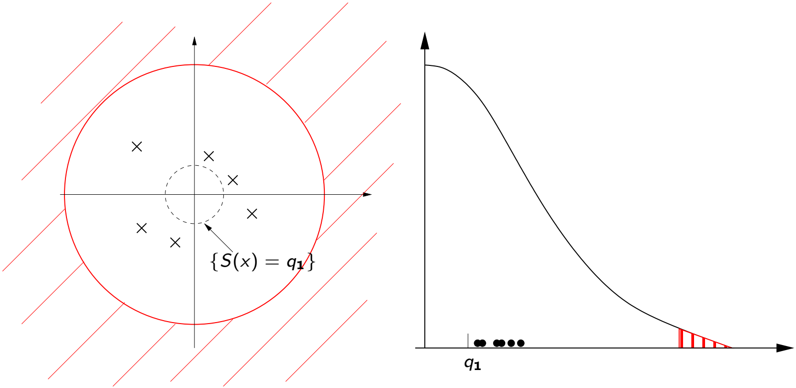
Illustration



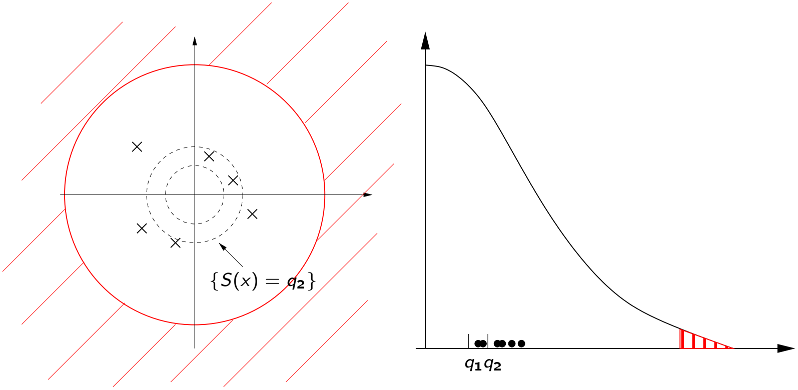
Illustration



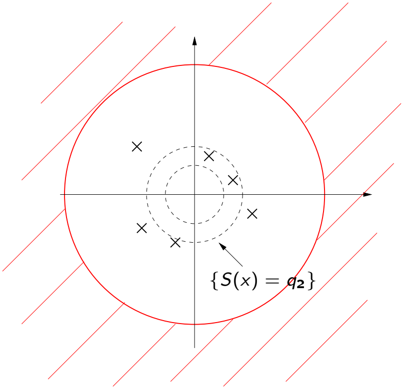
Illustration

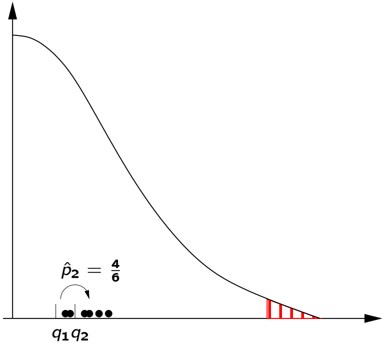


Illustration

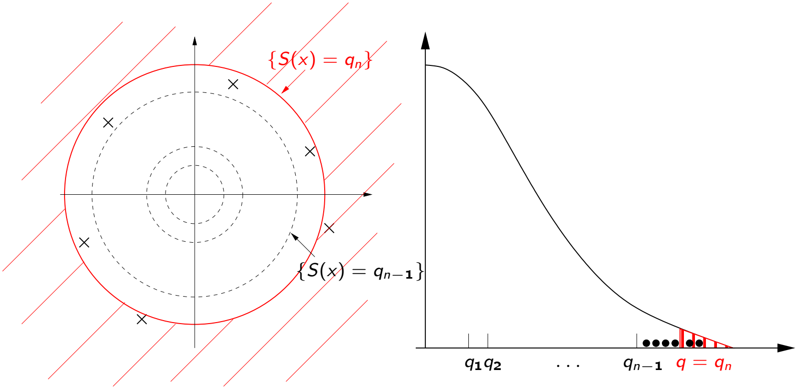


Illustration

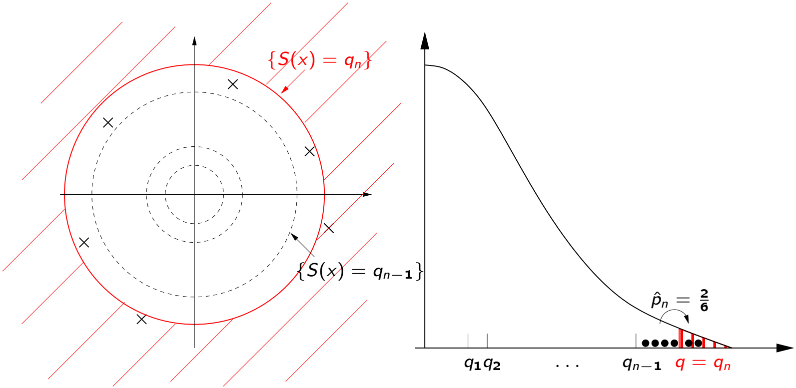




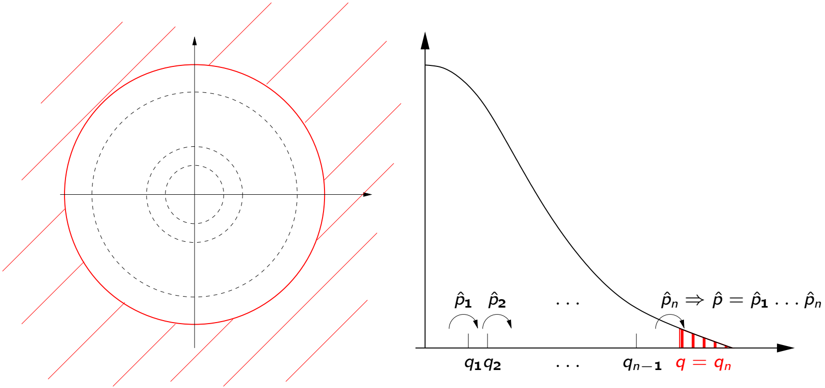
Illustration



Illustration



Illustration



Implementation: Step j

Ingredient: an η-symmetric transition kernel K_j.

- Example: if $X \sim \mathcal{N}(0, I_d)$, then $X' = \frac{X+\sigma_j W}{\sqrt{1+\sigma_j^2}}$.
- Step j: letting $\mathcal{A}_j = \{x, S(x) > q_j\}$, apply (iterate) the kernel

$$M_j(x, dx') = K_j(x, dx')\mathbb{1}_{\mathcal{A}_j}(x') + K_j(x, \bar{\mathcal{A}}_j)\delta_x(dx').$$

$\Rightarrow$ The law $\eta_j(dx) = \eta(dx)\mathbb{1}_{\mathcal{A}_j}(x)/\eta(\mathcal{A}_j)$ is the restriction of η "above" q_j and satisfies $\eta_j M_j = \eta_j$.

Remark: Importance of the tuning parameter σ_j

- σ_j too large $\Rightarrow$ most proposed transitions are refused.
- σ_j too small $\Rightarrow$ the particles move slowly.

Connection with Feynman-Kac Formulas

- Potential functions: $\mathbb{1}_{S(x)>q_j} = \mathbb{1}_{\mathcal{A}_j}(x) = G_{j-1}(x)$.
- Markov chain: let $(X_j)_{j\geq 0}$ a non-homogeneous Markov chain with initial distribution $\eta_0 = \eta$ and transitions M_{j+1}.

$\Rightarrow$ Unnormalized measures:

$$\gamma_n(\varphi) = \mathbb{E}\left[\varphi(X_n)\prod_{j=0}^{n-1} G_j(X_j)\right] = \mathbb{E}\left[\varphi(X)\mathbb{1}_{S(X)>q}\right],$$

Thus, $\gamma_n(1) = \mathbb{E}[\mathbb{1}_{S(X)>q}] = p$ is our quantity of interest.

$\Rightarrow$ Normalized measures:

$$\eta_n(\varphi) = \frac{\gamma_n(\varphi)}{\gamma_n(1)} = \mathbb{E}[\varphi(X)|S(X)>q].$$

Interacting Particle System

- Markov chain $(X_j^1, \dots, X_j^N)_{0 \leq j \leq n}$ with initial distribution $\eta_0^{\otimes N}$ and transitions described by the previous algorithm.
- $\hat{p}_{j+1} = \eta_j^N(G_j)$ is the proportion of the sample $(X_j^1, \dots, X_j^N)$ "above" q_{j+1}, *i.e.*, such that $S(X_j^i) \geq q_{j+1}$.

$\Rightarrow$ Empirical normalized measures

$$\eta_n^N(\varphi) = \frac{1}{N} \sum_{i=1}^{N} \varphi(X_n^i) \xrightarrow[N \to \infty]{a.s.} \eta_n(\varphi) = \mathbb{E}[\varphi(X) | S(X) > q]$$

$\Rightarrow$ Empirical unnormalized measures

$$\begin{cases} \gamma_n^N(1) = \prod_{j=0}^{n-1} \eta_j^N(G_j) = \hat{p}_1 \dots \hat{p}_n = \hat{p} \xrightarrow[N \to \infty]{a.s.} \gamma_n(1) = p \\ \\ \gamma_n^N(\varphi) = \gamma_n^N(1) \times \eta_n^N(\varphi) \xrightarrow[N \to \infty]{a.s.} \gamma_n(\varphi) = \mathbb{E}[\varphi(X) \mathbb{1}_{S(X) > q}] \end{cases}$$

Fluctuation Analysis [Del Moral & Jacod (2001)]

We introduce the family of operators $Q_{\ell,n}$ defined by

$$Q_{\ell,n}(\varphi)(x) = \mathbb{E}\left[\varphi(X_n)\prod_{j=\ell}^{n-1} G_j(X_j)\middle| X_\ell = x\right]$$

and their normalized versions $\overline{Q}_{\ell,n}(\varphi) = Q_{\ell,n}(\varphi)/\eta_\ell(Q_{\ell,n}(1))$.

Theorem

$$\sqrt{N}\left(\gamma_n^N(\varphi) - \gamma_n(\varphi)\right) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}(0, p^2\Sigma(\varphi)),$$

where $\Sigma(\varphi) = \sum_{\ell=0}^{n} \eta_\ell\left(\overline{Q}_{\ell,n}(\varphi)^2 - \eta_n(\varphi)^2\right)$. *Moreover,*

$$\sqrt{N}\left(\eta_n^N(\varphi) - \eta_n(\varphi)\right) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}(0, \Sigma(\varphi)).$$

Back to the Probability Estimate

Corollary

The estimator $\hat{p} = \hat{p}_1 \dots \hat{p}_n$ converges a.s. to p, and we have

$$\sqrt{N}\,\frac{\hat{p}-p}{p} \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}(0,\sigma^2),$$

where

$$\sigma^2 = \sum_{j=1}^{n} \frac{1-p_j}{p_j} + \sum_{j=1}^{n-1} \frac{1}{p_j}\mathbb{E}\left[\left(\frac{\mathbb{P}(X_{n-1}\in\mathcal{A}_n|X_{j-1})}{\mathbb{P}(X_{n-1}\in\mathcal{A}_n|X_{j-1}\in\mathcal{A}_j)}-1\right)^2\middle|X_{j-1}\in\mathcal{A}_j\right]$$

What is the best thing to do?

- Consequence: one has

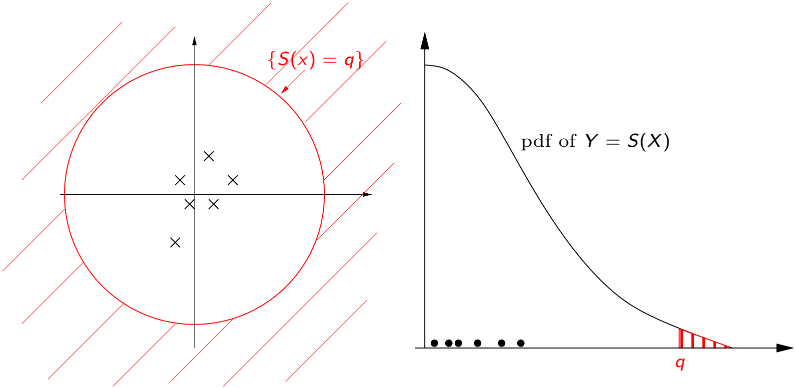
$$\sigma^2 \geq \sum_{j=1}^{n} \frac{1-p_j}{p_j} \quad \text{with} = \text{if the } X_j\text{'s are } \perp$$

- Constrained Minimization:

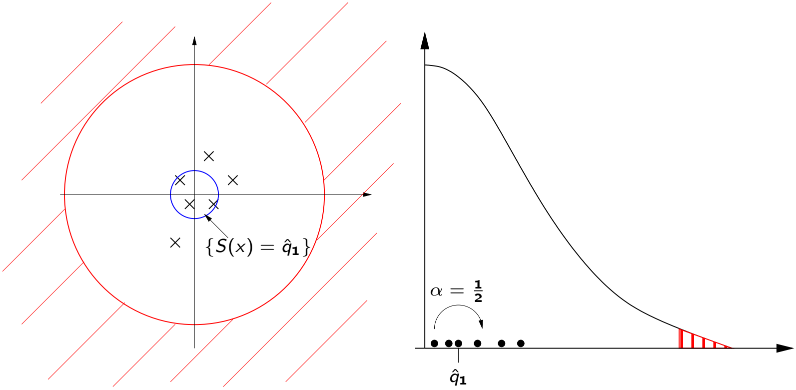
$$\arg \min_{p_1,\dots,p_n} \sum_{j=1}^{n} \frac{1-p_j}{p_j} \quad \text{s.t.} \quad \prod_{j=1}^{n} p_j = p$$

- Optimum: $p_1 = \cdots = p_n = p^{1/n}$.
- Conclusion: the levels have to be placed evenly in terms of conditional probabilities $\Rightarrow$ Adaptive Multilevel Splitting.

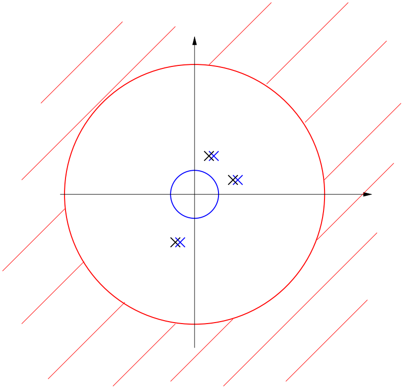
Implementation with $\alpha = 1/2$

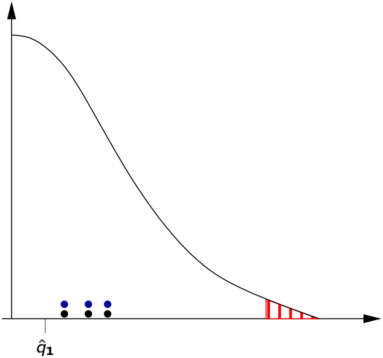


Implementation with $\alpha = 1/2$

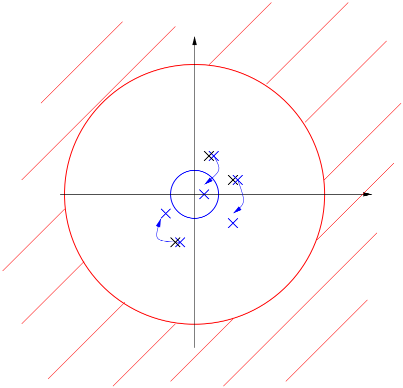


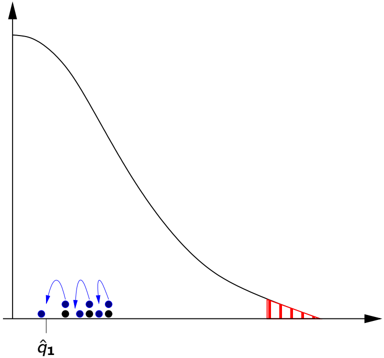
Implementation with $\alpha = 1/2$



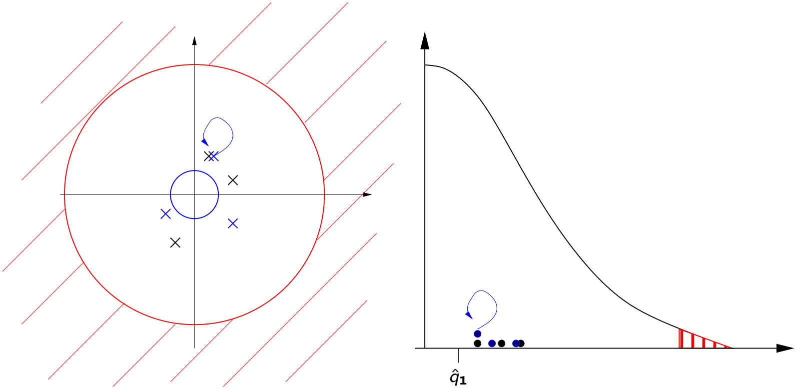


Implementation with $\alpha = 1/2$

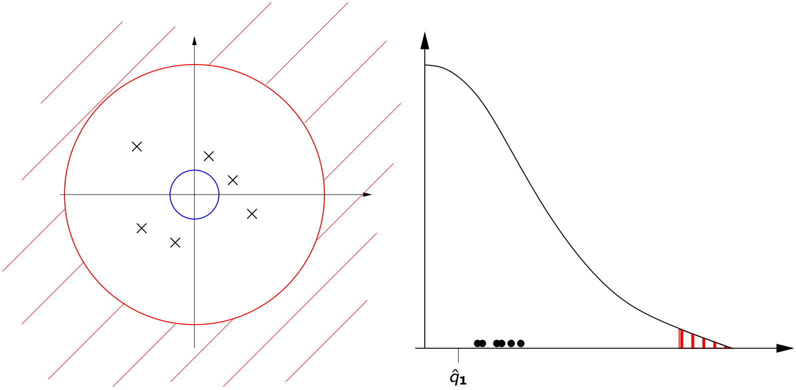




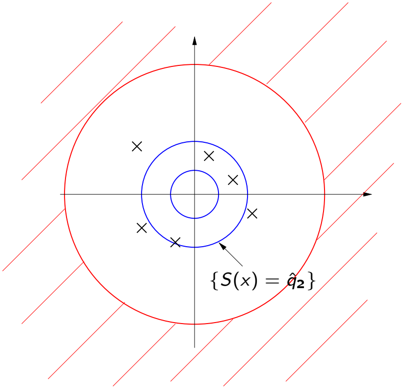
Implementation with $\alpha = 1/2$

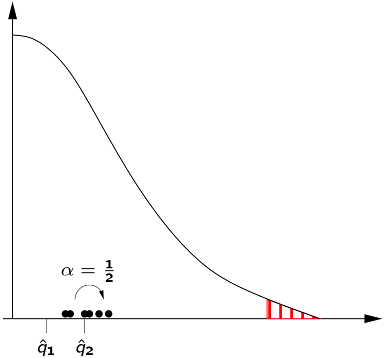


Implementation with $\alpha = 1/2$

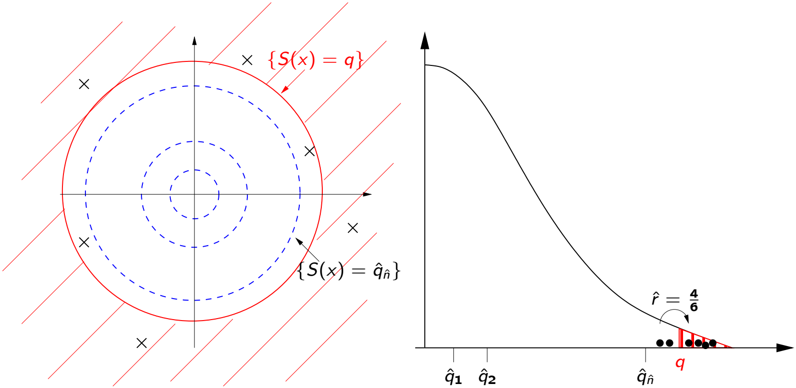


Implementation with $\alpha = 1/2$

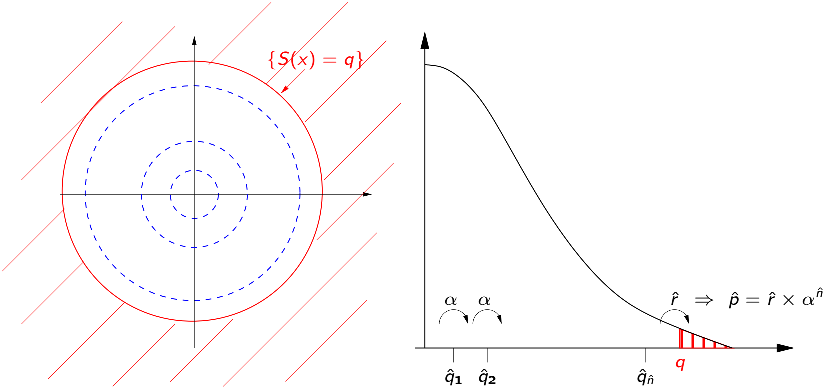




Implementation with $\alpha = 1/2$



Implementation with $\alpha = 1/2$



Interacting Particle System

- Markov chain $(X_j^1, \ldots, X_j^N)$ with initial distribution $\eta_0^{\otimes N}$ and transitions described by the previous algorithm.
- Decomposition: Fix α, then $p = r \times \alpha^n \;\Rightarrow\; \hat{p} = \hat{r} \times \alpha^{\hat{n}}$.
- Last step $\hat{n}$ corresponds to $\hat{q}_{\hat{n}+1} > q$.
- Take $\varphi(x) = \varphi(x) \times \mathbb{1}_{S(x)>q}$, *i.e.*, φ null below q.

$\Rightarrow$ Empirical normalized measures

$$\hat{\eta}_{\hat{n}}^N(\varphi) = \frac{1}{N}\sum_{i=1}^{N} \varphi(X_{\hat{n}}^i) \xrightarrow[N\to\infty]{a.s.} \eta_n(\varphi) = \mathbb{E}[\varphi(X) | S(X) > q]$$

$\Rightarrow$ Empirical unnormalized measures

$$\begin{cases} \hat{\gamma}_{\hat{n}}^N(\varphi) = \alpha^{\hat{n}} \times \frac{1}{N}\sum_{i=1}^{N} \varphi(X_{\hat{n}}^i) \xrightarrow[N\to\infty]{a.s.} \gamma_n(\varphi) = \mathbb{E}[\varphi(X)\mathbb{1}_{S(X)>q}] \\ \\ \hat{\gamma}_{\hat{n}}^N(\mathbb{1}_{S(\cdot)>q}) = \alpha^{\hat{n}} \times \hat{r} = \hat{p} \xrightarrow[N\to\infty]{a.s.} \gamma_n(\mathbb{1}_{S(\cdot)>q}) = p \end{cases}$$

Fluctuation Analysis [Cérou & Guyader (2015)]

Number of steps: $\hat{n} \to n = \lfloor \log p / \log \alpha \rfloor$ a.s. when $N \to \infty$.

Notation: $\eta_n^N(\varphi)$ and $\gamma_n^N(\varphi)$ are the multilevel splitting estimators for $p_1 = \cdots = p_n = \alpha$ (i.e., optimal fixed levels once α is given).

Theorem

Under some (mild) regularity assumptions on S, η and the kernels K_j, the asymptotic variances of $\hat{\eta}_n^N(\varphi)$ and $\hat{\gamma}_n^N(\varphi)$ are equal to those of $\eta_n^N(\varphi)$ and $\gamma_n^N(\varphi)$ respectively.

Sketch of the Proof

We have the decomposition

$$\hat{\eta}_n^N(\varphi) - \eta_n(\varphi) = \mathcal{M}_n^N + \mathcal{R}_n^N,$$

where:

- the first term splits as follows:

$$\mathcal{M}_n^N = \mathcal{M}_n^{N,1} + \mathcal{M}_n^{N,2}$$

  with martingales with respect to specific σ-fields generated by the particles X_j^i and the adaptive levels $\hat{q}_j$,
- the remaining term $\mathcal{R}_n^N$ is negligible, meaning that

$$\sqrt{n} \times \mathcal{R}_n^N \xrightarrow[N \to \infty]{\mathbb{P}} 0.$$

Conclusion

- Adaptive Multilevel Splitting allows us to place the levels in an optimal way without any loss of precision.
- The price to pay is only a low additional computational cost.
- In a different context, the take-home message here is the same as in the paper by Beskos *et al.*
- References:

  - A. Beskos, A. Jasra, N. Kantas, and A. Thiery. On the Convergence of Adaptive Sequential Monte Carlo Methods. *Annals of Applied Probability*, 2016.
  - F. Cérou and A. Guyader. Fluctuation Analysis of Adaptive Multilevel Splitting. *Annals of Applied Probability*, 2016.