Importance Sampling en grande dimension

Jason Beh[1,2,3] Florian Simatos[1,3] Jérôme Morio[2,3]

[1]ISAE-SUPAERO
[2]ONERA
[3]Fédération ENAC ISAE-SUPAERO ONERA,
Université de Toulouse, France

ETICS 2024, 25 septembre

Contexte

d entrées aléatoires → Boîte noire → Réponse du système

$X \sim f = N(0, I_d)$ $\qquad$ $\varphi : \mathbb{R}^d \mapsto \mathbb{R}$ $\qquad$ $\varphi(X)$

La défaillance du système a lieu si $\varphi(X) \geq 0$ ou
$X \in A = \{x \in \mathbb{R}^d; \varphi(x) \geq 0\}$.

But: En grande dimension (d grande), estimer la probabilité de défaillance

$$p = \mathbb{P}_f\left(X \in A\right) = \int \mathbb{1}\left(x \in A\right) f(x)\,\mathrm{d}x$$

Estimateur Monte-Carlo (MC): $(X_i)_{i=1\ldots n} \overset{\text{i.i.d.}}{\sim} f$,

$$\hat{p}_f = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left(X_i \in A\right), \qquad \frac{\sqrt{\mathbb{V}\mathrm{ar}(\hat{p}_f)}}{\mathbb{E}(\hat{p}_f)} \propto \frac{1}{\sqrt{np}}$$

Importance Sampling (IS)

g densité auxiliaire: $(Y_i)_{i=1\dots n} \overset{\text{i.i.d.}}{\sim} g$

$$p = \int \mathbb{1}\left(y \in A\right) \frac{f(y)}{g(y)} g(y)\,\mathrm{d}y \rightarrow \quad \hat{p}_g = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left(Y_i \in A\right) \frac{f(Y_i)}{g(Y_i)}$$

Il existe densité de variance nulle t.q. $\mathbb{V}\mathrm{ar}(\hat{p}_g) = 0$:

$$f|_A = \frac{\mathbb{1}\left(. \in A\right) f}{p}$$

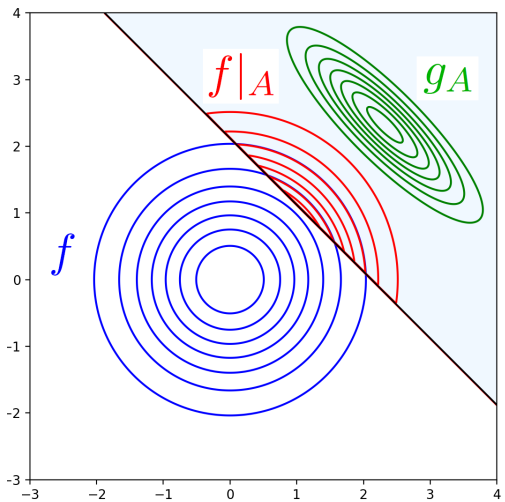
$\Rightarrow$ Bon choix de g: Réduction de variance comparé à $\hat{p}_f$

Cadre gaussien

Cadre d'étude de densité auxiliaire

$f = N(0, I)$, $f|_A = \mathbb{1}(\cdot \in A)\, f/p$ rend $\mathbb{V}\text{ar}(\hat{p}_{f|_A}) = 0$
Rechercher $g \in \mathcal{G} = \{g = N(\mu, \Sigma), \mu \in \mathbb{R}^d, \Sigma \in \mathcal{S}_d\}$
g_A optimale minimisant la divergence de Kullback-Leibler (DKL) avec $f|_A$

$$g_A = N(\mu_A, \Sigma_A) = \arg\min_{g \in \mathcal{G}} D(f|_A || g) = \arg\min_{g \in \mathcal{G}} \mathbb{E}_{f|_A}\left(\log\left(\frac{f|_A(X)}{g(X)}\right)\right)$$



Solution théorique :
$\mu_A = \mathbb{E}_{f|_A}(X), \Sigma_A = \mathbb{V}\text{ar}_{f|_A}(X) \Rightarrow$ estimer g_A ?

Problématique en grande dimension :

Estimation de moyenne et de matrice de covariance

Cadre d'étude

Régime en grande dimension: $d \to +\infty$
Toutes les grandeurs dépendent de d.
Deux régimes se distinguent pour la probabilité à estimer:

$$\inf_d p > 0$$

- [AB03; CD18; CHR22]
- Analyse théorique de la consistance d'un IS adaptatif (Cross-Entropy)

$$p \to 0$$

- [LEc+10; Can+10; LST11; LT11; GT20]
- Discutons le choix de densité auxiliaire IS avec un exemple fil rouge

$$\inf_{d} p > 0$$

Analyse théorique de la consistance d'un IS adaptatif (Cross-Entropy)

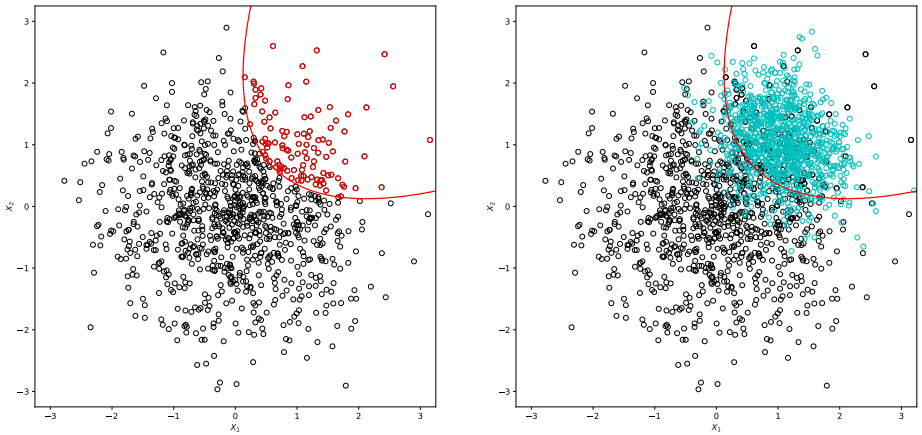
Exemple de $\inf_d p > 0$

[Uri+21; EPS24] s'intéressent à la solution u d'une EDP avec des coefficients dépendant d'un champ Gaussien E: $u(E)$ solution de $L(u, E) = 0$.

- Ils souhaitent estimer $p = \mathbb{P}(u(E) \in A) > 0$
- Pour ce faire, ils approximent E par expansion de Karhunen-Loève à l'ordre d: $E \simeq \exp(X^\top B)$ avec B une base de fonctions adaptées, et $X \sim N(0, I)$ en dimension d.
- Ils estiment $p_d = \mathbb{P}(u(e^{X^\top B}) \in A)$ qui vérifie $\inf_d p_d > 0$ puisque $p_d \to p$.

IS adaptatif, Cross-Entropy

CE estime progressivement g_A avec $\left(\hat{g}_t = N(\hat{\mu}_t, \hat{\Sigma}_t)\right)$. À l'étape t,



m échs $(Y_i)_{i=1\ldots m} \overset{\text{i.i.d.}}{\sim} \hat{g}_t$ ayant ρm plus grands $(\varphi(Y_i))$
$\Rightarrow$ zone de défaillance intermédiaire
$\hat{A}_t = \{x \in \mathbb{R}^d;\ \varphi(x) \geq \hat{q}_{(\lfloor (1-\rho)m \rfloor)}\}$.

Ensuite, estimer $\mu_{\hat{A}_t} = \mathbb{E}_{f|_{\hat{A}_t}}(X)$ et $\Sigma_{\hat{A}_t} = \mathbb{V}\text{ar}_{f|_{\hat{A}_t}}(X)$ par IS avec les poids $f(Y_i)/\hat{g}_t(Y_i) \Rightarrow \hat{\mu}_{t+1}, \hat{\Sigma}_{t+1}$.

Alors $\hat{g}_{t+1} = N(\hat{\mu}_{t+1}, \hat{\Sigma}_{t+1})$.

Estimer étape par étape $\Sigma_{\hat{A}_t}$ au lieu de directement Σ_A

Analyse théorique de la CE en grande dimension

https://arxiv.org/abs/2309.16828 (avec Yonatan Shadmi, Florian Simatos)

Cadre: $d \to +\infty$.

Hypothèse centrale: $\inf_d p_f(A) > 0$ + Hypothèses techniques

m : taille d'échantillon par itération de CE pour construire $\hat{g}_t$

Theorème

Pour tout $t \geq 0$, il existe $\kappa_t > 0$ t.q. si $m \gg d^{\kappa_t}$, alors

$$\frac{\hat{p}_{\hat{g}_t}}{p_f(A)} = \frac{1}{p_f(A)n} \sum_{i=1}^{n} \frac{f(Y_i)}{\hat{g}_t(Y_i)} \mathbb{1}\left(Y_i \in A\right) \Rightarrow 1 \text{ pour toute } n \to +\infty.$$

Pourquoi c'est intéressant ? L'idée répandue pour l'IS:

$n \gg \exp(d)$ pour avoir un estimateur consistant [BBL08]

Ce que nous avons prouvé:

en apprenant g_A préalablement par CE avec m polynomial en d,

l'estimateur CE est consistant pour toute vitesse $n \to +\infty$.

Le résultat en pratique

À notre connaissance, premier résultat de consistance de la CE lorsque $d \to +\infty$

Mais, pour $\hat{g}_t = N(\hat{\mu}_t, \hat{\Sigma}_t)$
κ_t liée à l'inverse de la plus petite valeur propre $\frac{1}{\lambda_1(\hat{\Sigma}_{t-1})}$

- Résultats similaires sur la CE couplée avec la projection sur $\hat{\Sigma}_t$ [Uri+21; EMS21; EMS24] et improved CE [PGS19] en cours de finalisation
- Ces résultats permettront de comparer m pour chaque algorithme et quantifier le gain de la projection sur $\hat{\Sigma}_t$

$p \to 0$

Discutons le choix de densité auxiliaire IS avec un exemple fil rouge

Exemples de $p \to 0$

- Highly Reliable Markovian Systems [LEc+10; LT11]
- Static Network Reliability Estimation [Can+10; LST11]
- Large Deviation Estimation [LEc+10; GT20]

Dans ces travaux, des densités adaptées qui procurent de bonnes propriétés asymptotiques pour l'estimateur IS (logarithmic efficiency/bounded relative moment (of order k) *etc.*) sont conçues.

Rq. Nous considérons que φ est une fonction boîte noire sans connaissance préalable: nous ne pouvons pas savoir quand utiliser ces densités spécifiques pour les φ cités au-dessus $\Rightarrow$ Cas restrictif de densités gaussiennes, en existe-t-il qui ont des "bonnes propriétés asymptotiques" ?

Fil rouge: Hyperplan généralisé [Buc04; GT20]

$$A = \{x \in \mathbb{R}^d : \sum_{j=1}^{d} x(j) \geq d^\gamma\} \quad p = p_f(A) = \mathbb{P}_{N(0,1)}\left(N \geq d^{\gamma - 1/2}\right)$$

$$\mu_A = \mathbb{E}_{f|_A}(X),\ \Sigma_A = \mathbb{V}\mathrm{ar}_{f|_A}(X)$$

On considère $\gamma > 1/2$ donc $p \to 0$

Établissons la CNS pour la consistance et un TCL

Rq. Pour $\gamma = 1$, plusieures densités ayant de bonnes propriétés (bounded relative moment of order $k \geq 1$ ou autre) sont conçues [LEc+10; Car+14], mais nous sommes restreints à des densités gaussiennes

Condition nécessaire et suffisante pour la consistance

$(Y_i) \overset{\text{iid}}{\sim} g$, $\dfrac{\hat{p}_g}{p} := \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \ell_A(Y_i)$ avec $\ell_A(y) = \dfrac{f|_A(y)}{g(y)}$.

Proposition

Lorsque $d \to \infty$,

$$\frac{\hat{p}_g}{p} \Rightarrow 1 \iff \frac{\ell_A(X)}{n} \Rightarrow 0,\ X \sim f|_A.$$

Idée de preuve: Convergence de suite triangulaire de variables aléatoires [JŠ03, VII Théorème 2.35]

Conséquence: un résultat similaire à [CD18, Theorem 1.3]

Pour $t \gg \sqrt{\mathbb{V}\text{ar}_{f|_A}\left[\log \ell_A(X)\right]}$,

$$n = \exp\left(D(f|_A||g) + t\right) \Longrightarrow \frac{\hat{p}_g}{p} \Rightarrow 1$$

$$n = \exp\left(D(f|_A||g) - t\right) \Longrightarrow \frac{\hat{p}_g}{p} \not\Rightarrow 1$$

Vitesse de convergence

Proposition

Supposons que g est tq $\forall d, \mathbb{V}\text{ar}_g[\ell_A(Y)] < +\infty$.Pour toute suite $n(d)$ t.q. $n \gg \mathbb{V}\text{ar}_g\left[\ell_A(Y)\right]$,

$$\sqrt{\frac{n}{\mathbb{V}\text{ar}_g\left[\ell_A(Y)\right]}}\left(\frac{\hat{p}_g}{p} - 1\right) \Rightarrow N(0,1) \text{ lorsque } d \to +\infty.$$

Idée de preuve: Theorème de Lindeberg-Feller [JŠ03, VII Theorem 5.2]

À trouver: Si g est telle qu'il existe $d : \mathbb{V}\text{ar}_g[\ell_A(Y)] = +\infty$, on devrait avoir une convergence similaire vers une distribution infiniment divisible

Fil rouge: Hyperplan généralisé [Buc04; GT20]

Traduction des conditions précédentes pour consistance et vitesse en taille d'échantillon n nécessaire

$$A = \{x \in \mathbb{R}^d : \sum_{j=1}^{d} x(j) \geq d^\gamma\} \quad p = p_f(A) = \mathbb{P}_{N(0,1)}\left(N \geq d^{\gamma - 1/2}\right)$$

$$\mu_A = \mathbb{E}_{f|_A}(X),\ \Sigma_A = \mathbb{V}\mathrm{ar}_{f|_A}(X)$$

si $\gamma > 1/2$
alors $p \to 0$

Choix de g	consistance	TCL vers $N(0,1)$?
$f = N(0, I)$	$n \gg d^{\gamma-1/2}\exp(d^{2\gamma-1})$	
$N(\mu_A, I)$	$n \gg d^{\gamma-1/2}$	
$N(\mu_A, \Sigma_A)$	$n \to \infty$	Non, $\lambda_1(\Sigma_A) < 1/2$

Quelle est la densité auxiliaire à utiliser?

Dépend du choix de la métrique d'erreur!

Métriques d'erreur

- Probabilité d'erreur d'ordre $\varepsilon > 0$ [HN16; San18; CHR22]

$$\mathbb{P}\left(\left|\frac{\hat{p}_g}{p} - 1\right| \geq \varepsilon\right)$$

- Erreur L_1 [CD18]

$$\mathbb{E}\left(\left|\frac{\hat{p}_g}{p} - 1\right|\right)$$

- Erreur L_2 [AB03; Aga+17; GT20; HR21]

$$\mathbb{E}\left(\left(\frac{\hat{p}_g}{p} - 1\right)^2\right)^{1/2}$$

$N(\mu_A, \Sigma_A)$ VS $N(\mu_A, I)$

$$g_A = N(\mu_A, \Sigma_A)$$

- $\hat{p}_{g_A}$ consistant $\forall n \to \infty$: densité la plus facilement consistante
- g_A minimise DKL par rapport à $f|_A$ densité zéro-variance parmi les densités gaussiennes, **mais $\hat{p}_{g_A}$ n'admet pas de variance**
  Pourtant, c'est la densité auxiliaire que vise la CE !
- g_A est un contre-exemple de la phrase suivante dans [HR02] (avec notre notation):
  «This suggests that, as $p \to 0$, the Variance Minimization and Cross Entropy problems tend to have the same solution»

$N(\mu_A, \Sigma_A)$ VS $N(\mu_A, I)$

$$g_I = N(\mu_A, I)$$

- $\hat{p}_{g_I}$ consistant ssi $n \gg d^{\gamma - 1/2}$
- $\hat{p}_{g_I}$ vérifie un TCL si $n \gg d^{\gamma - 1/2}$.
- Il est connu que $\hat{p}_{g_I}$ est logarithmiquement efficace à l'ordre 2 [Buc04; LEc+10; GT20]

Définition: Efficacité logarithmique [LEc+10; GT20, par ex]

$\hat{p}_g$ est logarithmiquement efficace à l'ordre 2 (LE-2) si

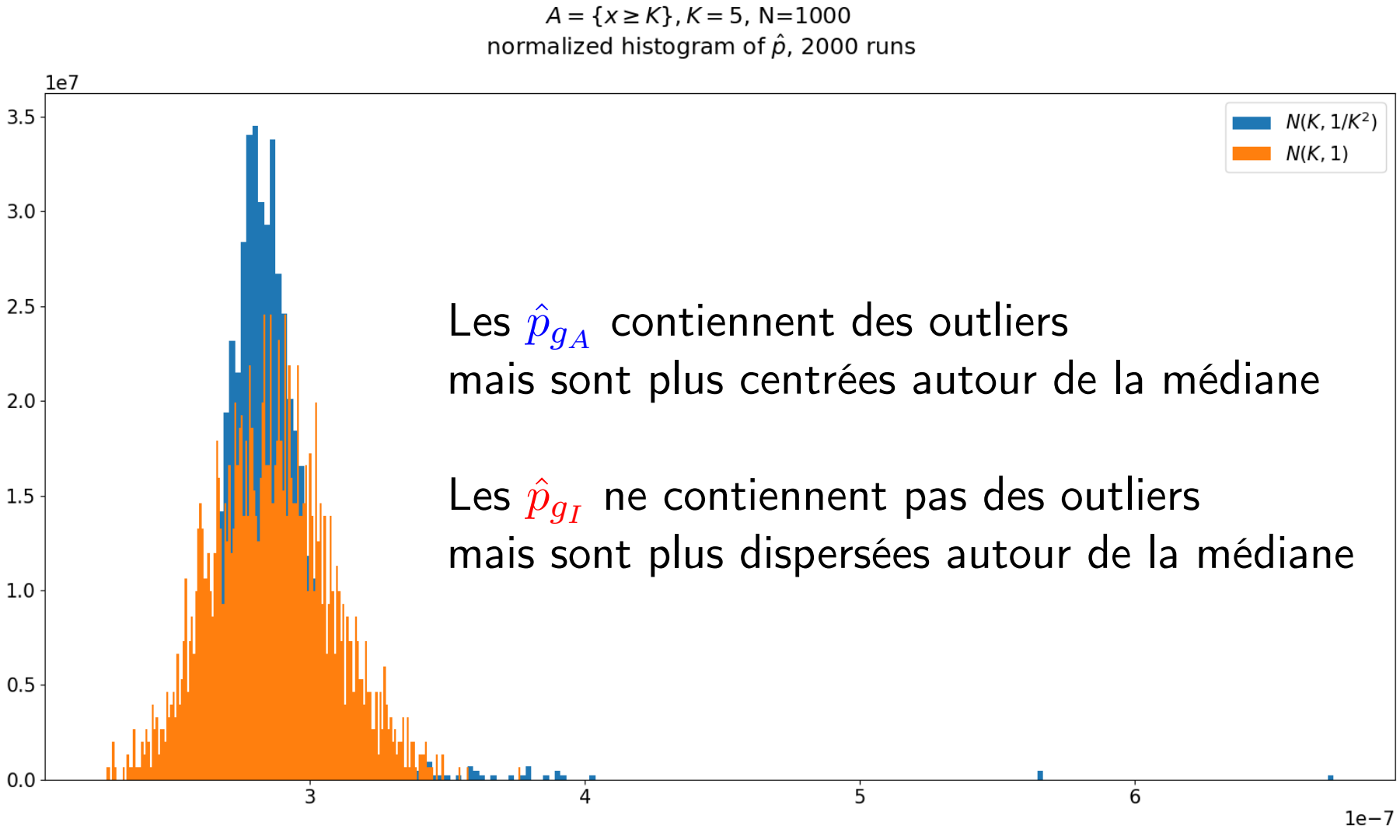
$$\lim_{d \to \infty} \frac{\log \mathbb{E}(\hat{p}_g^2)}{2 \log p} = 1$$

Interprétation: si $\hat{p}_g$ est LE-2, alors $\mathbb{V}\mathrm{ar}_g[\ell_A(Y)]$ ne croît pas en $\exp(d)$, ce qui est en effet le cas pour g_I: $\mathbb{V}\mathrm{ar}_{g_I}[\ell_A(Y)] \sim d^{\gamma - 1/2}$

Rappel: CNS TCL vers $N(0,1)$ requiert $n \gg \mathbb{V}\mathrm{ar}_g[\ell_A(Y)]$

En pratique: Non-asymptotique

Le choix n'est pas tranché: illustrons avec $d = 1$, $A = \{x \in \mathbb{R}, x \geq 5\}$

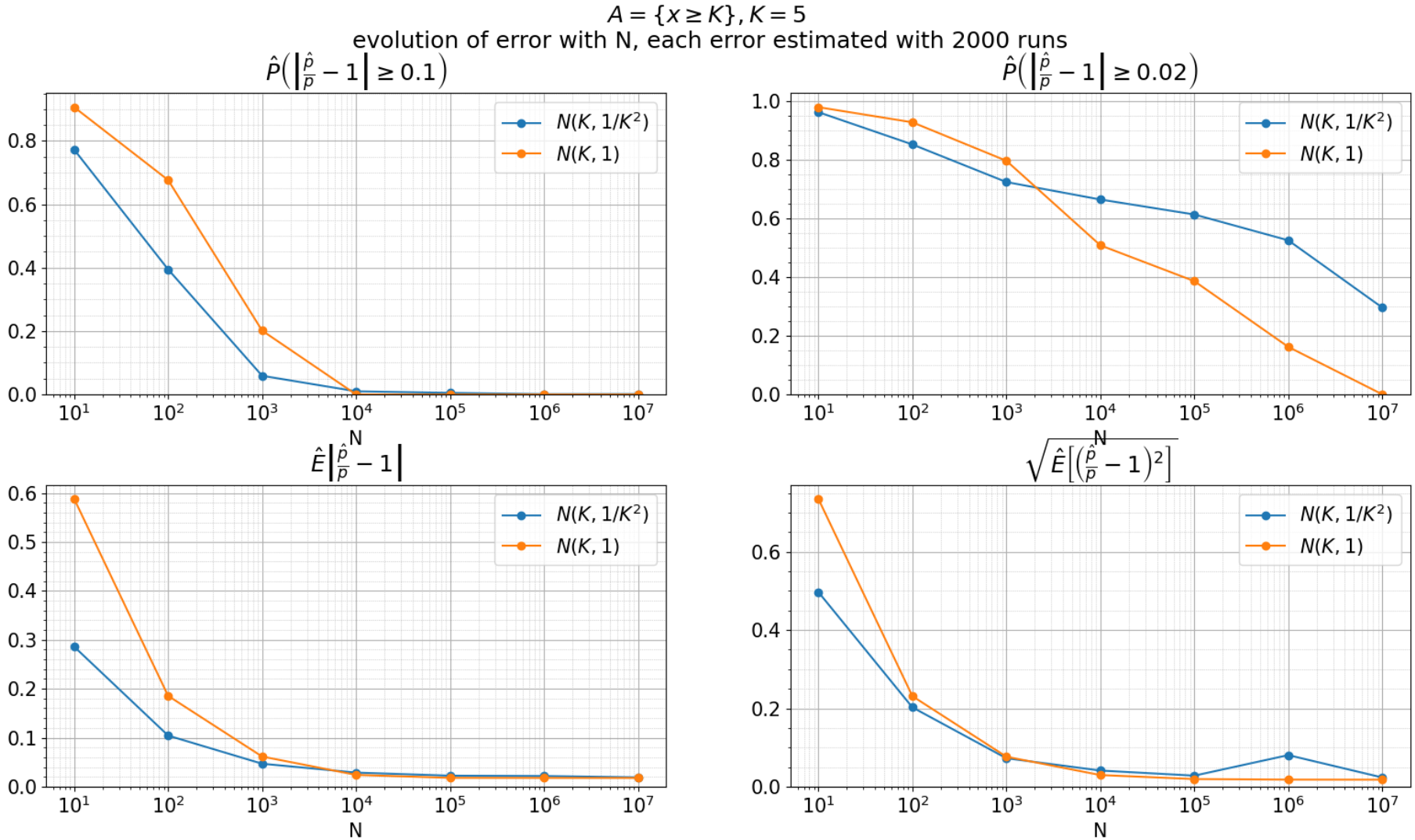


Les $\hat{p}_{g_A}$ contiennent des outliers
mais sont plus centrées autour de la médiane

Les $\hat{p}_{g_I}$ ne contiennent pas des outliers
mais sont plus dispersées autour de la médiane

En pratique: Non-asymptotique

Le choix de g dépend de la métrique et la taille d'échantillon



Coût d'estimation de μ_A, Σ_A

En Adaptive Importance Sampling, il faut prendre en compte le coût d'estimation de μ_A et Σ_A.

AIS hypothétique: si on a $(X_i)_{i=1\dots m} \overset{\text{i.i.d.}}{\sim} f|_A$ et

$$\hat{\mu}_A = \frac{1}{m}\sum_{i=1}^{m} X_i \ , \ \hat{\Sigma}_A = \frac{1}{m}\sum_{i=1}^{m} X_i X_i^\top ;$$

Conjecture (prouvé dans [BSS23] pour le cas $\inf_d p > 0$)

Choix de g	consistance	TCL vers $N(0,1)$?	Coût d'estimation
$f = N(0, I)$	$n \gg d^{\gamma-1/2}\exp(d^{2\gamma-1})$		0
$N(\mu_A, I)$	$n \gg d^{\gamma-1/2}$		$m \gg d$
$N(\mu_A, \Sigma_A)$	$n \to \infty$	Non, $\lambda_1(\Sigma_A) < 1/2$	$m \gg d^2$

$N(\mu_A, \Sigma_A)$ intéressante si budget $n + m = O(d^2)$ et si $\gamma > 2.5$

Illustration 2: Mixture Importance Sampling

Les conditions aident à étudier Mixture IS:

g_0 et g_1: 2 densités auxiliaires

(α) une suite réelle (en d) avec $\inf_d \alpha > 0$ et $\sup_d \alpha < 1$.

La mixture: $g_\alpha = (1-\alpha)g_0 + \alpha g_1$. Le CNS de consistance implique

Si $\hat{p}_{g_0}$ ou $\hat{p}_{g_1}$ est consistant, alors $\hat{p}_{g_\alpha}$ est consistant.

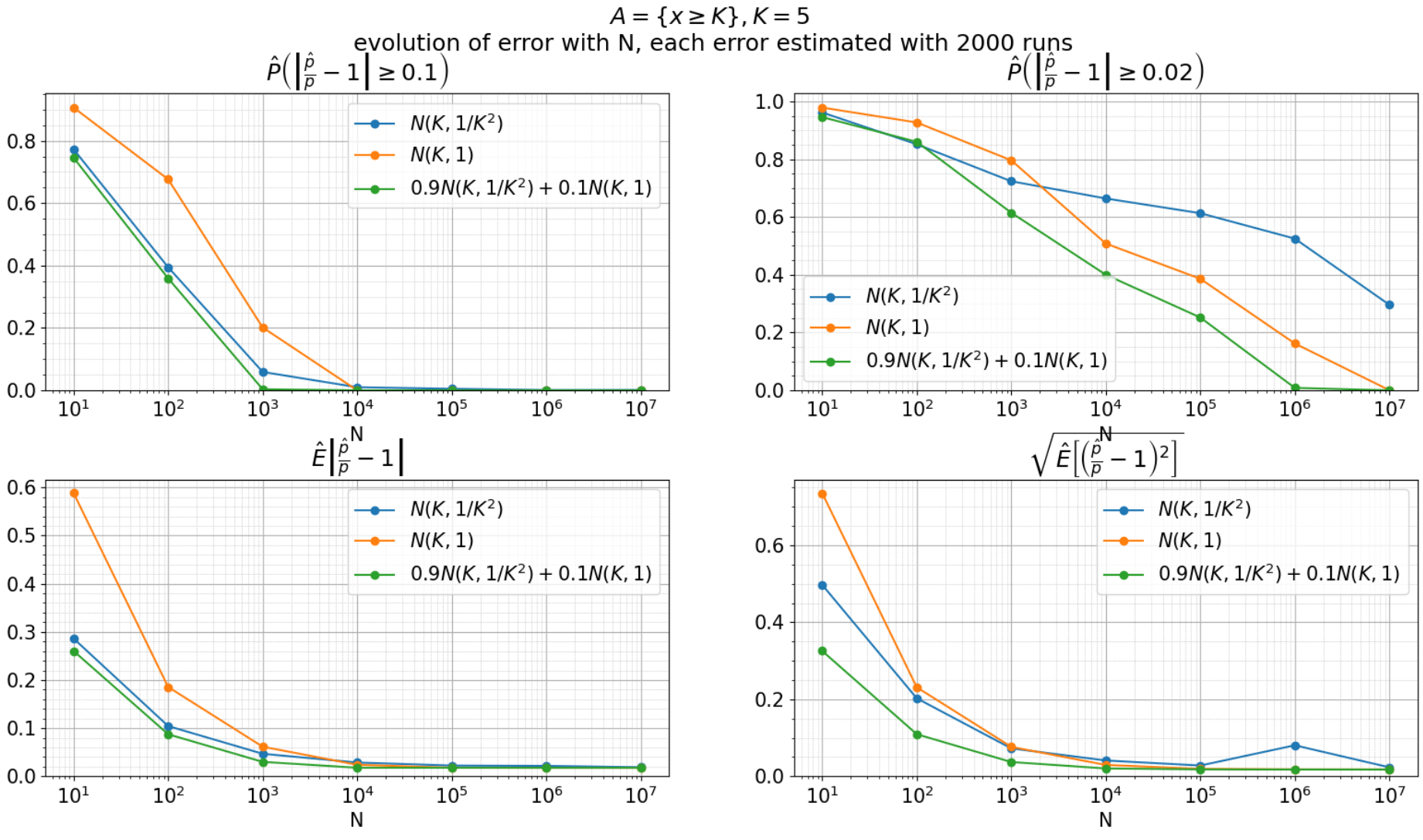
De plus, pour la TCL, la variance de rapport de vraisemblance vérifie:

$$\mathbb{V}\text{ar}_{g_\alpha}\left[\frac{f|_A(X)}{g_\alpha(X)}\right] \leq \min\left\{\frac{1}{1-\alpha}\mathbb{V}\text{ar}_{g_0}\left[\frac{f|_A(Y)}{g_0(Y)}\right] + \frac{\alpha}{1-\alpha}, \frac{1}{\alpha}\mathbb{V}\text{ar}_{g_1}\left[\frac{f|_A(Y)}{g_1(Y)}\right] + \frac{1-\alpha}{\alpha}\right\}$$

Si $\hat{p}_{g_0}$ ou $\hat{p}_{g_1}$ vérifie un TCL vers $N(0,1)$, alors $\hat{p}_{g_\alpha}$ aussi.

Mixture IS

Mixture = "Best of both worlds"



Perspective

- Les nouvelles CNS de consistance et TCL permettraient d'étudier l'IS avec une approche différente de [BSS23]
- Nous espérons pouvoir étendre les résultats lorsque l'hypothèse $\inf_d p > 0$ est enlevée
- Suite: Trouver les lois limites pour g sans $\mathbb{V}\mathrm{ar}_g[\ell_A(X)]$, étudier les cas $p \to 0$ de manière plus approfondie, et les simulations avec n qui croît avec la dimension
- Peut-on trouver des densités *gaussiennes* qui vérifient les propriétés asymptotiques définies dans [LEc+10] avec **le minimum de hypothèses sur** φ? Le cas écheant, mixture des gaussiennes?
  - Si une telle densité existe, quelle est le coût d'estimation de ces paramètres par la CE?
- Quelle métrique d'erreur est la plus pertinente en pratique et quelle est la densité gaussienne optimale selon la métrique?

References I

[AB03] S. Au and J. Beck. "Important sampling in high dimensions". In: *Structural Safety* 25.2 (2003), pp. 139–163.

[Aga+17] S. Agapiou et al. *Importance Sampling: Intrinsic Dimension and Computational Cost*. arXiv:1511.06196 [stat]. 2017.

[BBL08] T. Bengtsson, P. Bickel, and B. Li. "Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems". In: *Institute of Mathematical Statistics Collections*. Beachwood, Ohio, USA: Institute of Mathematical Statistics, 2008, pp. 316–334.

[BSS23] J. Beh, Y. Shadmi, and F. Simatos. "Insight from the Kullback–Leibler divergence into adaptive importance sampling schemes for rare event analysis in high dimension". In: *arXiv preprint* (2023). arXiv: 2309.16828 [math.ST].

[Buc04] J. A. Bucklew. *Introduction to rare event simulation*. Vol. 5. Springer, 2004.

[Can+10] H. Cancela et al. "Combination of conditional Monte Carlo and approximate zero-variance importance sampling for network reliability estimation". In: *Proceedings of the 2010 Winter Simulation Conference*. 2010, pp. 1263–1274.

[Car+14] Caron, Virgile et al. "Some Recent Results in Rare Event Estimation". In: *ESAIM: Proc.* 44 (2014), pp. 239–259.

[CD18] S. Chatterjee and P. Diaconis. "The sample size required in importance sampling". In: *The Annals of Applied Probability* 28.2 (2018).

References II

[CHR22] F. Cérou, P. Héas, and M. Rousset. *Entropy minimizing distributions are worst-case optimal importance proposals*. 2022. arXiv: 2212.04292 [math.NA].

[EMS21] M. El Masri, J. Morio, and F. Simatos. "Improvement of the cross-entropy method in high dimension for failure probability estimation through a one-dimensional projection without gradient estimation". In: *Reliability Engineering & System Safety* 216 (2021), p. 107991.

[EMS24] M. El Masri, J. Morio, and F. Simatos. "Optimal Projection for Parametric Importance Sampling in High Dimensions". In: *Computo* (11, 2024).

[EPS24] M. Ehre, I. Papaioannou, and D. Straub. *Stein Variational Rare Event Simulation*. 2024. arXiv: 2308.04971 [stat.ME].

[GT20] A. Guyader and H. Touchette. "Efficient Large Deviation Estimation Based on Importance Sampling". In: *Journal of Statistical Physics* 181.2 (2020), pp. 551–586.

[HN16] H. Hult and P. Nyquist. "Large deviations for weighted empirical measures arising in importance sampling". In: *Stochastic Processes and their Applications* 126.1 (2016), pp. 138–170.

[HR02] T. Homem-de-Mello and R. Y. Rubinstein. "Rare event estimation for static models via cross-entropy and importance sampling". In: (2002).

References III

[HR21] C. Hartmann and L. Richter. *Nonasymptotic bounds for suboptimal importance sampling*. arXiv:2102.09606 [cs, math, stat]. 2021.

[JŠ03] J. Jacod and A. N. Širjaev. *Limit theorems for stochastic processes*. 2nd ed. Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen 288. Springer Berlin, Heidelberg, 2003.

[LEc+10] P. L'Ecuyer et al. "Asymptotic robustness of estimators in rare-event simulation". In: *ACM Trans. Model. Comput. Simul.* 20.1 (2010).

[LST11] P. L'Ecuyer, S. Saggadi, and B. Tuffin. "Graph reductions to speed up importance sampling-based static reliability estimation". In: *Proceedings of the 2011 Winter Simulation Conference (WSC)*. 2011, pp. 429–438.

[LT11] P. L'Ecuyer and B. Tuffin. "Approximating zero-variance importance sampling in a reliability setting". In: *Annals of Operations Research* 189.1 (2011), pp. 277–297.

[PGS19] I. Papaioannou, S. Geyer, and D. Straub. "Improved cross entropy-based importance sampling with a flexible mixture model". In: *Reliability Engineering & System Safety* 191 (2019), p. 106564.

[San18] D. Sanz-Alonso. "Importance Sampling and Necessary Sample Size: An Information Theory Approach". In: *SIAM/ASA Journal on Uncertainty Quantification* 6.2 (2018), pp. 867–879. eprint: https://doi.org/10.1137/16M1093549.

References IV

[Uri+21] F. Uribe et al. "Cross-Entropy-Based Importance Sampling with Failure-Informed Dimension Reduction for Rare Event Simulation". In: *SIAM/ASA Journal on Uncertainty Quantification* 9.2 (2021), pp. 818–847.