Étude du biais et de la variance de l'arithmétique stochastique

El-Mehdi El ARAR
el-mehdi.el-arar@uvsq.fr

Laboratoire D'Informatique Parallélisme Réseaux Algorithmes Distribués (LI – PARAD)

Interflop, 06 Octobre 2021

Plan

- Introduction.
- Étude du biais de l'arithmétique stochastique sur la méthode d'intégration des rectangles.
- Étude de la variance d'un mode d'arrondi stochastique.

Introduction

- Estimation de l'erreur.
- Étude de la distribution en sortie.
  - Moment du premier ordre : Comparaison de deux modes d'arrondi stochastiques.
  - Moment du deuxième ordre : Produit scalaire, système linéaire.

Notations et terminologie

Définition 1.1.

Soient $\mathcal{F} \subset \mathbb{R}$ l'ensemble des représentants et $x \in \mathbb{R}$. On définit :

- $fl[x] = x(1+\delta)$, *où δ est l'erreur relative vérifie* $|\delta| < u = \frac{1}{2}\beta^{1-p}$.
- *L'arrondi vers le haut $\lceil x \rceil$ et l'arrondi vers le bas $\lfloor x \rfloor$ par :*

$$\lceil x \rceil = \min\{y \in \mathcal{F} : y \geqslant x\}, \quad \lfloor x \rfloor = \max\{y \in \mathcal{F} : y \leqslant x\}.$$

- $\varepsilon(x) = \beta^{e-p} = \lceil x \rceil - \lfloor x \rfloor$ *où β est la base, e est l'exposant de x et p est la précision machine.*
- $\theta(x) = \frac{x - \lfloor x \rfloor}{\lceil x \rceil - \lfloor x \rfloor}$ *est la fraction arrondie de $\varepsilon(x)$.*

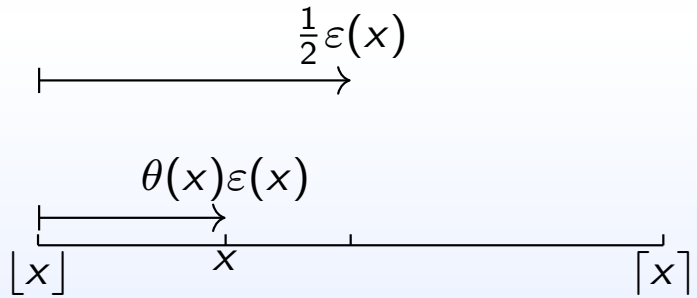


FIGURE – $\theta(x)$ est la fraction arrondie de $\varepsilon(x)$.

Définition 1.2.

On définit les deux modes d'arrondi stochastiques :

$$\textbf{\textit{random\_nearness}} : fl(x) = \begin{cases} \lceil x \rceil & \text{avec probabilité } \theta(x), \\ \lfloor x \rfloor & \text{avec probabilité } 1-\theta(x). \end{cases}$$

$$\textbf{\textit{random\_up\_or\_down}} : fl(x) = \begin{cases} \lceil x \rceil & \text{avec probabilité } 1/2, \\ \lfloor x \rfloor & \text{avec probabilité } 1/2. \end{cases}$$

D. S. Parker. *Monte Carlo Arithmetic : exploiting randomness in floating-point arithmetic*. University of California (Los Angeles). Computer Science Department, 1997.

Propriété 1.1.

Soit $x \in \mathbb{R}$ et $x \notin \mathcal{F}$. On note X la variable aléatoire de la distribution du résultat obtenu en arrondissant x par l'un des deux modes d'arrondi, i.e.,

$$X = \textbf{\textit{random\_round}}(x) = \textbf{\textit{round}}(x + \beta^{e_x - p}\xi),$$

*Le mode **random_nearness** est réalisé pour ξ suivant une loi uniforme sur $(-\frac{1}{2}; \frac{1}{2})$. Ce mode est sans biais, et on a :*

$$E[X] = \theta(x)\lceil x \rceil + (1 - \theta(x))\lfloor x \rfloor = x.$$

*Le mode **random_up_or_down** est réalisé pour ξ suivant une loi uniforme sur $(-\theta(x); 1 - \theta(x))$. Ce mode est biaisé, et on a :*

$$E[X] = x + \varepsilon(x)(\frac{1}{2} - \theta(x)).$$

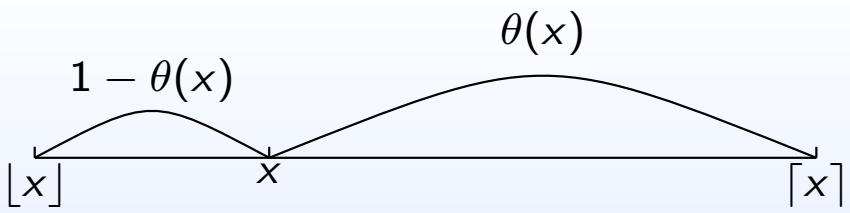


FIGURE – **Random_nearness**.

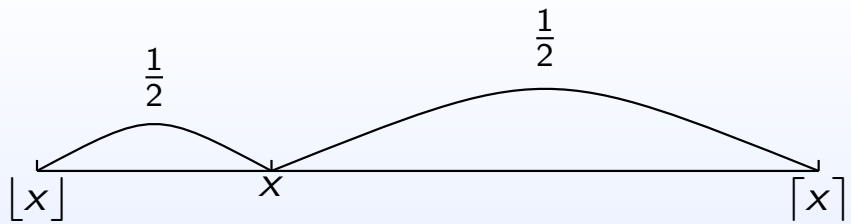


FIGURE – **Random_up_or_down**.

Intégration de la fonction constante

On s'intéresse au calcul de $\int_0^1 1dx = 1$ par la méthode des rectangles.

```
float dx = 1/N;
float s = 0.0;
for (int i=0; i < N; i++) {
    s += dx*1;
}
return s;
```

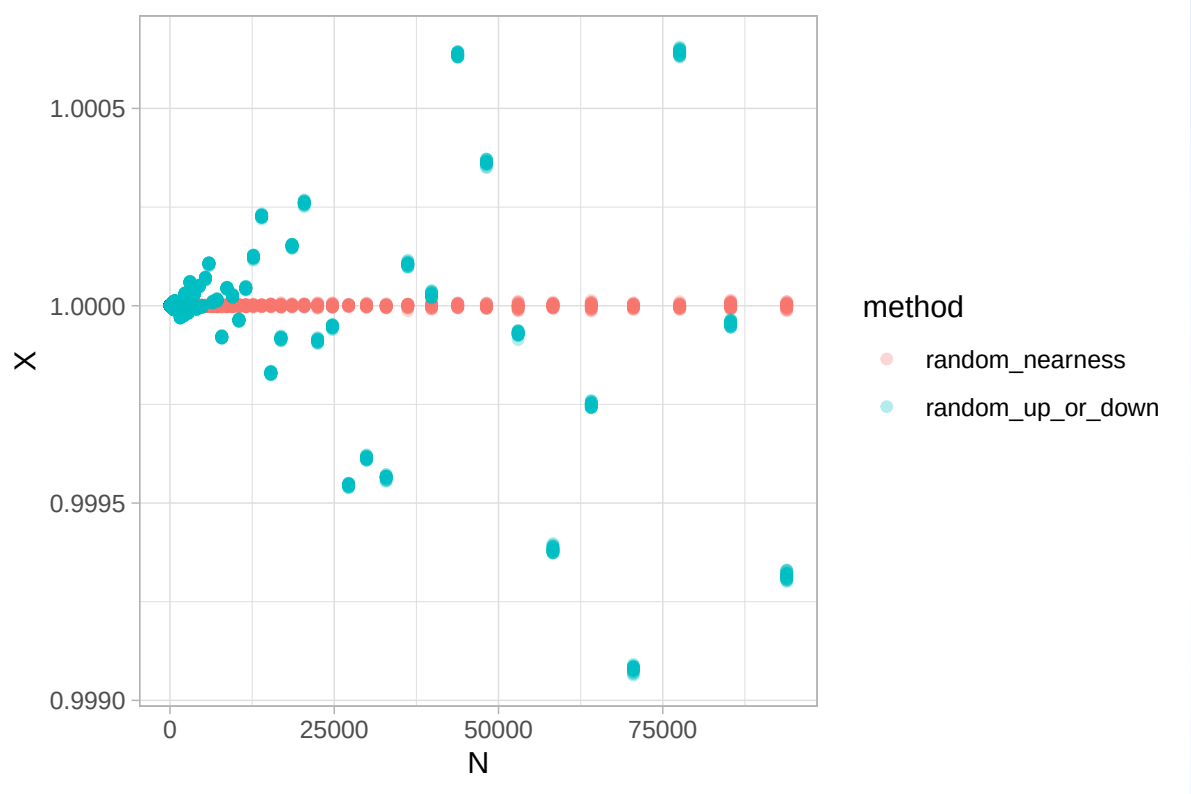


FIGURE – 30 échantillons d'arrondi stochastique de $\int_0^1 1\,dx$.

Lemme 2.1.

La distribution X est obtenu en addition N fois le pas dx. On note S_k la variable aléatoire de la somme partielle à l'étape $0 \leqslant k \leqslant N-1$ et s_k la somme partielle exacte (la valeur mathématique). Avec $S_{N-1} = X$. On a :

$$E[S_k - s_k] = \varepsilon(s_k)(\frac{1}{2} - \theta(s_k)),$$

et

$$|E[S_k - s_k]| < \frac{1}{2}\varepsilon(s_k).$$

Lemme 2.1.

La distribution X est obtenu en addition N fois le pas dx. On note S_k la variable aléatoire de la somme partielle à l'étape $0 \leqslant k \leqslant N-1$ et s_k la somme partielle exacte (la valeur mathématique). Avec $S_{N-1} = X$. On a :

$$E[S_k - s_k] = \varepsilon(s_k)(\frac{1}{2} - \theta(s_k)),$$

et

$$|E[S_k - s_k]| < \frac{1}{2}\varepsilon(s_k).$$

Remarque 2.1.

- *$\varepsilon(s_k)$ est constant entre deux puissances consécutives de la base, en particulier pour la base deux, à chaque changement de puissance de deux, il est doublé.*

Lemme 2.1.

La distribution X est obtenu en addition N fois le pas dx. On note S_k la variable aléatoire de la somme partielle à l'étape $0 \leqslant k \leqslant N-1$ et s_k la somme partielle exacte (la valeur mathématique). Avec $S_{N-1} = X$. On a :

$$E[S_k - s_k] = \varepsilon(s_k)(\frac{1}{2} - \theta(s_k)),$$

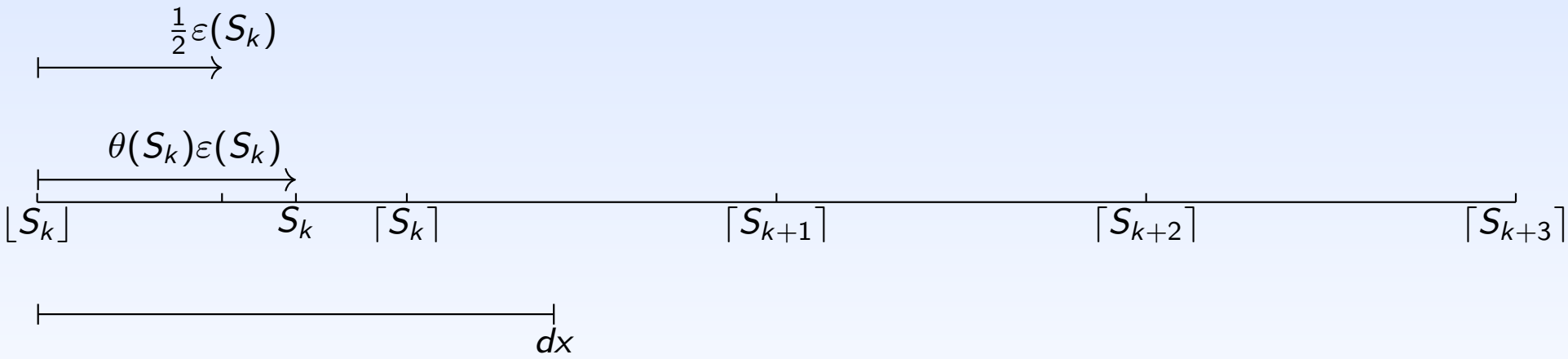
et

$$|E[S_k - s_k]| < \frac{1}{2}\varepsilon(s_k).$$

Remarque 2.1.

- *$\varepsilon(s_k)$ est constant entre deux puissances consécutives de la base, en particulier pour la base deux, à chaque changement de puissance de deux, il est doublé.*
- *Sauf la première valeur, $\theta(s_k)$ est constante entre deux puissances consécutives de la base.*

Le schéma suivant explique la remarque précédente :



Le schéma suivant explique la remarque précédente :

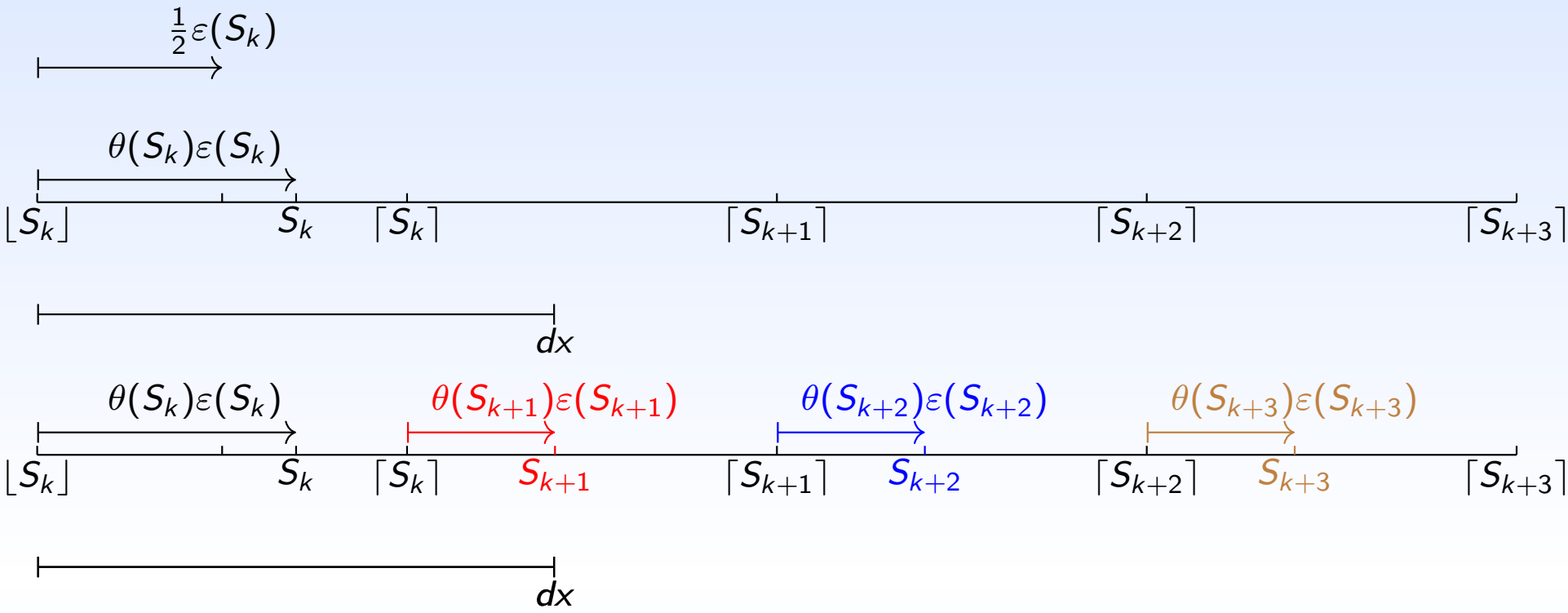


FIGURE – θ est constante entre deux puissance de deux successives.

Le schéma suivant explique la remarque précédente :

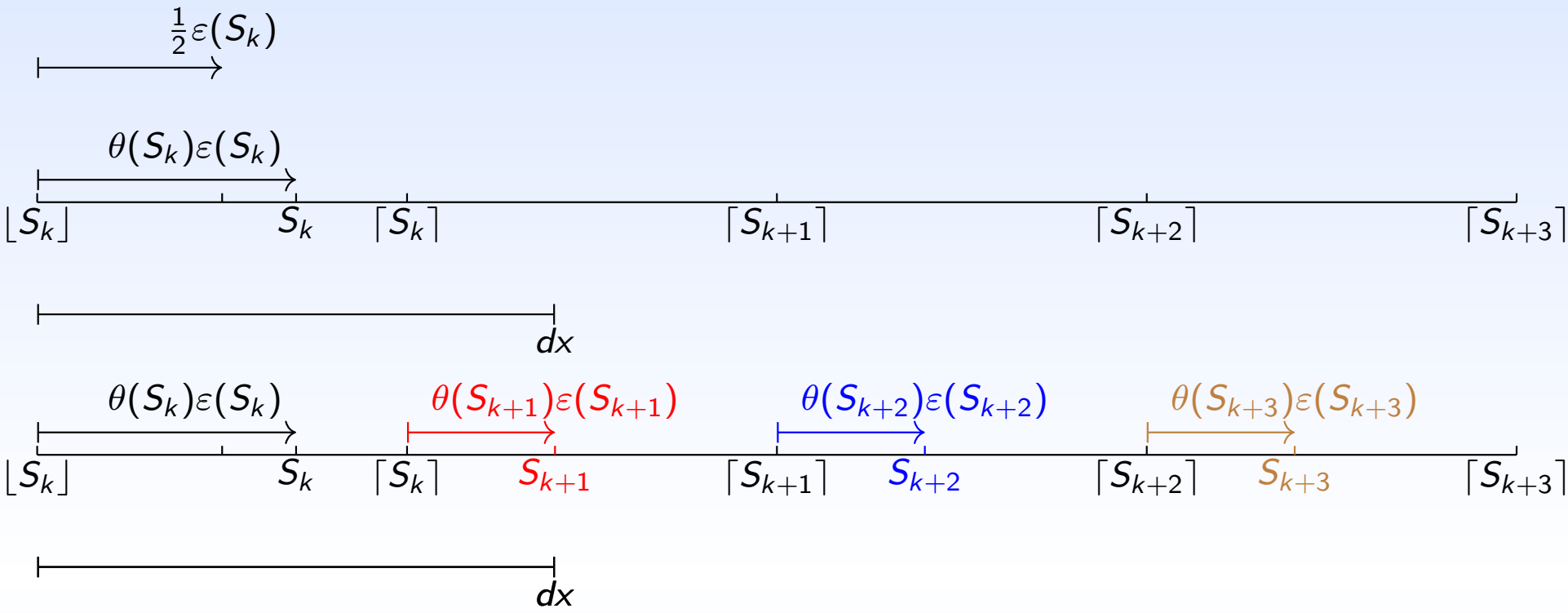


FIGURE – θ est constante entre deux puissance de deux successives.

Conclusion

- L'origine du biais.
- Le mode **nearness** se comporte mieux.

Étude de la variance

- M. P. Connolly, N. J. Higham and Théo Mary. Stochastic rounding and its probabilistic backward error analysis.
- Mean independant.

Définition 3.1.

On dit qu'une variable aléatoire Y est "mean independent" par rapport à une variable aléatoire X si et seulement si : pour tout x tel que la probabilité que $X = x$ est non nulle on a $E(Y/X = x) = E(Y)$ et on écrit $E(Y/X) = E(Y)$.
Des variables aléatoires $\delta_1, \delta_2, ...$ sont "mean independent" si $E(\delta_k/\delta_1, ..., \delta_{k-1}) = E(\delta_k)$ pour tout k.

Propriété 3.1.

*Soit $\delta_1, \delta_2, ...$, dans cet ordre obtenu dans un calcul itératif par le mode **nearness**,*

- *Les δ_k sont des variables aléatoires centrées.*
- *Pour tout k on a $E(\delta_k/\delta_1, ..., \delta_{k-1}) = E(\delta_k) = 0$.*

Théorème 3.1.

Soient $y = a^\top b$, avec $a, b \in \mathbb{R}^n$, Y la variable aléatoire de la distribution du résultat obtenu en arrondissant y par le mode ***nearness****. Alors*

$$V(Y) \leqslant y^2 \left(K^2(1+u)^n - 1\right) = y^2((K^2-1) + nuK^2) + O(u^2),$$

avec K est le conditionnement de $y = \sum_{i=1}^{n} a_i b_i$.

Idée de la preuve

Utiliser le résultat obtenu par N. J. Higham,

$$Y = \sum_{i=1}^{n} a_i b_i \prod_{k=1}^{n} (1 + \delta_{k,i}).$$

Théorème 3.2.

On considère le système linéaire suivant $Ax = b$, avec $b \in \mathbb{R}^n$et $A \in \mathbb{R}^{n\times n}$ est une matrice inversible triangulaire inférieure. Le système est résolu par substitution avec le mode ***nearness****, la solution X satisfait pour tout $i = 1 : n$*

$$V(X_i) \leqslant x_i^2 \left[(1+u)^{i+1} \left(K_i^2 + V\left(\frac{\sum_{j=1}^{i-1} |a_{ij}| X_j}{a_{ii} x_i}\right)\right) - 1 \right],$$

avec K_i est le conditionnement de $a_{ii}x_i = b_i - \sum_{j=1}^{i-1} a_{ij}x_j$.

Nombre de chiffres significatifs

Théorème 3.3.

Soit $y = a^\top b$, *avec* $a, b \in \mathbb{R}^n$, *en appliquant le mode* ***nearness*** *on a*

$$\mathbb{P}\left(\frac{|Y - y|}{|y|} \leqslant 2^{-\alpha_\theta}\right) \geqslant \theta,$$

avec $0 < \theta < 1$ *et* $\alpha_\theta = -\log_2\left(K \exp\left(u\sqrt{2n \ln \frac{2}{1-\theta}}\right) - 1\right)$ *est le nombre des chiffres significatifs de* $\frac{|Y-y|}{|y|}$ *avec probabilité* θ.

Idée de la preuve

- Les martingales.
- Inégalité de Azuma-Hoeffding

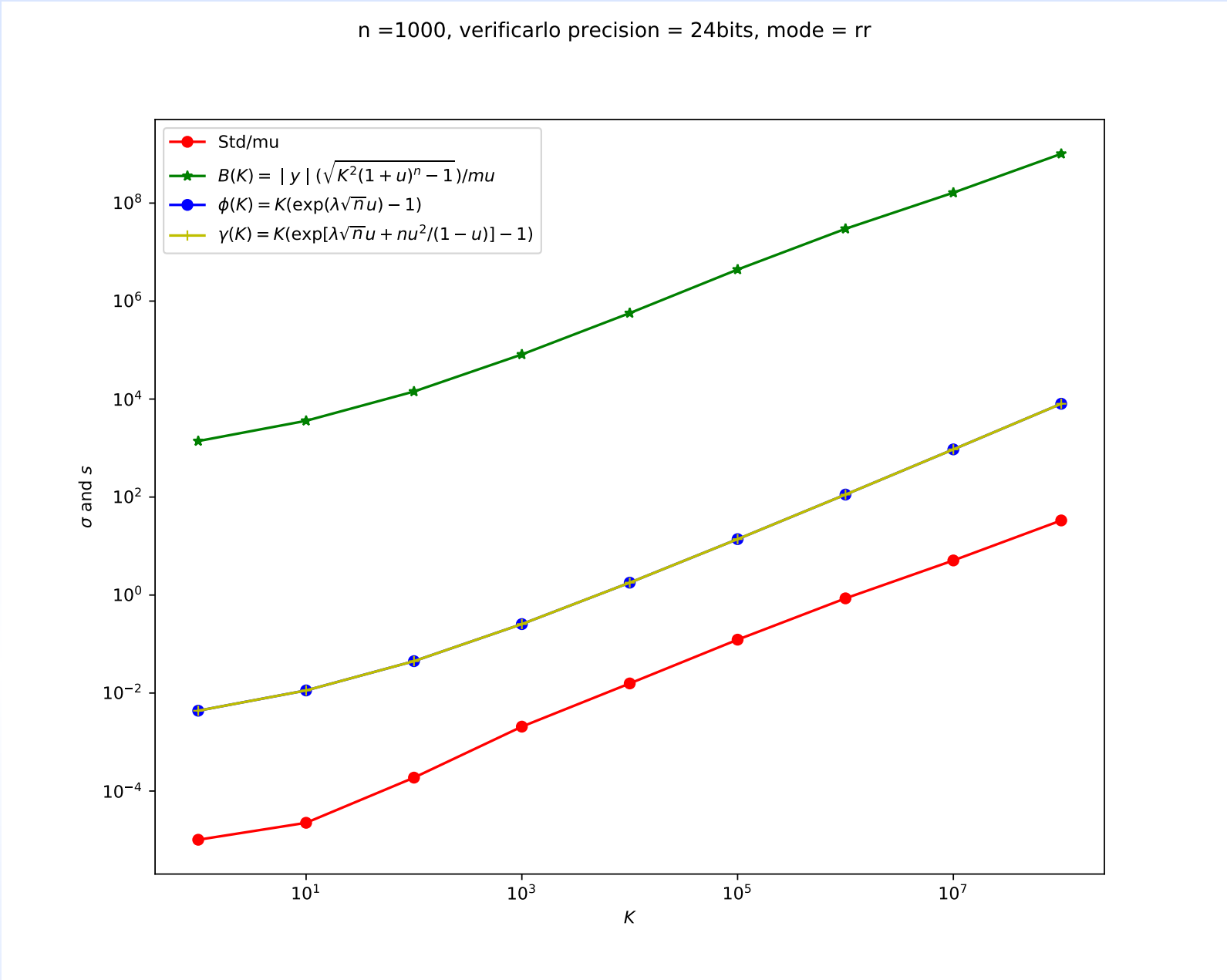


FIGURE – Le nombre de chiffres significatifs pour le produit scalaire avec 29 exécutions et $\lambda = 1.6651$.

Conclusion

Étapes réalisées

- Étude du biais.
- Borne de la variance.
- Borne probabiliste.

Conclusion

Étapes réalisées

- Étude du biais.
- Borne de la variance.
- Borne probabiliste.

Étape suivante

- Les effets des opérations élémentaires sur l'espérance et la variance.