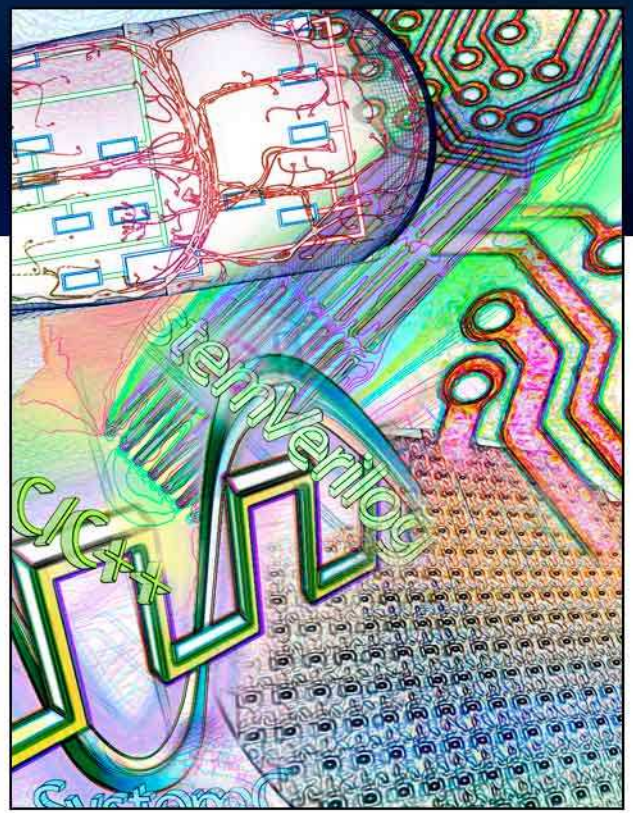
Méthodes d'analyse de la variabilité des circuits intégrés



Hubert FILIOL

Ingénieur R&D

Mentor Graphics, Eldo-AMS

Journée thématique GDR - Méthodes et Outils pour la Prise en Compte de la Variabilité des Procédés de Fabrication

5 Octobre 2012

Mentor Graphics®

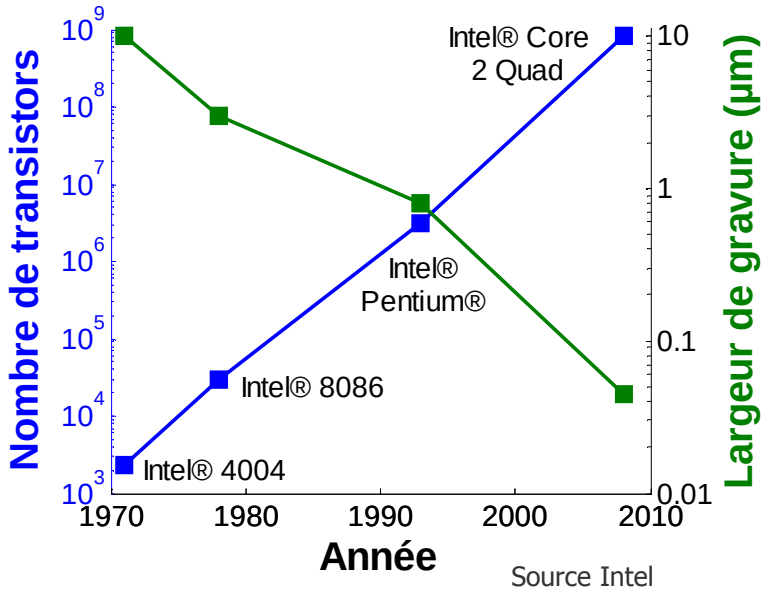
Plan

1. Loi de Moore et phénomènes de variabilité
2. Méthodes de caractérisation de la variabilité
3. Méthodes d'analyse de sensibilité

LOI DE MOORE ET PHÉNOMÈNES DE VARIABILITÉ

Evolution de la microélectronique

- Loi de Moore



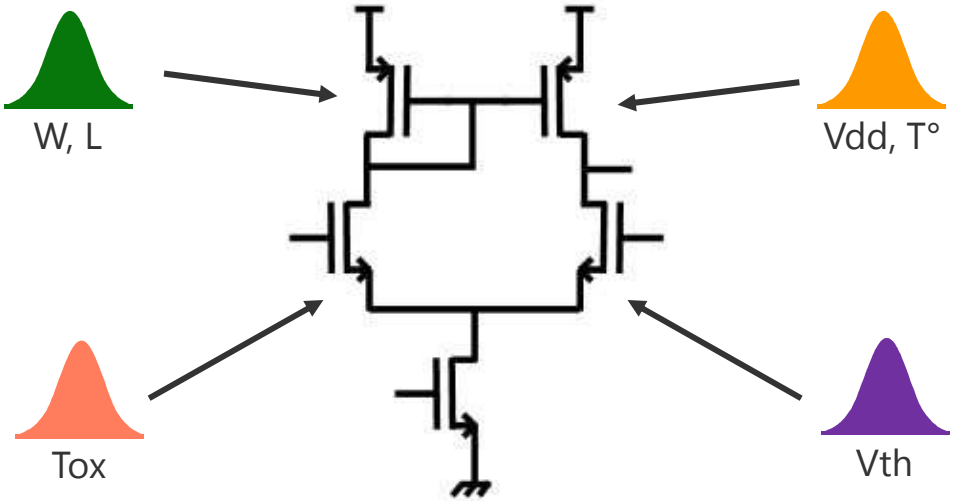
Source Intel

- Intérêts
  - Réduction de l'encombrement
  - Nouvelles fonctionnalités
  - Baisse du coût unitaire

- Limites physiques
  - Lithographie
  - Courants de fuite
  - Puissance dissipée
  - Fiabilité
  - …
  - **Variabilité**

Phénomènes de variabilité

- Variations des caractéristiques des circuits



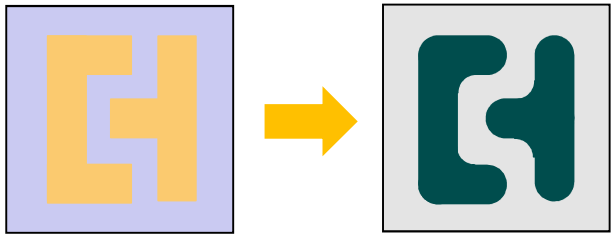
- Sources des variations

Variations Paramétriques	Variations Environnementales	Vieillissement
Dimensions **Concentration des dopants**	**Tension d'alimentation** **Température**	**Electromigration** **Injection de porteurs chauds**

Classification des variations

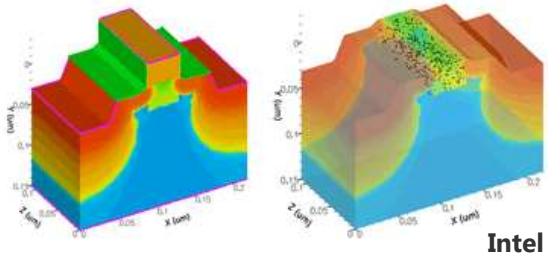
- Nature

Variations Systématiques



Ex : lithographie, dessin des masques

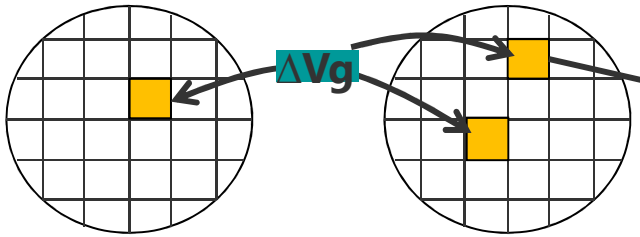
Variations Aléatoires



Ex : Fluctuation des dopants

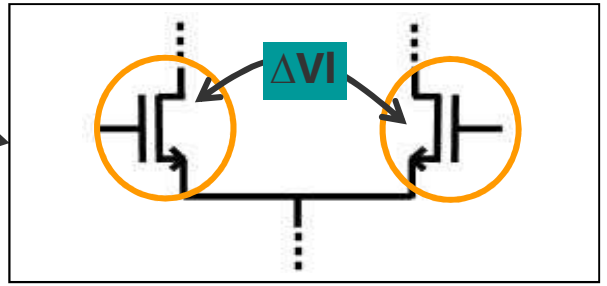
- Répartition spatiale

Variations Globales



Inter-puces, inter-wafers, inter-lot, etc.

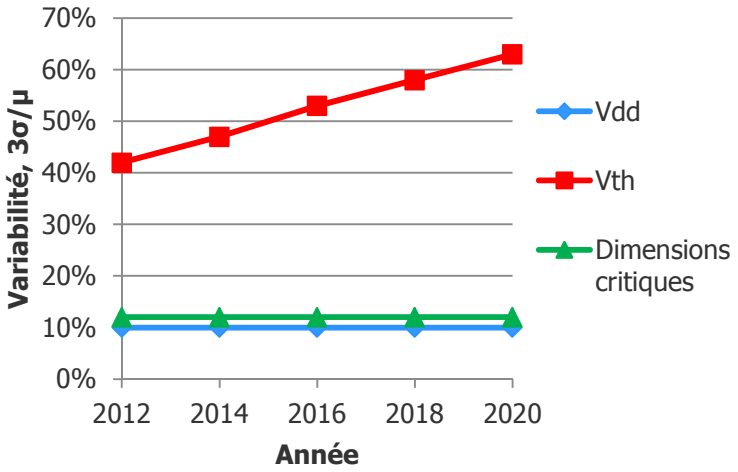
Variations Locales

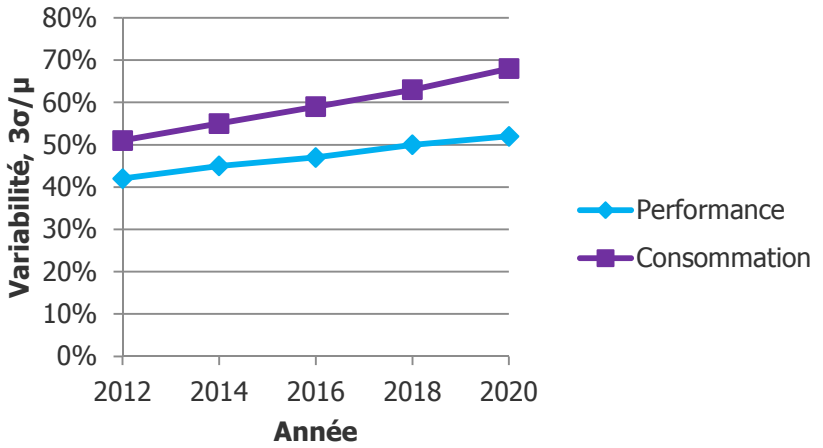


Intra-puce, mismatch

Evolution de la variabilité

- Prévision de l'ITRS (International Technology Roadmap for Semiconductors)

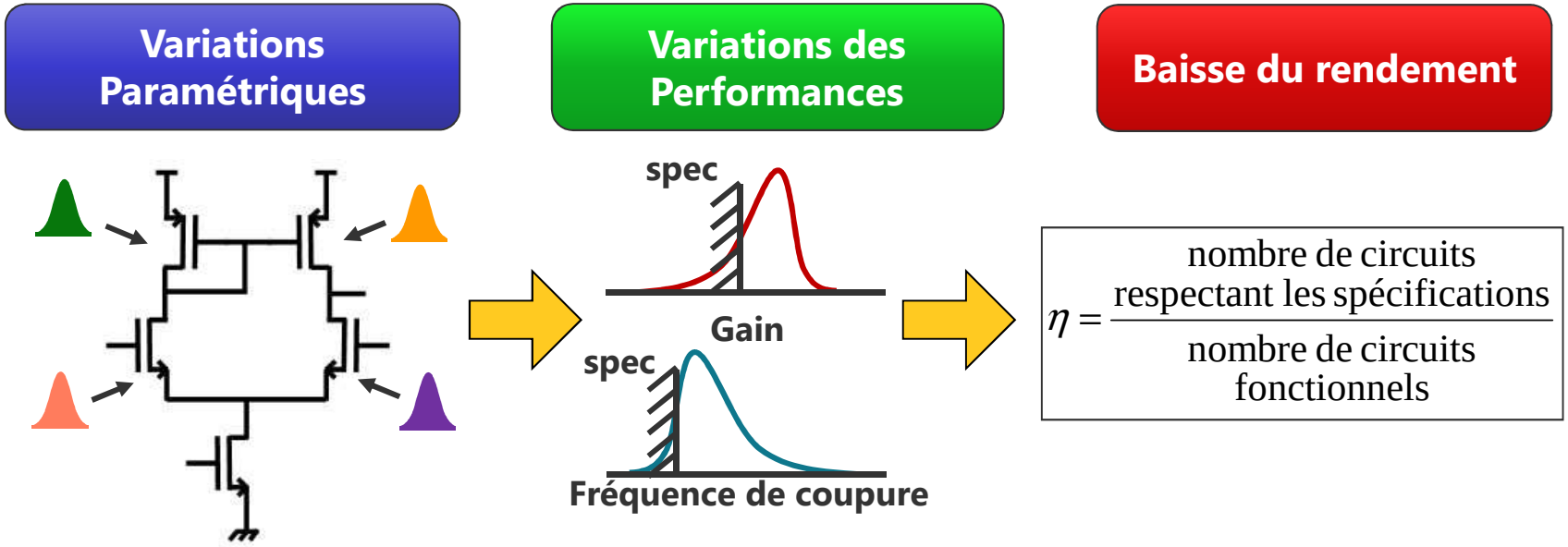




- Bilan
  - — ↗ de la variabilité avec la réduction des dimensions
  - — Variations systématiques
    - – Techniques lithographiques de compensation (OPC, PSM, etc.)
  - — Variations aléatoires (RDF)
    - – Ne peuvent pas être compensées lors de la fabrication
    - – Doivent être prises en compte dès le début du flot de conception

Problématique

- Baisse du rendement

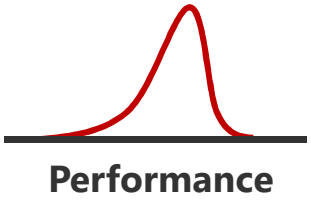


➔ Méthodes d'analyse de la variabilité

— Analyser l'impact des variations paramétriques sur les performances des circuits

Méthodes d'analyse de la variabilité

Caractériser les variations des performances

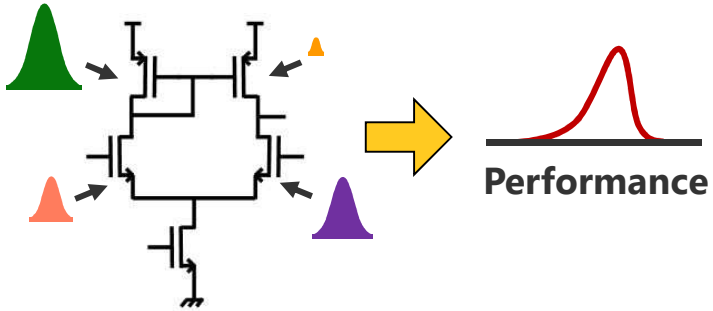


- Variable aléatoire
  - Densité de probabilité
  - Moments
  - …

➔ **Caractérisation de la variabilité = Estimation d'une variable aléatoire**

  - Histogramme
  - Variance

Identifier les principales variations paramétriques



- Quelles variations impactent le plus les performances ?

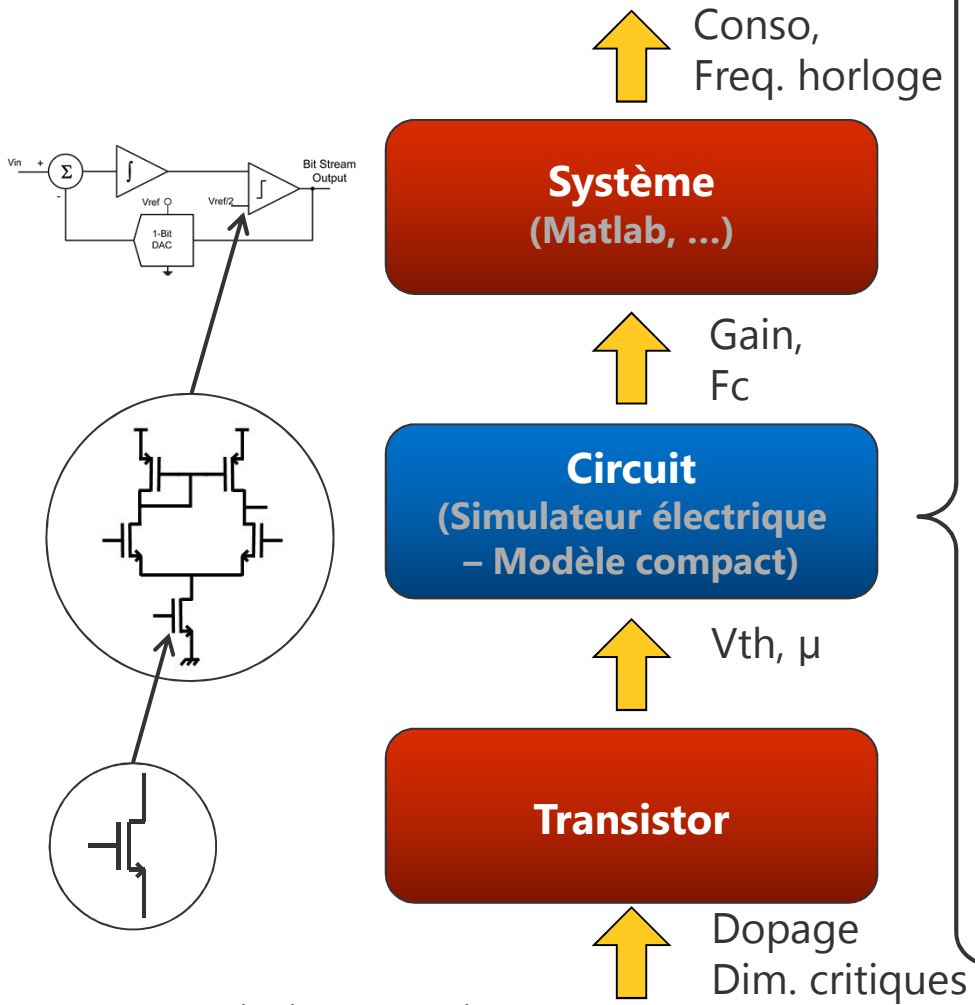
➔ **Analyse de sensibilité**

  - Indices de sensibilité

MÉTHODES DE CARACTÉRISATION DE LA VARIABILITÉ

Modélisation des performances des circuits

- Niveaux de modélisation



- Modèle des performances :

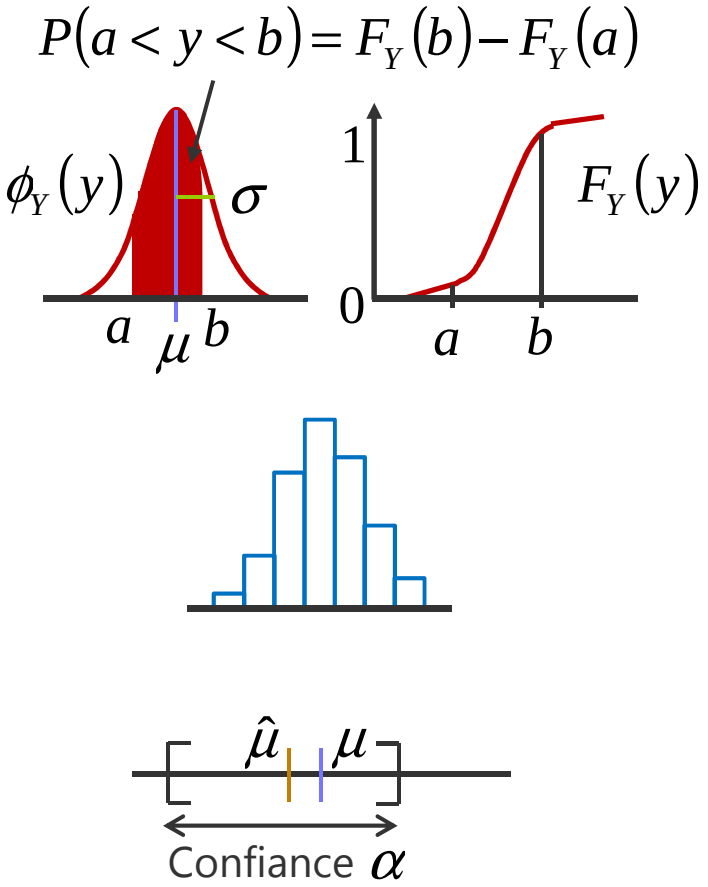
$$Y = f(X_1, X_2, \ldots, X_p)$$

- f : Simulation électrique
  - Modèle compact (BSIM, …)
  - Spice, Eldo, …
- X_i : Variations paramétriques
  - définies par la carte modèle d'une technologie
  - Variables aléatoires
    - gaussiennes / uniformes
    - corrélées / indépendantes

➔ Y = Variable aléatoire

Variables aléatoires – Estimation des variations des performances

- Variable aléatoire définie par :
  - Densité de probabilité / Fonction de répartition
  - Moments (espérance, variance, ...)
- Estimation des paramètres d'une variable aléatoire + Intervalle de confiance (précision)
  - Densité de probabilité ⇔ Histogramme
  - Espérance ⇔ Moyenne arithmétique



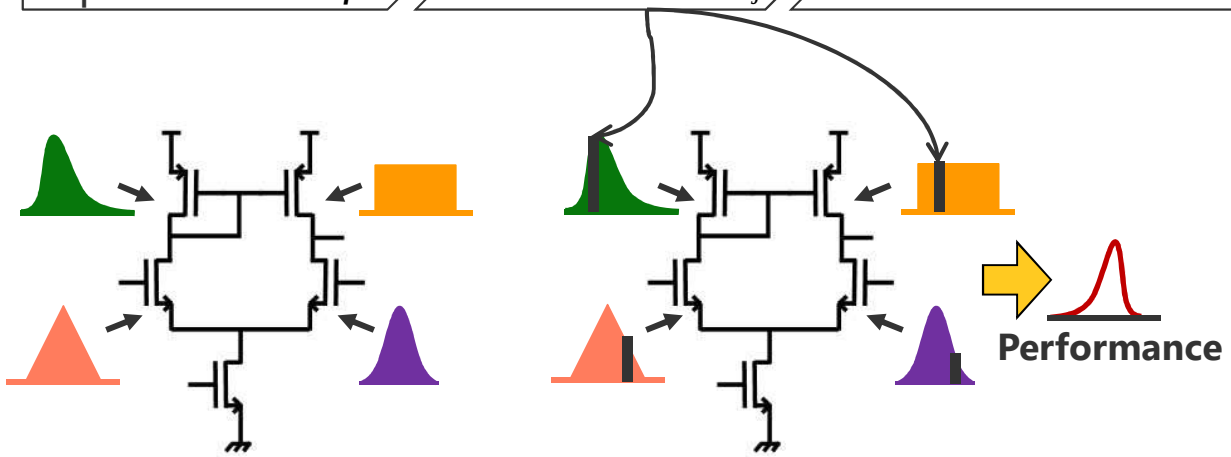
Méthode de Monte Carlo

- Loi des variations paramétriques
- Modèle de la performance f

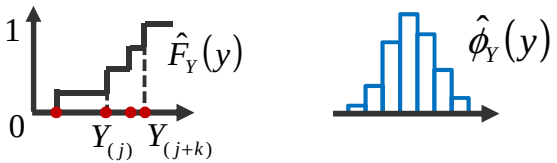
M vecteurs aléatoires $(X_1,\ldots,X_p)_j$

M simulations $Y_j = f\left((X_1,\ldots,X_p)_j\right)$

Tri croissant des $Y_j : Y_{(j)}$



- Fonction de répartition, Histo



- Moyenne, variance estimée :

$$\hat{\mu}_Y = \frac{1}{M}\sum_{j=1}^{M} Y_j,\ \hat{\sigma}_Y^2 = \frac{1}{M-1}\sum_{j=1}^{M}\left(Y_j - \hat{\mu}_Y\right)^2$$

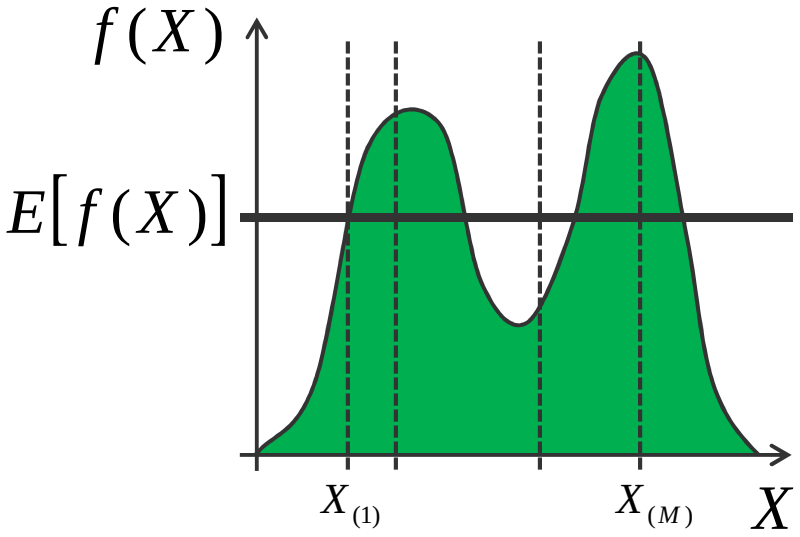
Hypothèses	Points +	Limites
• Nombre de tirages M suffisamment grand	• Générale – f non-linéaire – lois asymétriques • Précision indép. dimension	• Très lente – ↗ M pour ↗ précision

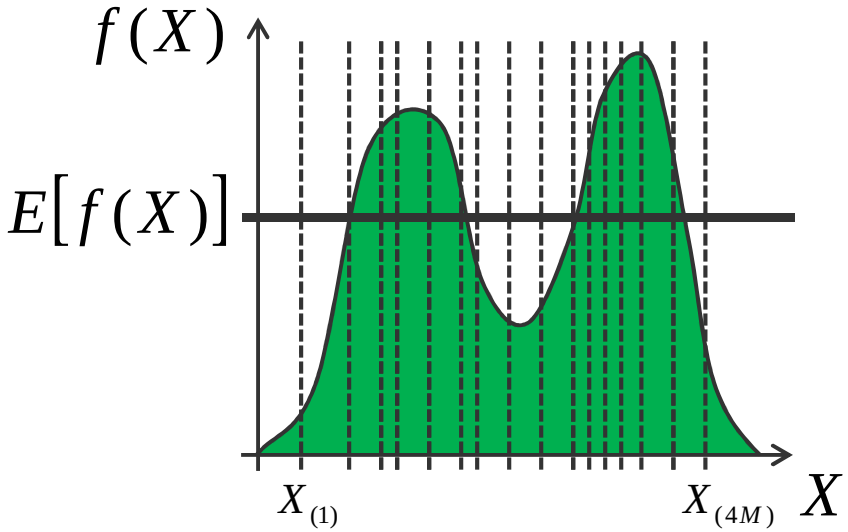
Monte Carlo – Réduction de la variance

- Soient une variable aléatoire X et une fonction $f : \Re \to \Re$
- Objectif : calculer l'espérance $\mu = E\left[f(X)\right]$
  - Générer M variables aléatoires $(X_j)_{1 \le j \le M}$ indépendantes et de même loi que X
  - Estimateur MC de l'espérance : $\hat{\mu}_{MC} = \frac{1}{M}\sum_{j=1}^{M} f(X_j)$
- Loi des grands nombres : $\lim_{M \to +\infty} \hat{\mu}_{MC} = \mu$
- Théorème central limite : $M \to +\infty, \frac{\hat{\mu}_{MC} - \mu}{\sigma / \sqrt{M}} \Rightarrow N(0,1) \ \ avec\ \sigma^2 = Var\left[f(X)\right]$
- Estimateur de la variance : $\hat{\sigma}^2 = \frac{1}{M-1}\sum_{j=1}^{M}\left(f(X_j) - \hat{\mu}_{MC}\right)^2$
- Intervalle de confiance à 95% : $\left[\hat{\mu}_M - \frac{1.96\,\hat{\sigma}}{\sqrt{M}}, \hat{\mu}_M + \frac{1.96\,\hat{\sigma}}{\sqrt{M}}\right]$
- Améliorer la précision de l'estimation : $\nearrow M$ ou $\searrow \sigma$

MC - Augmentation du nombre d'échantillons

- intervalle de confiance de l'estimation proportionnel à $1/\sqrt{M}$



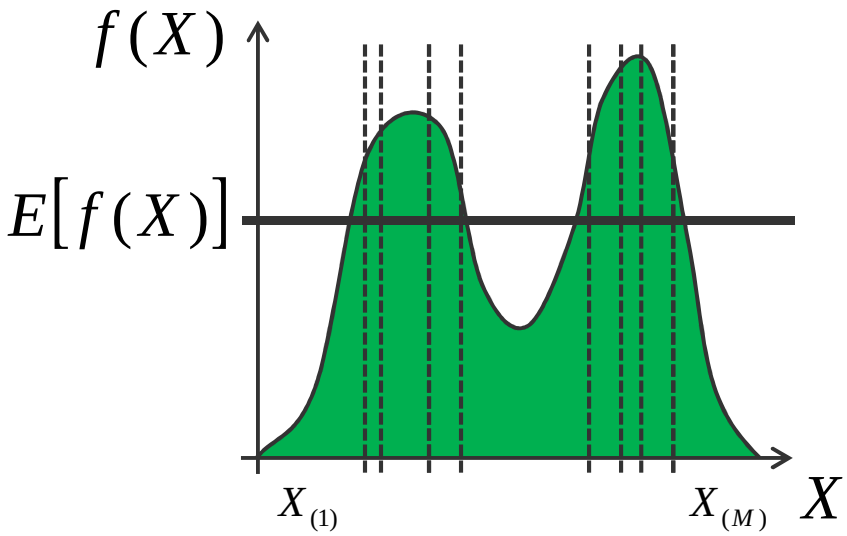


- Pour diviser l'erreur par 2 il faut 4 x plus d'échantillons

➔ Coût calcul

MC - Echantillonnage préférentiel (1)

- Réduction de la variance
  - Echantillonner en priorité dans les régions d'importance de la fonction



MC - Echantillonnage préférentiel (2)

- Principe : modification de l'échantillonnage
  - Soit une densité de probabilité $\tilde{\phi} > 0$
  - Nouvelle écriture de l'espérance à estimer :

$$\mu = E[f(X)] = \int \frac{f(x)\phi_X(x)}{\tilde{\phi}(x)} \tilde{\phi}(x)dx = E\left[\frac{f(Z)\phi_X(Z)}{\tilde{\phi}(Z)}\right]$$

Où Z est une variable aléatoire de densité de probabilité $\tilde{\phi}$

- Variance :

$$Var\left[\frac{f(Z)\phi_X(Z)}{\tilde{\phi}(Z)}\right] = E\left[\frac{f(Z)^2\phi_X(Z)^2}{\tilde{\phi}(Z)^2}\right] - \mu^2 = \int \frac{f(x)^2\phi_X(x)^2}{\tilde{\phi}(x)}dx - \mu^2$$

➔ Annulation de la variance si $\tilde{\phi}(x) = \frac{f(x)\phi_X(x)}{\mu}$

Ce que l'on veut estimer !

- Heuristique :
  - $\tilde{\phi}(x) \propto f(x)\phi_X(x)$

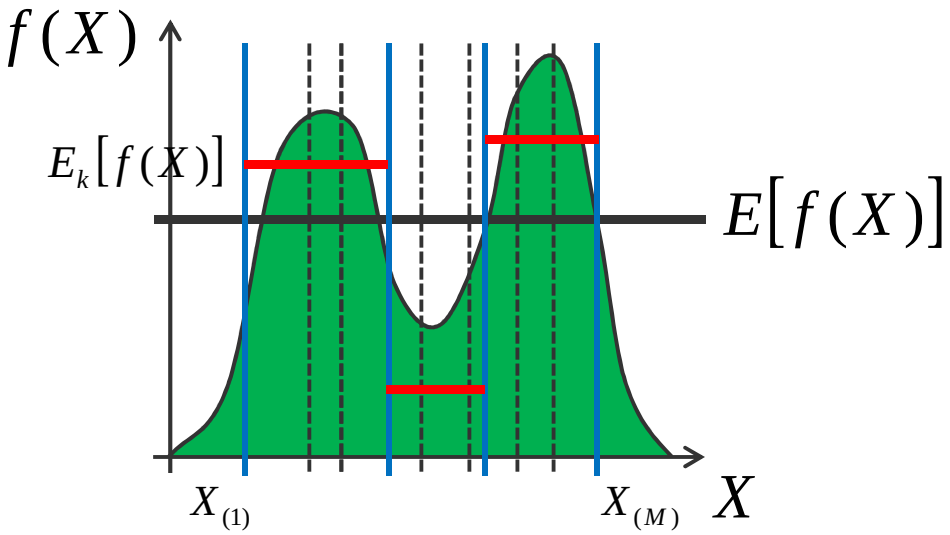
MC - Echantillonnage préférentiel (3)

- Algorithme :
  - Choisir une fonction d'importance $\tilde{\phi}$
    - facile à générer
    - $\tilde{\phi}(x) \propto f(x)\phi_X(x)$
    - Approche simple → $\tilde{\phi}$ = la distribution originale ϕ mais centrée sur le maximum de $f(x)\phi_X(x)$
  - Générer M variables aléatoires $(Z_j)_{1\le j\le M}$ indépendantes de densité de probabilité $\tilde{\phi}$
  - Estimateur de l'espérance :

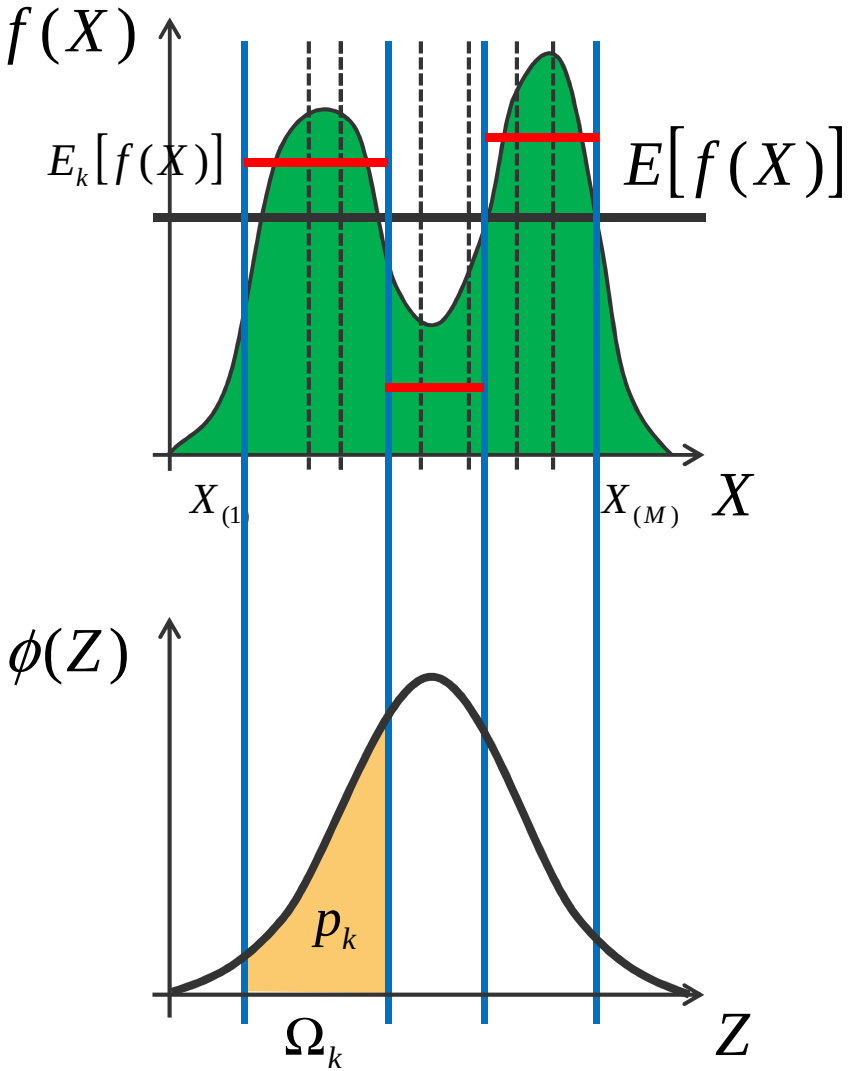
$$\hat{\mu}_{pref} = \frac{1}{M}\sum_{j=1}^{M} \frac{\phi_X(Z_j)}{\tilde{\phi}(Z_j)} f(Z_j)$$

MC - Echantillonnage stratifié (1)

- Réduction de la variance :
  - Partitionner le domaine
  - Estimer l'espérance sur chaque sous-domaine avec l'estimateur MC
  - Pondérer les estimations locales par une probabilité pour obtenir l'estimation globale de l'espérance



MC - Echantillonnage stratifié (2)



- Algorithme
  - Partitionner l'espace de Z
  - Calcul de la probabilité p_k de chaque strate
  - Génération des variables aléatoires $(X \mid Z \in \Omega_k)$
  - Calcul de $f(X) \mid Z \in \Omega_k$
- Espérance

$$\mu = E[f(X)] = \sum_{k=1}^{K} p_k \; E[f(X) \mid Z \in \Omega_k]$$

- Estimateur

Estimateur MC

$$\hat{\mu}_{strat} = \sum_{k=1}^{K} p_k \frac{1}{m_k} \sum_{j=1}^{m_k} f(X_j^{(k)})$$

MC - Echantillonnage stratifié (3)

- Choix des strates Ω_k :
  - Dans chaque strate, la variabilité de f doit être faible

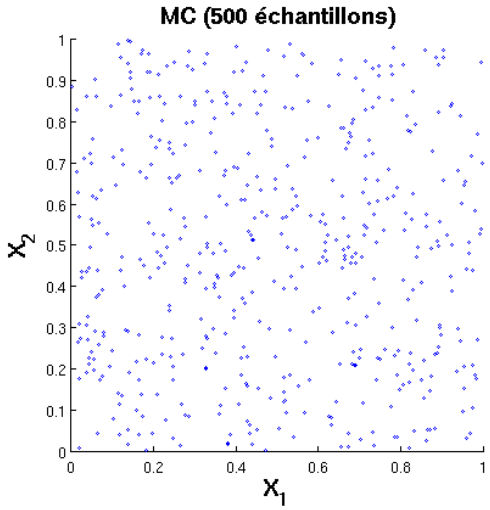
- Choix du nombre de tirages m_k dans chaque strate :
  - Allocation proportionnelle $m_k = p_k M$

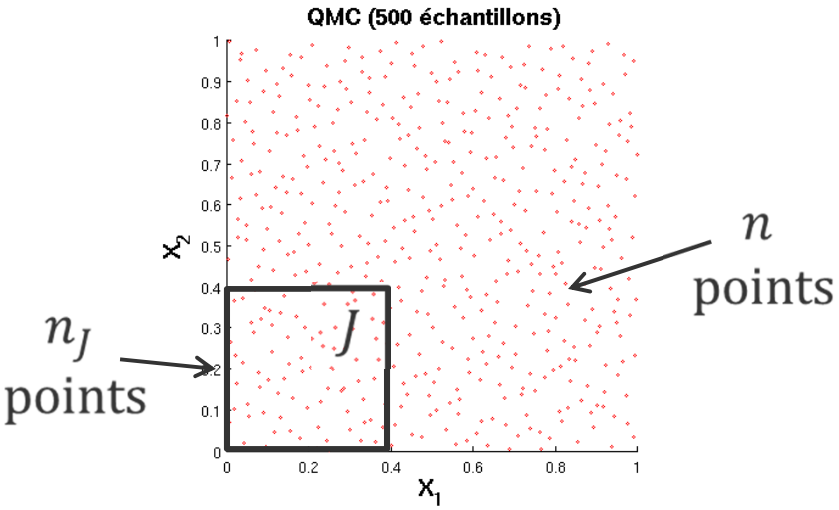
$$Var[\hat{\mu}_{MC}] - Var[\hat{\mu}_{strat}] = \frac{1}{M}\sum_{k=1}^{K} p_k \left(E[f(X|Z \in \Omega_k)] - \mu\right)^2 > 0$$

➔ Variance toujours plus faible que l'estimateur MC classique

Quasi Monte Carlo (1)

- Améliorer l'estimation avec des tirages plus « uniformes »





- Mesure d'uniformité : la discrépance

$$D_n = sup_J \left| \frac{n_j}{n} - Vol(J) \right|$$

Proportion de points dans J

Proportion du volume occupé par J

Quasi Monte Carlo (2)

- Majoration de l'erreur commise sur l'estimation (inégalité de Koksma-Hlawka) :

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - E[f(X)] \right| \leq c(f) D_n(X_i)$$

Discrépance

- Suites à faible discrépance $D_n(X_i)$:
  - Suites de Sobol, Halton, Faure, …

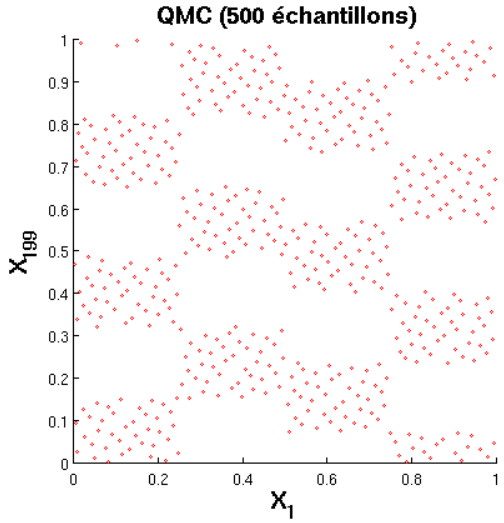
- Erreur

Dimension

$$\text{QMC : } O\left(\frac{\log(n)^d}{n}\right) < \text{MC : } O\left(\frac{1}{\sqrt{n}}\right)$$

Quasi Monte Carlo (3)

- Suites déterministes :
  - Impossible d'appliquer le Théorème Central Limite (pour calculer des intervalles de confiance)
- Problème en grande dimension

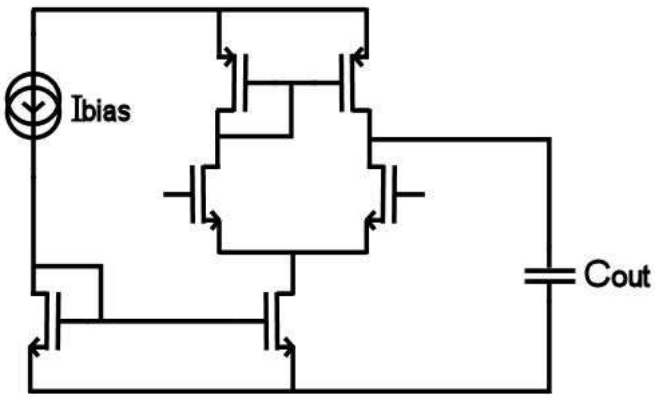


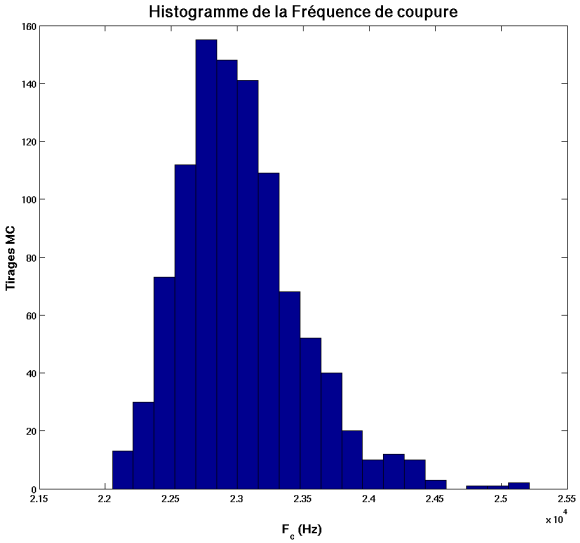
200 dimensions
Projection
sur X1 et X199

  - Besoin de beaucoup de points pour assurer une bonne répartition uniforme dans l'espace

Circuit de test – comparaison des méthodes

- Amplificateur à transconductance (CMOS 65nm)
  - 5 variations paramétriques
  - Fréquence de coupure
- Estimation de l'espérance
  - 100 tirages aléatoires
- Variance de l'estimateur
  - 50 essais
- Estimateurs
  - MC classique
  - MC préférentiel (loi normale)
  - MC stratifié (3 strates, alloc propor.)
  - QMC (suite de Halton)





Circuit de test – comparaison des méthodes

- Résultats (100 tirages MC, 50 essais)

Méthode	Moyenne	Variance de la moyenne	Var(MC) / Var
MC classique	22 997,8 Hz	2 198,3 Hz	1
MC préférentiel	22 997,5 Hz	2 070,4 Hz	1,06
MC stratifié	23 009,8 Hz	1 417,5 Hz	1,55
QMC	22 996,0 Hz	942, 1 Hz	2,33

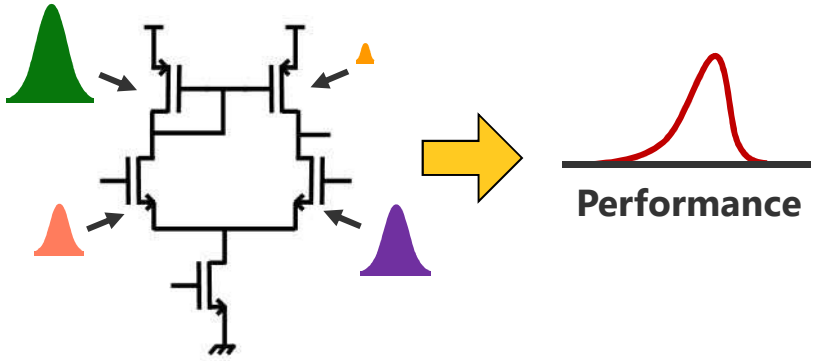
- Réduction de variance (faible)

  ➔ Rechercher les paramètres optimaux des méthodes de réduction de variance

MÉTHODES D'ANALYSE DE SENSIBILITÉ

Méthodes d'analyse de la sensibilité

- Impact des variations paramétriques sur une performance du circuit ?



- Objectifs :
  - Identifier les composants du circuits dont les variations influencent le plus la performance
  - Modélisation/Simulation avec un jeu réduit de variations paramétriques (QMC)

Analyse de sensibilité en grandes dimensions

- Contexte
  - ↗ du nombre de variations paramétriques (p = 100~1000)
- Problématique
  - Coût CPU élevé, nombre limité de simulations MC : $\boxed{p >> M}$
  - ➔ Un nombre réduit p_{imp} de variations sont vraiment influentes
- Objectif
  - Modèle linéaire avec seulement les variations influentes

$$Y = \underbrace{\beta_s X_s + \ldots + \beta_u X_u}_{p_{imp}\ \text{termes}} = X_{imp}^T \beta_{imp}$$

- Minimisation des moindres carrés

$$\min_{\beta} \sum_{k=1}^{M} \left(Y_k - \mathbf{X}_k \beta\right)^2 = \min_{\beta} \left\| Y - \mathbf{X}\beta \right\|^2 \quad \rightarrow \quad \hat{\beta} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T Y$$

- Problème : $\boxed{p \leq M}$

➔ Solution : régularisation par pénalisation L1

Pénalisation L1 - Régression LASSO

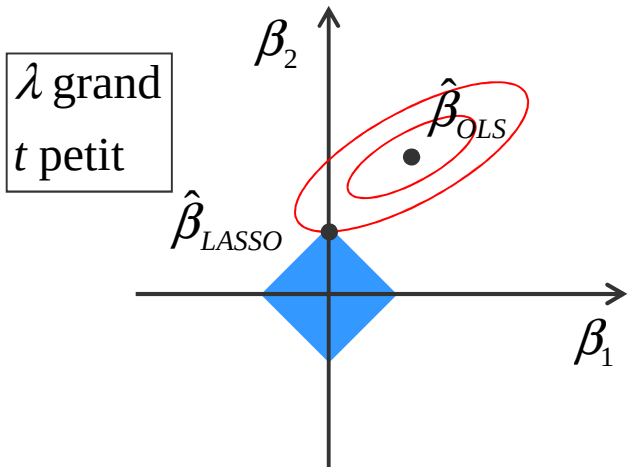
- Pénalisation L1 d'un problème de régression linéaire (LASSO)

$$\min_{\beta} \left\| Y - \mathbf{X}\beta \right\|_2^2 \quad \Leftrightarrow \quad \min_{\beta} \left\| Y - \mathbf{X}\beta \right\|_2^2 + \lambda \left\| \beta \right\|_1$$

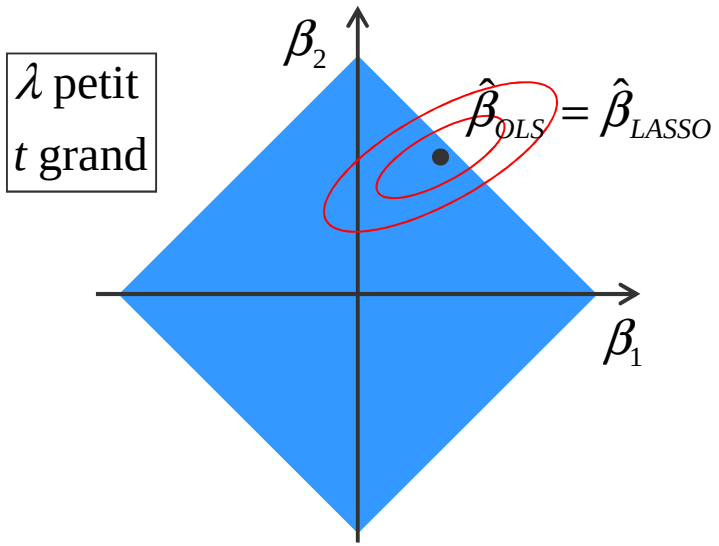
$$\text{tel que} \sum_{i=1}^{p} |\beta_i| \leq t$$

Paramètres de régularisation

$$t = f(\lambda)$$



➔ Seuillage de β_1 à 0

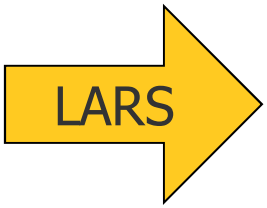


➔ Même solution que les moindres carrés

Algorithme LARS (Least Angle regression Stagewise)

- Construction itérative du modèle linéaire
  - Ajout des variables influentes une par une

LASSO $\min_{\beta} \left\| Y - \hat{Y} \right\|_2^2 + \lambda \left\| \beta \right\|_1$ avec $\hat{Y} = X\beta$



$$\begin{array}{ll} i = 0 & \hat{Y} = 0 \\ i = 1 & \hat{Y} = \hat{Y} + \hat{Y}_1 \\ i = 2 & \hat{Y} = \hat{Y} + \hat{Y}_2 \\ \vdots & \vdots \\ i = K & \hat{Y} = \mathbf{X}_{imp}\beta_{imp} \end{array} \qquad \lambda \searrow$$

- Itération LARS
  1. Trouver la variable à ajouter (la plus corrélée avec les résidus)
  2. Calculer la direction de descente
  3. Calculer le pas de descente

Algorithme LARS – Conditions d'arrêt

- Pas de critère d'arrêt
  - Cas 1 : Nombre de variables ≤ Nombre de simulations MC $(p \leq M)$
    - ➔ Modèle linéaire complet (= moindres carrés) $Y = \underbrace{\beta_s X_s + \ldots + \beta_u X_u}_{p \text{ termes}}$ $\longrightarrow$ Influence décroissante
  - Cas 2 : Nombre de variables > Nombre de simulations MC $(p > M)$
    - ➔ Modèle linéaire réduit $Y = \underbrace{\beta_s X_s + \ldots + \beta_u X_u}_{M \text{ termes}}$ $\longrightarrow$ Influence décroissante
- Seuil sur les résidus
  - Cas 3 : $\left(\left\| Y - \hat{Y} \right\|_2^2 < \gamma \right)$
    - ➔ Modèle linéaire réduit $Y = \underbrace{\beta_s X_s + \ldots + \beta_u X_u}_{p_{imp} \text{ termes}}$ $\longrightarrow$ Influence décroissante

Analyse de sensibilité de Sobol vs LARS (1)

- Circuit de test
  - 1864 variations paramétriques
- Analyse de sensibilité de référence
  - Méthode de Sobol (10000 simulations Monte Carlo, 8h09min)
- Sortie Eldo

Indices de sensibilité

6 variations importantes !

Localisation des variations paramétriques

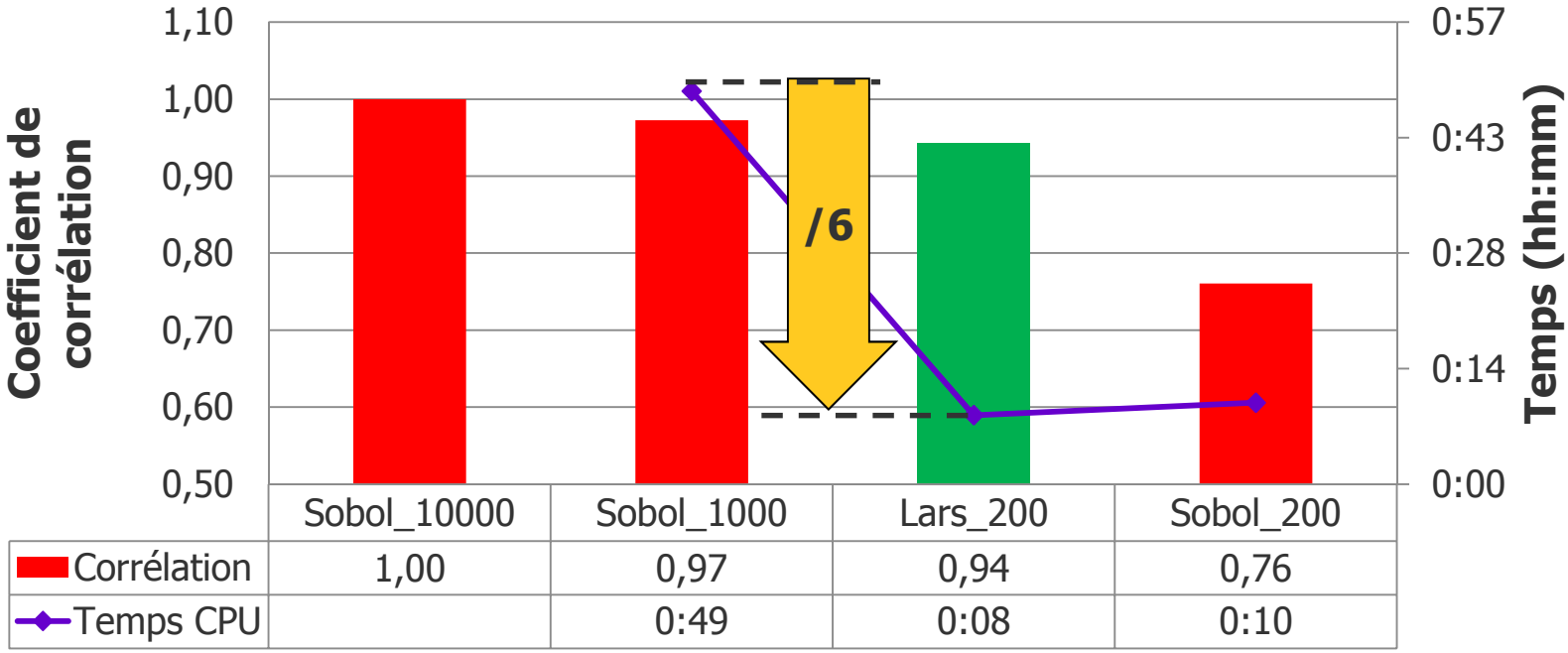
```
Estimation of Sensitivity Indices for Response Y : MAX(W(ISAP))
  Mode of Computation        : accurate
  Number of Runs             : 10001
  Number of Random Variables : 1864

  List of Important Parameters :
  Index(I)    Histo S(I)         (percent / cumulative)    Variable
   Value                                                    Name

    696     |------------------   15.73%  15.73%      PEM(PSELECT,XICOMP1V55.XI3.XM1.M1)
    695     |-----------------    14.09%  29.82%      PEM(PSELECT,XICOMP1V55.XI3.XM2.M1)
   1864     |-----------------    13.86%  43.69%      P(TEMPPAR)
      1     |----------------     12.59%  56.27%      P(PWR)
    117     |-----------          10.19%  66.47%      PEM(PSELECT,XICOMP1V55.XI3.XM0.M1)
    118     |----------            8.60%  75.06%      PEM(PSELECT,XICOMP1V55.XI3.XM15.M1)
    599     |-                     0.97%  76.03%      PEM(PSELECT,XICOMP.XMREF.XM1.M1)
    694     |                      0.44%  76.47%      PEM(PSELECT,XICOMP1V55.XI3.XM32.M1)
   1495     |                      0.36%  76.83%      PEM(PSELECT,XICOMP1V55.XI1.XM18.M1)
     48     |                      0.35%  77.18%      PEM(PSELECT,XISA2.XISA.XMN2.XM2.M1)
    643     |                      0.32%  77.50%      PEM(PSELECT,XISA2.XISA.XMP1.XM1.M1)
    641     |                      0.26%  77.75%      PEM(PSELECT,XICONTROL.XI1.XM12.M1)
   1488     |                      0.25%  78.00%      PEM(PSELECT,XICOMP1V55.XI15.XM102.M1)
   1122     |                      0.25%  78.25%      PEM(PSELECT,XICOMP1V55.XI15.XM5.M1)
    644     |                      0.24%  78.49%      PEM(PSELECT,XISA1.XISA.XMP1.XM1.M1)
```

Analyse de sensibilité de Sobol vs LARS (2)

- Mesures pour différentes tailles d'échantillons MC
  - Coefficient de corrélation entre les indices de sensibilités
  - Temps CPU

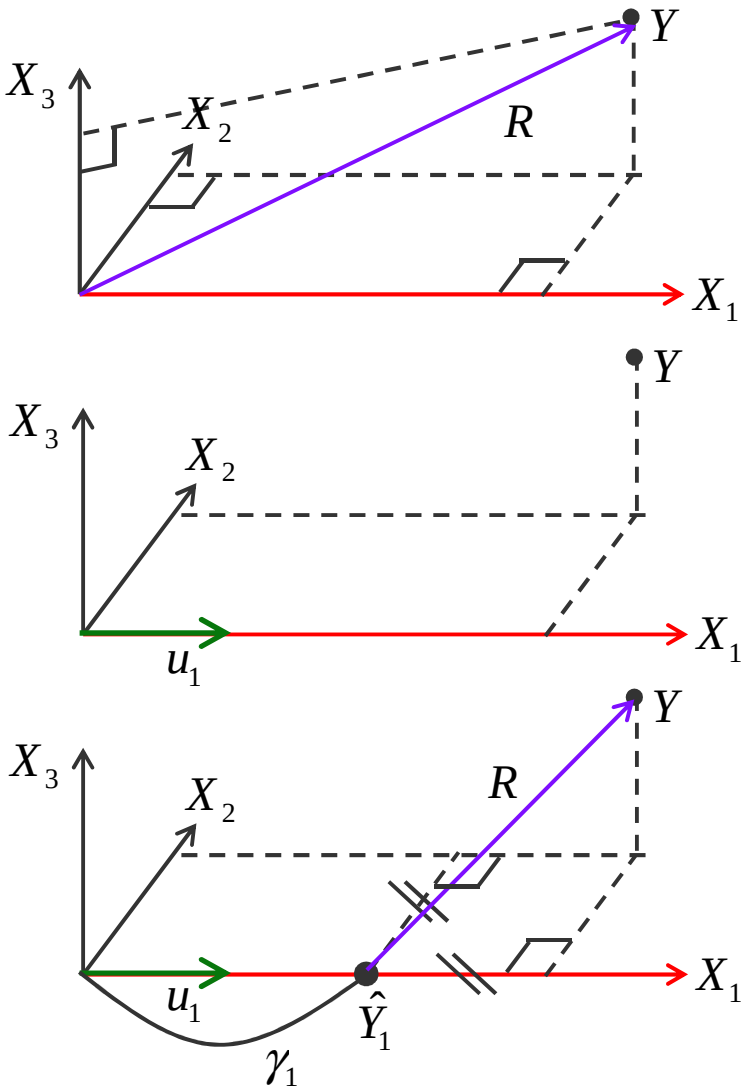


	Sobol_10000	Sobol_1000	Lars_200	Sobol_200
Corrélation	1,00	0,97	0,94	0,76
Temps CPU		0:49	0:08	0:10

➔ Lars_200 aussi précis que Sobol_1000 mais temps CPU divisé par 6


Mentor Graphics®
www.mentor.com

Algorithme LARS – Itération 1



- Recherche de la variable la plus corrélée avec les résidus :

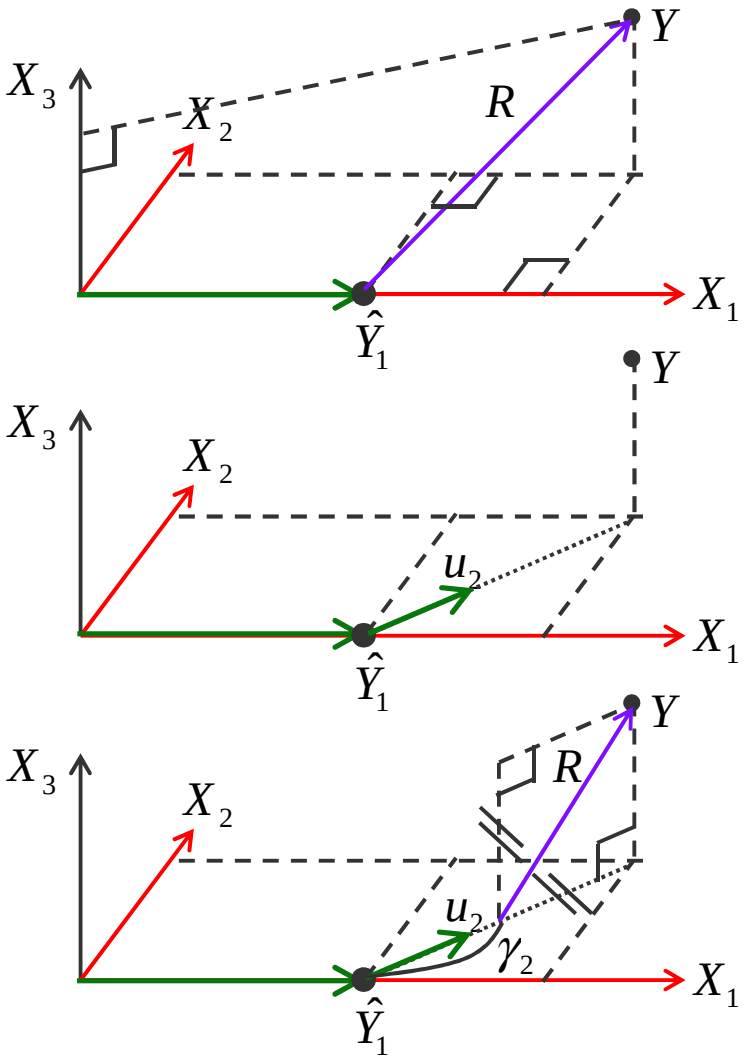
$$X^T\left(Y-\hat{Y}\right)=X^T R \quad \rightarrow \quad X_1$$

- Calcul de la direction u_1

- Calcul du pas γ_1
  - Valeur max jusqu'à ce qu'une variable inactive soit autant corrélée avec les résidus

$$\hat{Y}=\gamma_1 u_1$$

Algorithme LARS – Itération 2



- Recherche de la variable la plus corrélée avec les résidus :

$$X^T\left(Y-\hat{Y}\right)=X^T R \quad \rightarrow \quad X_2$$

- Calcul de la direction u_2
  - Vecteur équiangulaire entre les variables actives

- Calcul du pas γ_2
  - Valeur max jusqu'à ce qu'une variable inactive soit autant corrélée avec les résidus

$$\hat{Y}=\gamma_1 u_1+\gamma_2 u_2$$

Progression = compromis entre les variables les plus corrélées avec les résidus