Optimisation par algorithmes évolutionnaires

Quelques aspects pratiques et théoriques

Plan

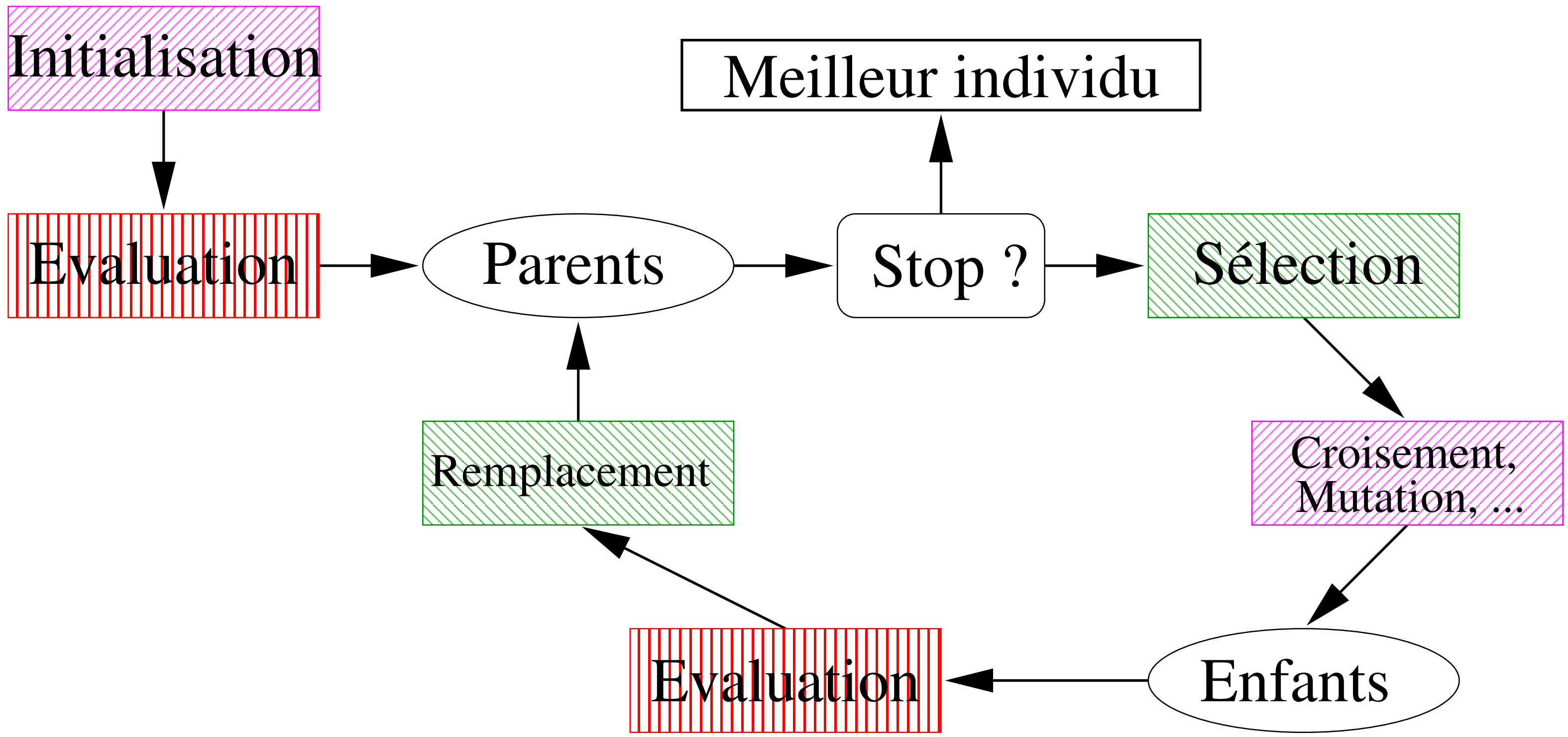
1. Introduction générale aux algorithmes évolutionnaires.
2. Eléments de théorie des algorithmes évolutionnaires.

I Les Algorithmes Evolutionnaires: Introduction générale

Algorithmes Evolutionnaires : les racines

- Algorithmes Génétiques
  J. Holland, 75, D.E. Goldberg 89
  AI and biology – US, east coast
- Strategies d'Évolution
  I. Rechenberg 65, 73, H.P. Schwefel 65, 81
  Engineers – Germany
- Programmation Evolutionnaire
  L.J. Fogel, 66, D.B. Fogel, 91, 95
  Automata and time series – US, west cost
- Programmation Génétique
  J. Koza, 92

Algorithmes évolutionnaires



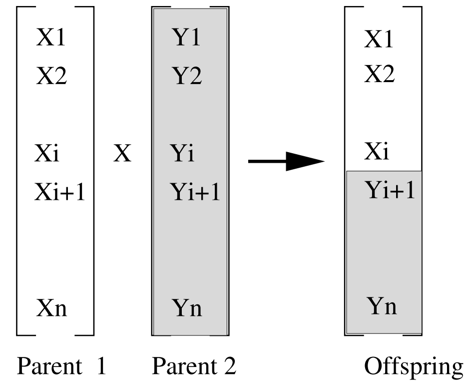
Initialisation et opérateurs stochastiques

Darwinisme (stochastique ou deterministe)

Le coût CPU

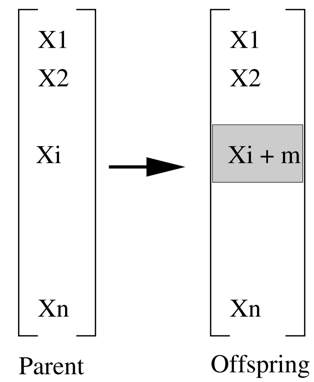
Quelques exemples d'opérateurs

- SELECTION: Quels sont les individus qui vont générer des enfants ?
  - exemple : Choisir aléatoirement avec un poids plus fort pour les “meilleurs”
- CROISEMENTS: Le croisement à un point



Quelques exemples d'opérateurs

- MUTATIONS



- REMPLACEMENT : Quels seront les parents de la génération suivante?
  - Exemple : Les meilleurs enfants remplacent les parents.

Le Dilemme:Exploitation vs Exploration

- Exploitation des bons individus.
  recherche locale: chercher dans le voisinage des meilleurs individus de la population.
- Exploration des zones inconnues de Ω.
  recherche globale: il faut pouvoir aller partout.

Excès d'exploitation $\Longrightarrow$

Convergence prématurée

enlisement dans un optimum local

Excès d'exploration $\Longrightarrow$

Pas de convergence

$\approx$ marche aléatoire

II Eléments de théorie des Algorithmes Evolutionnaires

Chaînes de Markov

Définition 1 $(X_0, ..., X_n, ...)$ *est une chaîne de Markov si:*

$$\mathbb{P}\left(X_{n+1} \in A | X_0 = x_0, ..., X_n = x_n\right) = \mathbb{P}\left(X_{n+1} \in A | X_n = x_n\right)$$

- “on peut se contenter de retenir du passé uniquement le présent”
- La chaîne est homogène si:
  $\forall n, \tilde{n} \in \mathbb{N}, \mathbb{P}\left(X_{n+1} \in A | X_n = x\right) = \mathbb{P}\left(X_{\tilde{n}+1} \in A | X_{\tilde{n}} = x\right)$
- Probabilité conditionnelle de transition de x à n’importe où dans $\mathbb{R}^N$: $\mathbb{P}\left(X_{n+1} \in A | X_n = x\right) = \int_A p(y, x_n) dy$
  où $p(y, x_n)$ est le noyau de transition associé X_n.

Noyau de transition

- Espace d'état $\mathcal{X}$ (typiquement $\mathbb{R}^N$)
- $\mathcal{B}$,σ-algèbre de sous ensembles de $\mathcal{X}$

Définition 2 *Un noyau de transition est une fonction P définie sur $\mathcal{B} \times \mathcal{X}$ telle que*

1. *P(.,x) est une mesure de probabilité,* $\forall x \in \mathcal{X}$
2. *P(A,.) est une fonction mesurable,* $\forall A \in \mathcal{B}$

pour un $x \in \mathcal{X}$ le noyau de transition détermine la probabilité conditionnelle de transition de x à n'importe où dans $\mathcal{X}$.

$$\mathbb{P}(X \in A|x) = \int_A p(y,x)dy, A \in \mathcal{B}$$

où $p(.,x)$ est la densité associée à P(.,x)

Cadre pour l'étude théorique

- Algorithmes évolutionnaires: $(1, \lambda)$-ES et $(1 + \lambda)$-ES
- Etude de la phase d'exploitation:
  On minimise la fonction ☺ sphère ☺ en dimension N:
  $\mathcal{F} : x = (x_1, ..., x_N) \longrightarrow \sum_{i=1}^{N} x_i^2$

Stratégies d'évolutions: (μ, λ)*-ES et* $(\mu + \lambda)$*-ES*

Minimisation de $\mathcal{F} : \mathbb{R}^N \longrightarrow \mathbb{R}$.

- Population de taille μ.
- Initialisation uniforme des X_i^0 pour $i = 1, ..., \mu$
- Succession des générations :
  Population X_i^n pour $i = 1, ..., \mu$
  - *Mutation* de X_i^n choisi aléatoirement:
    Création de λ enfants $X_j^{'}$ pour $j = 1, ..., \lambda$
  - *Evaluation* des $\mathcal{F}(X_j^{'})$ pour $j = 1, ..., \lambda$
  - *Sélection* déterministe des : X_i^{n+1} pour $i = 1, ..., \mu$
    Parmi les $\lambda X_j^{'}$ ....(μ, λ)-ES
    Parmi les μ X_i^n et λ $X_j^{'}$.........$(\mu + \lambda)$-ES

Cadre de cet exposé

- *Mutation*:
  - *Adaptative* gaussienne $\widetilde{X}^j_{n+1} = X_n + \mathcal{N}_j(0, \sigma_n)$ pour $j = 1..\lambda$
    - $\mathcal{N}_j(0, \sigma_n)$ variable aléatoire gaussienne
    - $\sigma_n := \sigma^\star r_n$ où $r_n = |X_n|$ et $\sigma^\star = cste$ (à déterminer)
  - *Auto adaptative* log normale
    $\widetilde{X}^j_{n+1} = X_n + \sigma_n \exp(\tau \mathcal{N}_j(0,1)) \widetilde{\mathcal{N}}_j(0,1)$ pour $j = 1..\lambda$

    $$\sigma_{n+1} = \sigma_n \exp(\tau \mathcal{N}_\star(0,1))$$

    où $\star$ est telle que
    $X^j_{n+1} = X_n + \sigma_n \exp(\tau \mathcal{N}_\star(0,1)) \widetilde{\mathcal{N}}_\star(0,1)$

Modélisation mathématique (1)

Modélisation du $(1, \lambda)$*-ES avec mutation adaptative:*

- $x \in \mathbb{R}^N$,$\mathcal{N}^1(0,1), ..., \mathcal{N}^\lambda(0,1)$, λ variables aléatoires gaussiennes centrées réduites indépendantes.
- $x^{'} = \arg\min\{\mathcal{F}(x + \sigma\mathcal{N}^1(0,1)), ..., \mathcal{F}(x + \sigma\mathcal{N}^\lambda(0,1))\}$
- $p_\sigma(x^{'}, x)$, noyau de transition pour le pas $x \longrightarrow x^{'}$
- Modélisation du $(1, \lambda)$-ES: chaîne de Markov homogène sur $\mathbb{R}^N$ de noyau de transition p_σ.

$$X_{n+1} = \arg\min\{\mathcal{F}(X_n + \sigma_n\mathcal{N}^1(0,1)), ..., \mathcal{F}(X_n + \sigma_n\mathcal{N}^\lambda(0,1))\}$$

avec,

$$\sigma_n = \sigma^\star r_n \text{ et } r_n = |X_n|.$$

Modélisation mathématique (2)

Modélisation du $(1,\lambda)$*-ES avec mutation auto adaptative:*

$$X_{n+1} = \arg\min\{\mathcal{F}(X_n + \sigma_n \eta^1 \mathcal{N}^1(0,1)), ..., \mathcal{F}(X_n + \sigma_n \eta^\lambda \mathcal{N}^\lambda(0,1))\}$$

- avec $\mathcal{N}^1(0,1), ..., \mathcal{N}^\lambda(0,1)$, λ variables aléatoires gaussiennes centrées réduites indépendantes.
- et, $\eta^1, ..., \eta^\lambda$, λ variables aléatoires de distribution log normale: $\eta = \exp(\tau \mathcal{N}(0,1))$.
- σ_n est mis a jour de la façon suivante:

  $\sigma_{n+1} = \sigma_n \eta^\star$ avec $X_{n+1} = X_n + \sigma_n \eta^\star \mathcal{N}^\star(0,1)$.

- X_n est une chaîne de Markov non homogène sur $\mathbb{R}^N$

- MOTIVATIONS: Comment choisir de façon “optimale” les paramètres $\lambda, \sigma^{\star}, \tau$?
- APPROCHES EXISTANTES :
  - Approche locale: on regarde ce qui se passe au niveau d’une itération de l’algorithme *Approche due à Rechenberg, Schwefel(*“Evolution and Optimum seeking”*) et Beyer(*“The Theory of Evolution Strategies”*)*
  - Approche globale: on s’interesse à la convergence globale des algorithmes.
    On choisit les paramètres de façon à maximiser la vitesse de convergence
    Approche due à O. Francois et A.Bienvenüe.

Approche locale

Approche locale

$x \in \mathbb{R}^N$,Y une variable aléatoire telle que $Y = X_{n+1}$ si $X_n = x$ (densité de probabilité $p_\sigma(., x)$)

- Taux de progression en x.

$$\varphi(x) = \mathbb{E}\left(|x| - |Y|\right) = \left(\int |x| - |y| \, p_\sigma(y, x) dy\right)$$

- Taux de progression normalisé

$$\varphi^\star(\sigma^\star, N) = \frac{N}{|x|}\mathbb{E}\left(|x| - |Y|\right) = \left(\int \frac{|x| - |y|}{|x|} p_\sigma(y, x) dy\right)$$

On montre que $\varphi^\star$ est indépendant de $|x|$.

Approche locale

- Gain de qualité ou espérance de progrès

$$\mathcal{Q}(x) = \mathbb{E}\left(|\mathcal{F}(x)| - |F(Y)|\right)$$

- Probabilité de succès

$$\mathcal{P} = \mathbb{P}\left(\mathcal{F}(Y) < \mathcal{F}(x)\right)$$

- L'approche locale consiste à maximiser ces trois quantités.

$(1+1)$-ES. Mutation adaptative

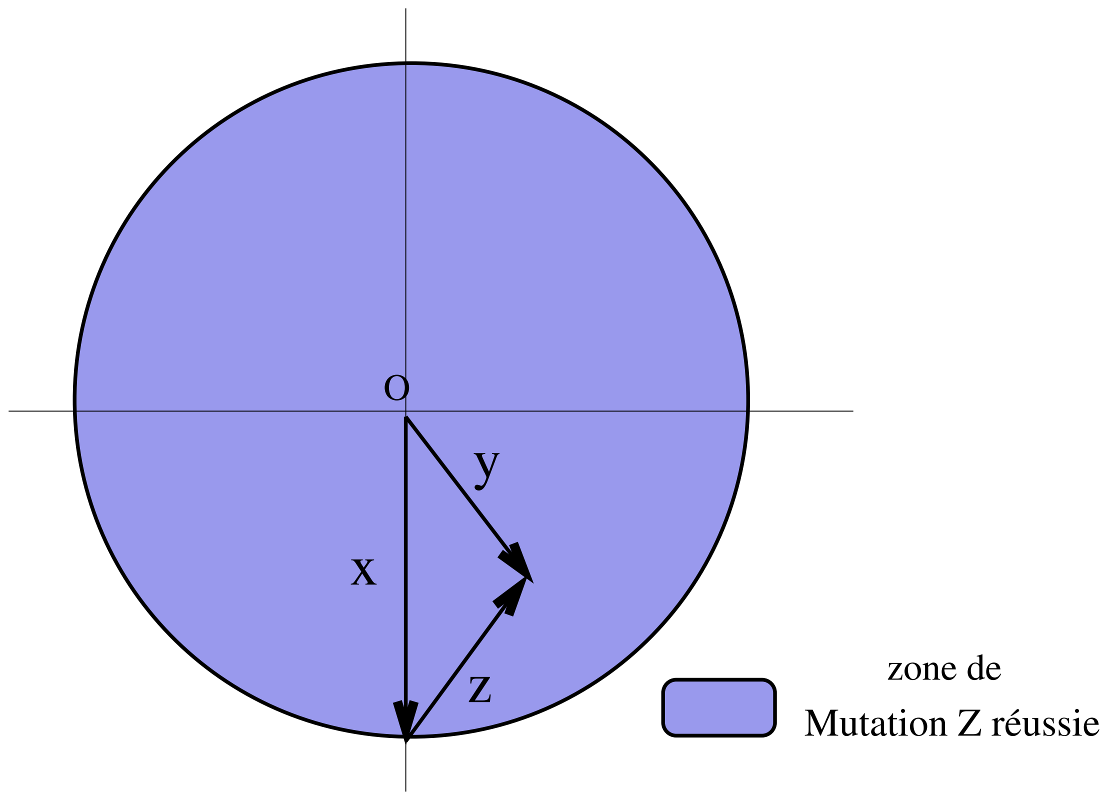
$$Y = \arg\min\left(\mathcal{F}(x), ..., \mathcal{F}(x + \sigma^\star |x| \mathcal{N}(0,1))\right)$$

$$Z \sim \alpha\delta_0 + \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2\sigma^2} z_i^2) \mathbb{1}_{\mathcal{A}}(z)$$

$$\begin{aligned}
\varphi(x) &= \mathbb{E}\left(|x| - |x + Z|\right) \\
&= \alpha(|x| - |x + 0|) \\
&\quad + \int (|x| - |x + z|) \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2\sigma^2} z_i^2) \mathbb{1}_{\mathcal{A}}(z) dz_1, .., dz_N
\end{aligned}$$

$(1+1)$-ES. Mutation adaptative

- $\mathcal{A} = \{w \in \mathbb{R}^N$ tq $\mathcal{F}(w) < \mathcal{F}(x)\}$ la zone de mutation réussie.



$(1+1)$-ES. Probabilité de succés

$$
\begin{aligned}
\mathcal{P} &= \mathbb{P}\left(\mathcal{F}(Y) < \mathcal{F}(x)\right) \\
&= \mathbb{P}(Z \in \mathcal{A}) \\
&= \int_{\mathcal{A}} \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2} z_i^2\right) \mathbb{1}_{\mathcal{A}}(z) dz_1, .., dz_N \\
&= \int_{\mathcal{A}} \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2} z_i^2\right) dz_1, .., dz_N
\end{aligned}
$$

$(1+1)$-ES. Mutation adaptative

- Formules intégrales exactes:

$$\begin{aligned}
\mathcal{P}_{1+1}(\sigma^\star, N) &= \frac{1}{\sqrt{2\pi}} \int_{y=0}^{y=\frac{2N}{\sigma^\star}} \exp(-\frac{1}{2}y^2) \frac{1}{\Gamma(\frac{N-1}{2})} \\
&\quad \int_{t=0}^{t=\frac{N}{\sigma^\star}y-\frac{y^2}{2}} \exp(-t) t^{\frac{N-1}{2}-1} dt dy \\
&= \int_{\widetilde{A}} q(t,y) dt dy
\end{aligned}$$

$$\varphi^\star_{1+1}(\sigma^\star, N) = \int_{\widetilde{A}} \left(1 - \sqrt{(1-\frac{\sigma^\star}{N}y)^2 + 2(\frac{\sigma^\star}{N})^2 t}\right) q(t,y) dt dy$$

$(1+1)$-*ES. Mutation adaptative*

- Formules asymptotiques (avec F la fonction de répartition d'une gaussienne).

$$\mathcal{P}_{1+1}(\sigma^\star) = 1 - F(\frac{\sigma^\star}{2}) - \mathcal{O}(\frac{1}{N})$$

$$\varphi^\star_{1+1}(\sigma^\star) = \frac{\sigma^\star}{\sqrt{2\pi}} exp(-\sigma^{\star 2}) - \frac{\sigma^{\star 2}}{2}\mathcal{P}_{1+1}(\sigma^\star) + \mathcal{O}(\frac{1}{N})$$

- Remarque sur l'interet de formules asymptotiques:
  - obtenir une estimation qualitative
  - éviter des inteǵrations numériques

Approche locale pour le $(1, \lambda)$*-ES*

- Même démarche que pour le $(1+1)$-ES: On peut calculer des formules intégrales exactes et en déduire des formules asymptotiques:

$$\varphi^\star_{1,\lambda}(\sigma^\star) = c_{1,\lambda}\sigma^\star - \frac{\sigma^{\star 2}}{2} + \mathcal{O}(\frac{1}{N})$$

$$c_{1,\lambda} := \frac{\lambda}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t \exp(-\frac{1}{2}t^2) F(t)^{\lambda-1} dt$$

- Efficacité en fitness: Intérêt pratique, choix du nombre d'enfants.

$$\eta_{1,\lambda} := \frac{max_{\sigma^\star}[\varphi^\star(\sigma^\star)]}{\lambda}$$

$$\eta_{1,\lambda} = \frac{c^2_{1,\lambda}}{2\lambda} + \mathcal{O}(\frac{1}{N})$$

$(1, \lambda)$-ES. Résultats numériques

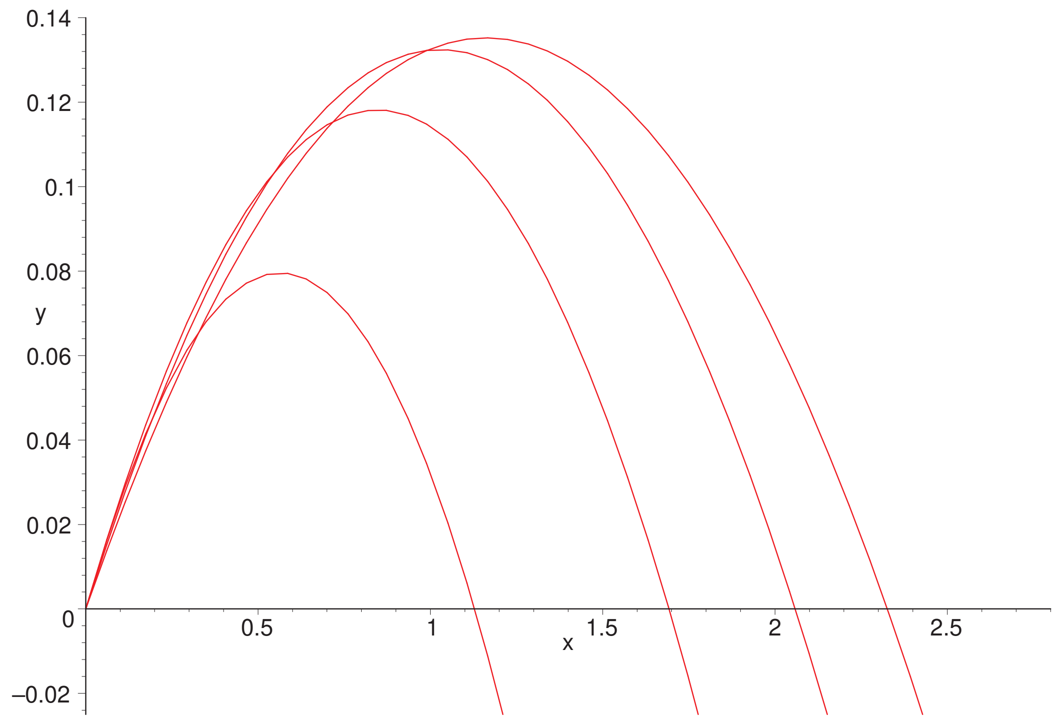


Figure 1: Taux de progrés par évaluation $\frac{\varphi^\star}{\lambda}$ en fonction de σ^{star} De gauche à droite $\lambda = 2, 3, 4, 5$ $\lambda_{opt} = 5$ $\sigma_{opt} = 5$

$(1, \lambda)$-*ES. Résultats*

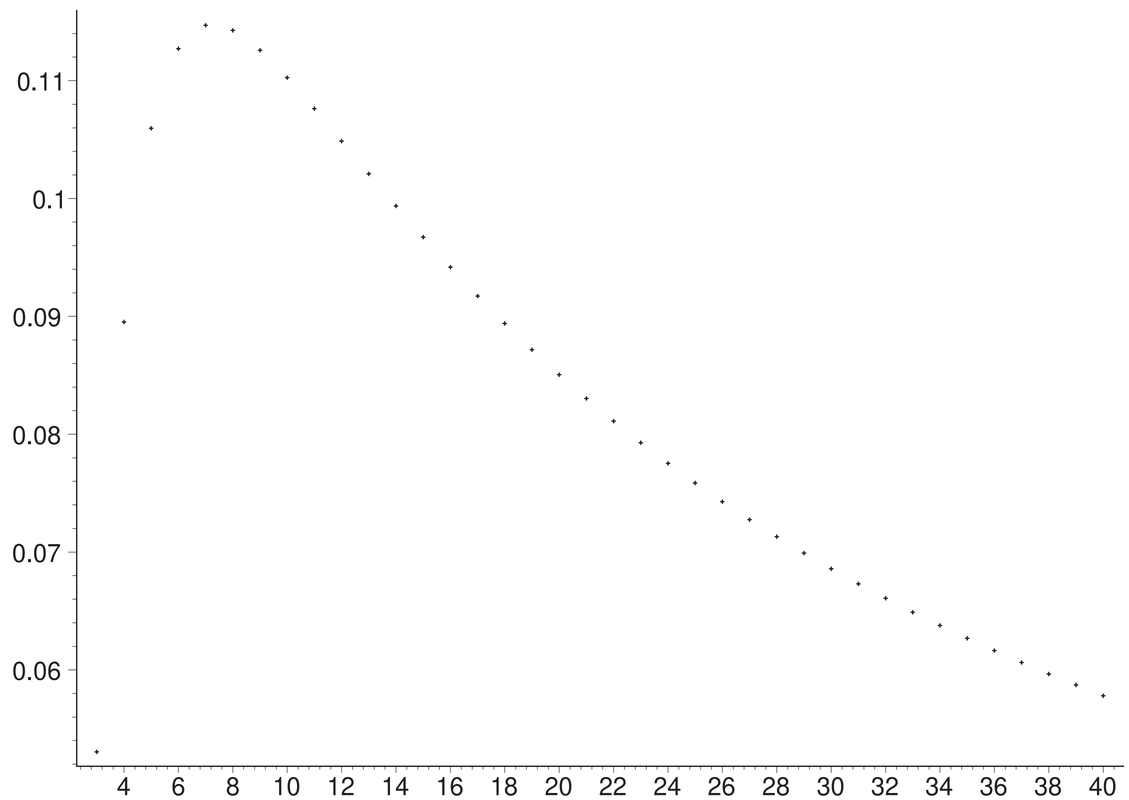


Figure 2: $\eta_\lambda = \frac{\max_{\sigma^\star} \varphi^\star}{\lambda}$ en fonction de λ

Approche locale pour le $(1+\lambda)$-ES

- Pas de formules exactes (pour $\lambda \geq 2$) intégrales mais des développements asymptotiques!

$$\varphi^\star_{1+\lambda}(\sigma^\star) = \sigma^\star d_{1+\lambda}\left(\frac{\sigma^\star}{2}\right) - \frac{\sigma^{\star 2}}{2}\left[1 - F(\frac{\sigma^\star}{2})^\lambda\right] + \mathcal{O}(\frac{1}{N})$$

Approche locale pour le $(1+\lambda)$*-ES*

avec,

$$d_{1+1}(x) := \frac{\exp(-\frac{1}{2}t^2)}{\sqrt{2\pi}}$$

$$d_{1+2}(x) := \frac{1}{\sqrt{\pi}}\left(1 - F(\sqrt{2}x)\right) + \frac{2}{\sqrt{2\pi}}\exp(-\frac{1}{2}t^2)F(x)$$

et, pour $\lambda \geq 2$:

$$d_{1+\lambda}(x) =:= \frac{\lambda}{\sqrt{2\pi}}\int_x^{\infty} t\exp(-\frac{1}{2}t^2)F(t)^{\lambda-1}dt$$

$(1+\lambda)$-*ES. Résultats numériques*

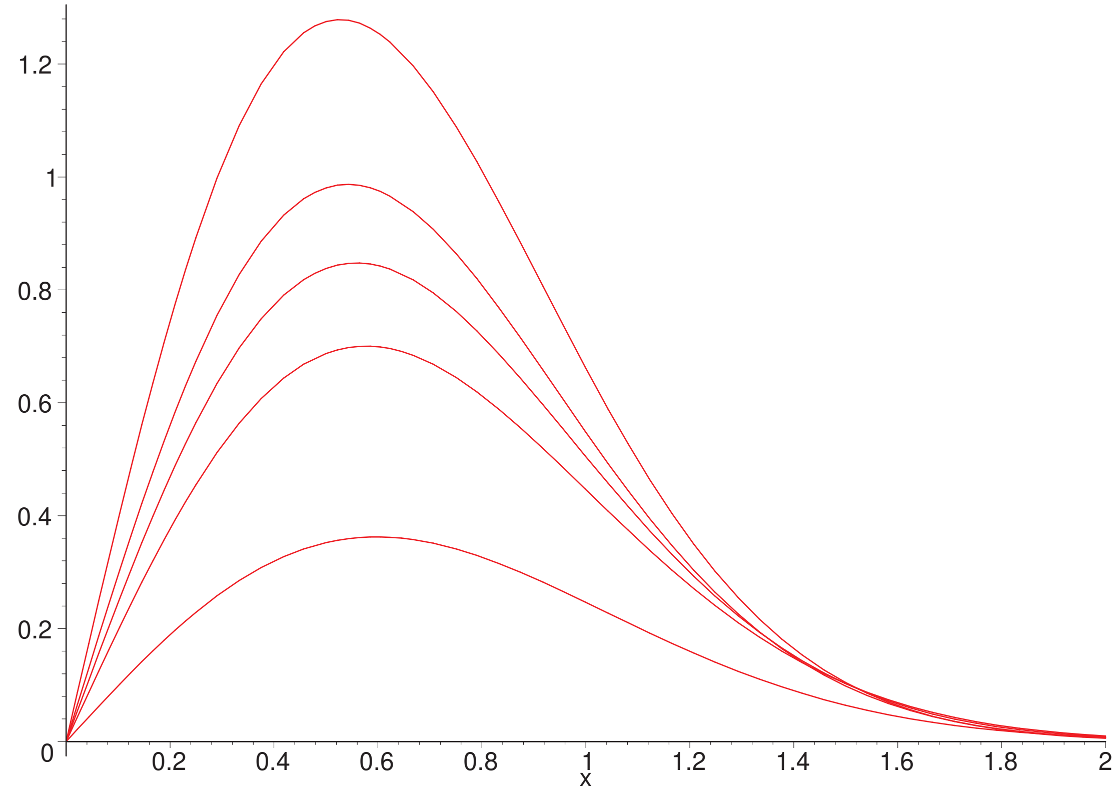


Figure 3: Allure de $\frac{\varphi^\star}{\lambda}$ en fonction de $\sigma^\star$ De haut en bas $\lambda = 1, 2, 3, 4$ $\lambda_{opt} = 1$

Approche globale

Mutation adaptative $(1, \lambda)$*-ES*

$$X_{n+1} = \arg\min\{\mathcal{F}(X_n + \sigma_n \mathcal{N}^1), ..., \mathcal{F}(X_n + \sigma_n \mathcal{N}^\lambda)\} \qquad (1)$$

- $\hat{Y}_{X_n}$ le vecteur tel que $X_{n+1} = X_n + \hat{Y}_{X_n}$
- La distribution de $\hat{Y}_x$ est $p_{\sigma^\star|x|}(x + ., x)$.
- On a la propriété suivante: $\alpha \hat{Y}_x = \hat{Y}_{\alpha x}$

Mutation adaptative $(1, \lambda)$*-ES*

- Densité de probabilité de $\hat{Y}_1$ est donnée par:

$$g_{\lambda,\sigma^\star}(t) = \frac{\lambda}{\sigma^\star} p_{\mathcal{N}}(\frac{t}{\sigma^\star}) \left[F_{\mathcal{N}} \left(\frac{-|t+1|-1}{\sigma^\star} \right) + 1 - F_{\mathcal{N}} \left(\frac{|t+1|-1}{\sigma^\star} \right) \right]^{\lambda-1}$$

où $p_{\mathcal{N}}(t)$ est le densité de probabilité gaussienne et $F_{\mathcal{N}}$ sa fonction de répartition.

$$\begin{aligned} p_{\mathcal{N}}(t) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} t^2 \right) \\ F_{\mathcal{N}}(t) &= \int_{-\infty}^{t} p_{\mathcal{N}}(u) du \end{aligned}$$

Mutation adaptative $(1, \lambda)$-ES

Théoreme 1 *Soit (X_n) la chaîne de Markov définie par (**??**), si $\mathbb{E}(Y_1) < \infty$, alors X_n converge ou diverge linéairement.*

$$\frac{1}{n}\ln(r_n) \longrightarrow \kappa(\sigma^\star, \lambda) si\ n \longrightarrow \infty$$

où,

$$\kappa(\sigma^\star, \lambda) = \int_{-\infty}^{+\infty} \ln(|1 + y|) g_{\lambda,\sigma^\star}(y) dy.$$

Mutation adaptative $(1, \lambda)$*-ES*

Preuve:

- $r_{n+1} = |X_n + \hat{Y}_{X_n}| = r_n|1 + sign(X_n)\hat{Y}_1|$
- $\ln r_{n+1} = \ln r_0 + \sum_{k=1}^{n} \ln|1 + \hat{Y}_1^k|$ avec $\hat{Y}_1^k$ des variables aléatoires indépendantes de même loi que $\hat{Y}_1$
- La loi forte des grands nombres nous donne que

$$\frac{1}{n}\sum_{k=1}^{n} \ln|1 + \hat{Y}_1^k| \longrightarrow \int_{-\infty}^{+\infty} \ln(|1 + y|) g_{\lambda,\sigma^\star}(y)dy.$$

Application pratique

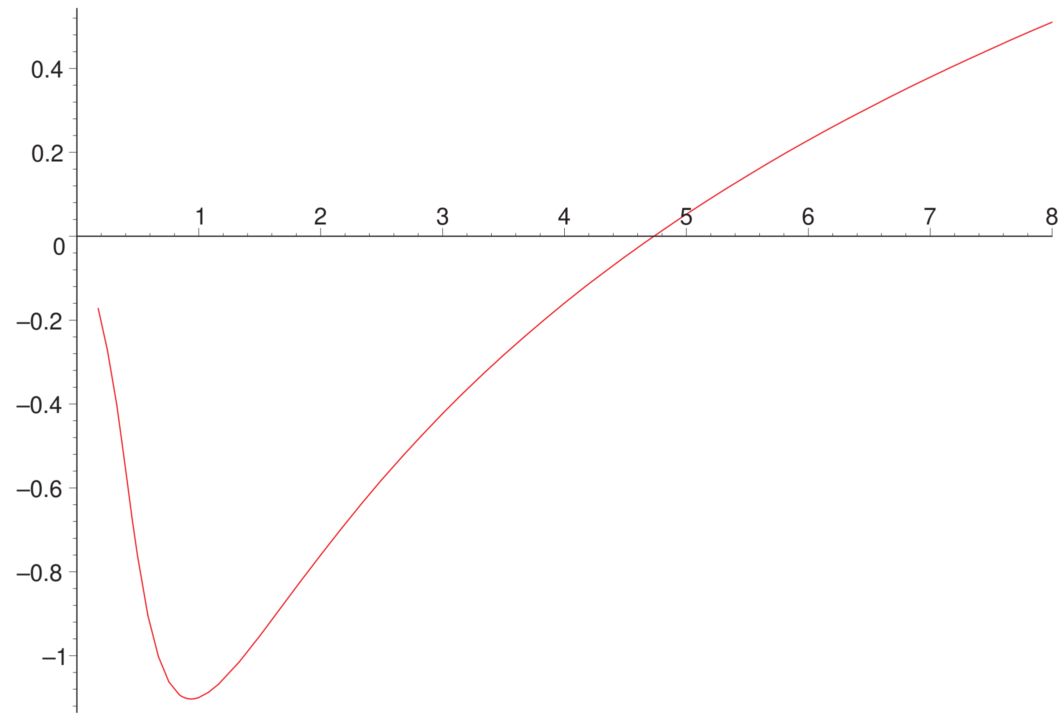


Figure 4: $\kappa(\sigma^\star, 3)$ lorsque $\sigma^\star$ (en abscisse) varie

- $\sigma^\star_{opt} \approx 0.94$
- l'algorithme diverge pour $\sigma^\star > \sigma^\star_c \approx 4.73$

Cas de la mutation auto adaptative

$$X_{n+1} = \arg\min\{\mathcal{F}(X_n + \sigma_n \eta^1 \mathcal{N}^1), ..., \mathcal{F}(X_n + \sigma_n \eta^\lambda \mathcal{N}^\lambda)\} \quad (2)$$

- $Z_n = \frac{r_n}{\sigma_n}$ avec $r_n = |X_n|$ est une chaîne de Markov homogène.
- Si on suppose que la chaîne homogène Z_n est récurrente au sens d'Harris, on a l'existence d'une mesure stationnaire pour Z_n et, on a alors le théorème de convergence global suivant:

Cas de la mutation auto adaptative

Théoreme 2 *Soit* (X_n) *la chaîne de Markov définie par (**??**) et* $Z_n = \frac{r_n}{\sigma_n}$*. On suppose que la chaîne est récurrente au sens d'Harris. Alors* X_n *converge ou diverge linéairement. Le taux de convergence est donné par:*

$$\frac{1}{n}\ln(r_n) \longrightarrow \int_0^{+\infty} \mathbb{E}(ln(\frac{r_z^1}{z}))d\mu(z)$$

où, μ *est la mesure stationnaire de* (Z_n)*.*

Cas de la mutation auto adaptative

- Intégration numérique du taux de convergence difficile à moins de connaître la mesure stationnaire de Z_n
- Intérêt de ce théorème: montre que l'étude de la convergence de X_n se ramène a l'étude d'une chaîne plus simple