Propriétés probabilistes dans les algorithmes d'optimisation sans et avec dérivées

Clément Royer - University of Wisconsin-Madison

Séminaire SPOC
Institut de Mathématiques de Bourgogne

12 avril 2017

Introduction: Aléatoire et optimisation

L'aléatoire est de plus en plus présent en optimisation numérique.

Pour de multiples raisons :

- *Problèmes de grande taille :* Méthodes classiques trop coûteuses.
- *Calcul distribué :* Données stockées sur plusieurs machines/processeurs.
- *Applications :* Problèmes d'apprentissage.

Introduction: Aléatoire et optimisation

L'aléatoire est de plus en plus présent en optimisation numérique.

Pour de multiples raisons :

- *Problèmes de grande taille :* Méthodes classiques trop coûteuses.
- *Calcul distribué :* Données stockées sur plusieurs machines/processeurs.
- *Applications :* Problèmes d'apprentissage.

Questions sur l'aléatoire

- Comment l'analyse des méthodes est-elle affectée ?
- Améliore-t-on les variantes déterministes ?
- Aléatoire en optimisation sans dérivées ?

Complexité des algorithmes d'optimisation

Analyse de Complexité

- Etudier le taux de convergence d'un critère donné.
- Borner le comportement d'une méthode au pire cas.
- Avec aléatoire : résultats en espérance/probabilité.

Utilité de la complexité

- Quelles indications données par la complexité ?
- Quel lien avec la pratique ?
- Importance pour les méthodes sans dérivées ?

Objectifs du séminaire

Trame

1. Introduire des aspects aléatoires dans des algorithmes sans dérivées.
2. Fournir des garanties théoriques (ex : complexité).
3. Comparer la complexité et le comportement numérique.

Objectifs du séminaire

Trame

1. Introduire des aspects aléatoires dans des algorithmes sans dérivées.
2. Fournir des garanties théoriques (ex : complexité).
3. Comparer la complexité et le comportement numérique.

- Dans cet exposé : méthodes de recherche directe;
- Les résultats s'appliquent à d'autre méthodes, comme les régions de confiance.

Outline

Outline

Problème et hypothèses

Soit le problème suivant :

$$\min_{x \in \mathbb{R}^n} f(x).$$

Hypothèses sur f

- f minorée, a priori non convexe.
- f de classe $\mathcal{C}^1$, ∇f fonction lipschitzienne.

Problème et hypothèses

Soit le problème suivant :

$$\min_{x \in \mathbb{R}^n} f(x).$$

Hypothèses sur f

- f minorée, a priori non convexe.
- f de classe $\mathcal{C}^1$, ∇f fonction lipschitzienne.

Optimisation différentiable

Depuis $x \in \mathbb{R}^n$, on peut décroître f dans la direction de $-\nabla f(x)$!

- Principe de base des méthodes de *premier ordre/gradient*.
- Objectif : converger vers un **point stationnaire d'ordre 1**:

$$\liminf_{k \to \infty} \|\nabla f(x_k)\| = 0.$$

Le contexte sans dérivées

Le gradient de f existe mais ne peut pas être utilisé en pratique.

- *Code de simulation :* le gradient est trop coûteux.
- *Fonction f disponible sous forme de boîte noire :* pas de code pour la dérivée.
- *Différentiation automatique* inapplicable.

Examples : météorologie, industrie pétrolière, médecine...

Le contexte sans dérivées

Le gradient de f existe mais ne peut pas être utilisé en pratique.

- *Code de simulation :* le gradient est trop coûteux.
- *Fonction f disponible sous forme de boîte noire :* pas de code pour la dérivée.
- *Différentiation automatique* inapplicable.

Examples : météorologie, industrie pétrolière, médecine...

Mesure de performance : Nombre d'évaluations de l'objectif.

Algorithmes d'Optimisation sans Dérivées (DFO)

Méthodes sans dérivées déterministes

- Méthodes à modèles, comme les régions de confiance.
- Méthodes directionnelles, comme la recherche directe.

Introduction to Derivative-Free Optimization
A.R. Conn, K. Scheinberg, L.N. Vicente. (2009)

- Théorie de convergence (vers des optima locaux).
- Bornes de complexité/Vitesses de convergence activement étudiées.

Algorithmes d'Optimisation Sans Dérivées (2)

Méthodes "DFO" stochastiques

- Ont pour but de trouver des optima globaux:
  Ex) Stratégies Evolutionnaires, Algorithmes Génétiques.
- Souvent sans analogues déterministes.

- Cet exposé ne concerne pas les méthodes stochastiques;
- Nos algorithmes sont basés sur des éléments probabilistes.

Méthodes "DFO" avec propriétés probabilistes

- Développées à partir d'algorithmes déterministes.
- Bénéficient de garanties théoriques grâce à cela.
- Le côté aléatoire améliore la performance.

Outline

Algorithmes de recherche directe

- Variantes sans dérivées des méthodes de gradient.
- Introduites vers 1960, théorie de convergence vers 1990.
- Simples à implémenter, fort potentiel de parallélisme.

- **Optimization by direct search: new perspectives on some classical and modern methods.**
  Kolda, Lewis and Torczon (*SIAM Review*, 2003).

Un algorithme basique de recherche directe

1. **Initialisation:** $x_0 \in \mathbb{R}^n, \alpha_0 > 0, 0 < \theta < 1 \leq \gamma$.
2. **Pour** $k = 0, 1, 2, \ldots$
   - Choisir un ensemble D_k de r vecteurs.
   - Si il existe $d_k \in D_k$ tel que

$$f(x_k + \alpha_k d_k) < f(x_k) - \alpha_k^2,$$

     alors *(k réussie)* poser $x_{k+1} := x_k + \alpha_k d_k$ et $\alpha_{k+1} := \gamma \alpha_k$.
   - Sinon *(k non réussie)* poser $x_{k+1} := x_k$ et $\alpha_{k+1} := \theta \alpha_k$.

Un algorithme basique de recherche directe

1. **Initialisation:** $x_0 \in \mathbb{R}^n, \alpha_0 > 0, 0 < \theta < 1 \leq \gamma$.
2. **Pour** $k = 0, 1, 2, \ldots$
   - Choisir un ensemble D_k de r vecteurs.
   - Si il existe $d_k \in D_k$ tel que

$$f(x_k + \alpha_k d_k) < f(x_k) - \alpha_k^2,$$

     alors *(k réussie)* poser $x_{k+1} := x_k + \alpha_k d_k$ et $\alpha_{k+1} := \gamma \alpha_k$.
   - Sinon *(k non réussie)* poser $x_{k+1} := x_k$ et $\alpha_{k+1} := \theta \alpha_k$.

Choisir les directions de sondage

On cherche à choisir les ensembles des directions/de sondage D_k pour garantir la convergene de l'algorithme.

Choisir les directions de sondage

On cherche à choisir les ensembles des directions/de sondage D_k pour garantir la convergene de l'algorithme.

Une mesure de qualité

Pour un ensemble de vecteurs D, la mesure cosinus de D est donnée par

$$\mathrm{cm}(D) = \min_{v \in \mathbb{R}^n \setminus \{0\}} \max_{d \in D} \frac{d^\top v}{\|d\| \, \|v\|}.$$

Choisir les directions de sondage

On cherche à choisir les ensembles des directions/de sondage D_k pour garantir la convergene de l'algorithme.

Une mesure de qualité

Pour un ensemble de vecteurs D, la mesure cosinus de D est donnée par

$$\mathrm{cm}(D) = \min_{v \in \mathbb{R}^n \setminus \{0\}} \max_{d \in D} \frac{d^\top v}{\|d\| \, \|v\|}.$$

- Si $\mathrm{cm}(D) > 0$, pour tout v il existe $d \in D$ tel que (d, v) est un angle aigu.
- Avec $v = -\nabla f(x) \neq 0$, D contient une direction de descente pour f en x.

Génération positive

Comment garantir $\mathrm{cm}(D) > 0$?

Ensemble de générateurs positifs (PSS)

D est un PSS s'il génère $\mathbb{R}^n$ par combinaisons linéaires à coefficients positifs ou nuls.

- D est un PSS $\Leftrightarrow \mathrm{cm}(D) > 0$.
- Un PSS contient au moins $n+1$ vecteurs.

Génération positive

Comment garantir $\mathrm{cm}(D) > 0$?

Ensemble de générateurs positifs (PSS)

D est un PSS s'il génère $\mathbb{R}^n$ par combinaisons linéaires à coefficients positifs ou nuls.

- D est un PSS $\Leftrightarrow \mathrm{cm}(D) > 0$.
- Un PSS contient au moins $n+1$ vecteurs.

Exemple

$D_\oplus = \{e_1, \ldots, e_n, -e_1, \ldots, -e_n\}$ est un PSS avec

$$\mathrm{cm}\left(D_\oplus\right) = \frac{1}{\sqrt{n}}.$$

Convergence de la recherche directe déterministe

Lemma

Si l'itération k n'est pas réussie et $\mathrm{cm}(D_k) \geq \kappa > 0$,

$$\kappa \|\nabla f(x_k)\| \leq \mathcal{O}(\alpha_k).$$

Lemma

Indépendamment de $\{D_k\}$,

$$\lim_{k \to \infty} \alpha_k = 0.$$

Convergence de la recherche directe déterministe

Lemma

Si l'itération k n'est pas réussie et $\mathrm{cm}(D_k) \geq \kappa > 0$,

$$\kappa \|\nabla f(x_k)\| \leq \mathcal{O}(\alpha_k).$$

Lemma

Indépendamment de $\{D_k\}$,

$$\lim_{k \to \infty} \alpha_k = 0.$$

Théorème de convergence

Si $\forall k,\ \mathrm{cm}(D_k) \geq \kappa$,

$$\liminf_{k \to \infty} \|\nabla f(x_k)\| = 0.$$

Théorème de complexité

Soient $\epsilon \in (0,1)$ et N_ϵ le nombre d'appels à f nécessaires pour satisfaire $\inf_{0 \leq l \leq k} \|\nabla f(x_l)\| < \epsilon$. On a

$$N_\epsilon \leq \mathcal{O}\left(r\,(\kappa\,\epsilon)^{-2}\right).$$

En choisissant $D_k = D_\oplus$, on a $\kappa = 1/\sqrt{n}, r = 2n$, la borne devient

$$N_\epsilon \leq \mathcal{O}\left(n^2\,\epsilon^{-2}\right).$$

Outline

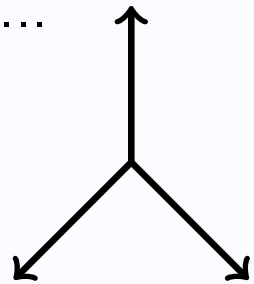
Idée (Gratton et Vicente, 2013)

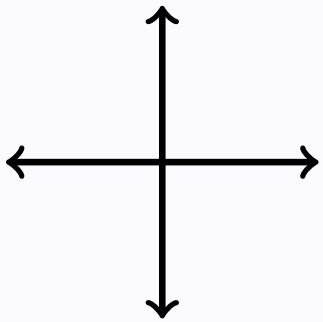
Utiliser des vecteurs tirés aléatoirement et indépendamment, typiquement moins que $n+1$!

Idée (Gratton et Vicente, 2013)

Utiliser des vecteurs tirés aléatoirement et indépendamment, typiquement moins que $n+1$!

From PSS...

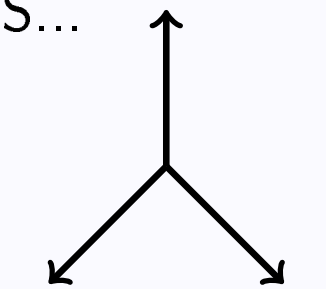


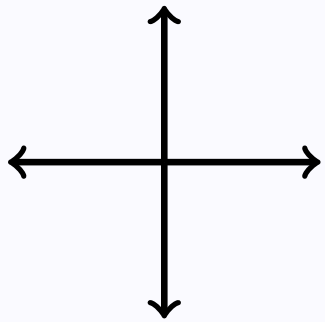


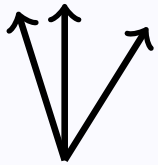
Idée (Gratton et Vicente, 2013)

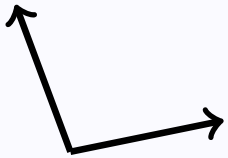
Utiliser des vecteurs tirés aléatoirement et indépendamment, typiquement moins que $n+1$!

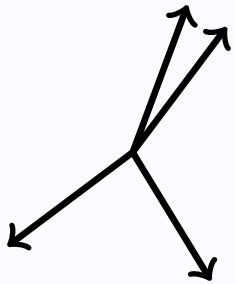
From PSS...











...to random sets

Motivation numérique

- Test de convergence: $f(x_k) < f_{\text{low}} + 10^{-3}\,(f(x_0) - f_{\text{low}})$;
- Budget: $2000\,n$ appels à f max.

Problème	$D_\oplus$	$Q\,D_\oplus$	$2\,n$	$n+1$	$n/2$	2	1
	Déterministe		Probabiliste				
arglina	3.42	16.67	10.30	6.01	3.21	1.00	–
arglinb	20.50	11.38	7.38	2.81	2.35	1.00	2.04
broydn3d	4.33	11.22	6.54	3.59	2.04	1.00	–
dqrtic	7.16	19.50	9.10	4.56	2.77	1.00	–
engval1	10.53	23.96	11.90	6.48	3.55	1.00	2.08
freuroth	56.00	1.33	1.00	1.67	1.33	1.00	4.00
integreq	16.04	18.85	12.44	6.76	3.52	1.00	–
nondquar	6.90	17.36	7.56	4.23	2.76	1.00	–
sinquad	–	2.12	1.31	1.00	1.60	1.23	–
vardim	1.00	3.30	1.80	2.40	2.30	1.80	4.30

Table: Ratio du nombre d'appels à f (moyenne sur 10 réalisations, taille $n = 40$)

Algorithme de recherche directe probabiliste

Notations probabilistes

- Ensembles/Directions de sondage : $D_k = \mathfrak{D}_k(\omega)$, $d_k = \mathfrak{d}_k(\omega)$;
- Itérés : $x_k = X_k(\omega)$;
- Longueurs de pas : $\alpha_k = A_k(\omega)$.

1. **Initialisation:** $x_0 \in \mathbb{R}^n, \alpha_0 > 0, 0 < \theta < 1 \leq \gamma$.
2. **Pour** $k = 0, 1, 2, ...$
   - Choisir un ensemble $\mathfrak{D}_k$ de r vecteurs **tirés indépendamment au hasard.**
   - Si il existe $\mathfrak{d}_k \in \mathfrak{D}_k$ tel que

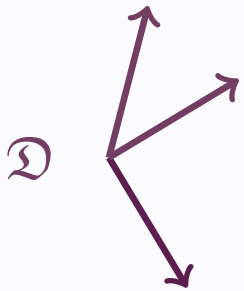
$$f(X_k + \_k\, d_k) < f(X_k) - A_k^2,$$

     alors *(k réussie)* poser $X_{k+1} := X_k + A_k\, \mathfrak{d}_k$ et $A_{k+1} := \gamma\, A_k$.
   - Sinon *(k non réussie)* poser $X_{k+1} := X_k$ et $A_{k+1} := \theta\, A_k$.

Outline

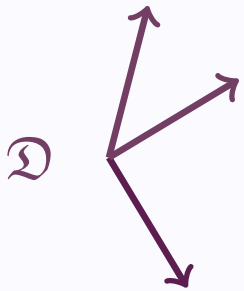
Qu'est-ce qu'un bon ensemble de sondage ?

- $\mathfrak{D}$ n'est pas un PSS...

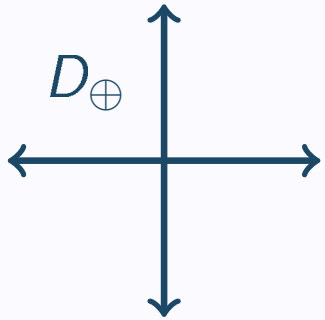


Qu'est-ce qu'un bon ensemble de sondage ?

$\mathfrak{D}$ n'est pas un PSS...

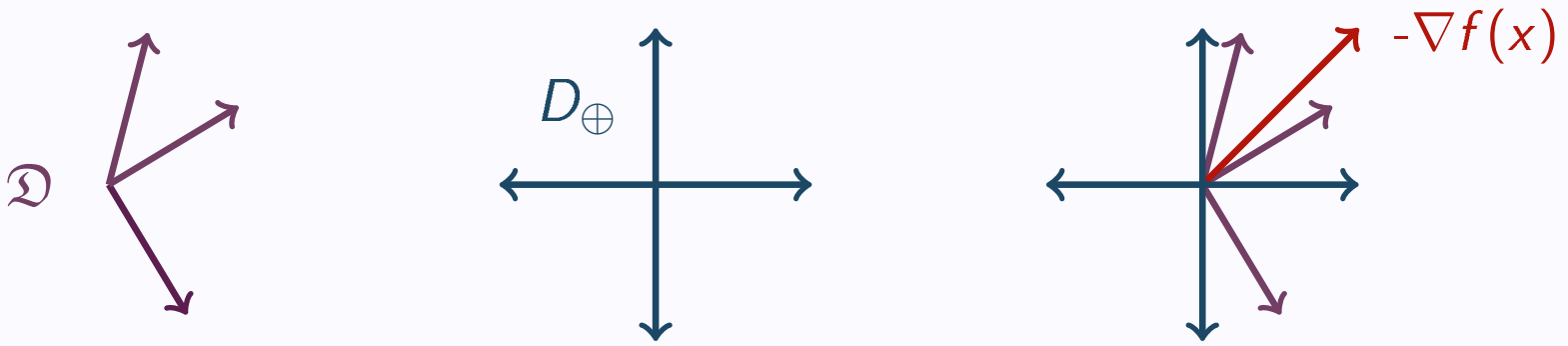


...$D_{\oplus}$ si...



Qu'est-ce qu'un bon ensemble de sondage ?

$\mathfrak{D}$ n'est pas un PSS... ...$D_\oplus$ si... ...$-\nabla f(x)$ plus proche de $\mathfrak{D}$!



Etre proche de l'opposé du gradient : un gage de qualité ?

Une nouvelle mesure de qualité

Propriétés dans le cas déterministe

- On a requis

$$\mathrm{cm}(D_k) = \min_{v \neq 0} \max_{d \in D_k} \frac{d^\top v}{\|d\| \, \|v\|} \geq \kappa.$$

- Il suffirait d'avoir

$$\mathrm{cm}\left(D_k, -\nabla f(x_k)\right) = \max_{d \in D_k} \frac{d^\top [-\nabla f(x_k)]}{\|d\| \|\nabla f(x_k)\|} \geq \kappa.$$

- Avec de l'aléatoire, la seconde propriété peut être vraie en probabilité.
- Quels sont les bons outils probabilistes pour exprimer cela ?

Plusieurs types de résultats

Déterministe/Pour toute réalisation

$\Downarrow$

Avec probabilité 1/Presque sûr

$\Downarrow$

Avec une certaine probabilité

Sous-martingale

Une sous-martingale est une suite de variables aléatoires $\{V_k\}$ telle que $\mathbb{E}\left[|V_k|\right] < \infty$ et

$$\mathbb{E}\left(V_k | V_0, V_1, \ldots, V_{k-1}\right) \geq V_{k-1}.$$

Ensembles de directions et descente (p, κ)

- On cherche à étudier

$$\mathbb{P}\left(\mathrm{cm}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) \geq \kappa\right).$$

  où X_k dépend de $\mathfrak{D}_0, \ldots, \mathfrak{D}_{k-1}$ mais pas de $\mathfrak{D}_k$.
- On va utiliser les probabilités conditionnelles/le conditionnement au passé.

Ensembles de directions et descente (p, κ)

- On cherche à étudier

$$\mathbb{P}\left(\mathrm{cm}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) \geq \kappa\right).$$

où X_k dépend de $\mathfrak{D}_0, \ldots, \mathfrak{D}_{k-1}$ mais pas de $\mathfrak{D}_k$.

- On va utiliser les probabilités conditionnelles/le conditionnement au passé.

Propriété de descente probabiliste

Une suite d'ensembles aléatoires $\{\mathfrak{D}_k\}$ est dite à descente (p, κ) si:

$$\begin{aligned} \mathbb{P}\left(\mathrm{cm}\left(\mathfrak{D}_0, -\nabla f(x_0)\right) \geq \kappa\right) &\geq p \\ \forall k \geq 1, \quad \mathbb{P}\left(\mathrm{cm}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) \geq \kappa \mid \mathfrak{D}_0, \ldots, \mathfrak{D}_{k-1}\right) &\geq p, \end{aligned}$$

Résultats de convergence

Lemma

Pour toute réalisation $\{\alpha_k\}$ *de* $\{A_k\}$, *indépendamment de* $\{\mathfrak{D}_k\}$,

$$\lim_{k\to\infty} \alpha_k = 0.$$

Lemma

Si l'itération k *n'est pas réussie,*

$$\{\mathrm{cm}\,(\mathfrak{D}_k, -\nabla f(X_k)) \geq \kappa\} \subset \{\kappa \|\nabla f(X_k)\| \leq \mathcal{O}(A_k)\}.$$

Il s'agit de prouver que $\{\mathrm{cm}\,(\mathfrak{D}_k, -\nabla f(X_k)) \geq \kappa\}$ se produit suffisamment souvent.

Résultats de convergence (2)

Soit $\{\mathfrak{D}_k\}$ à descente (p, κ) et $Z_k = \mathbf{1}\left(\mathrm{cm}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) \geq \kappa\right)$.

Proposition

Soit

$$S_k = \sum_{i=0}^{k-1} \left[Z_i - p_0\right], \quad p_0 = \frac{\ln \theta}{\ln(\theta/\gamma)}.$$

1. $\{\liminf_k \|\nabla f(X_k)\| > 0\} \subset \{S_k \to -\infty\}$.
2. Si $p > p_0$, $\{S_k\}$ est une sous-martingale avec $\mathbb{P}\left(\limsup S_k = \infty\right) = 1$.

Résultats de convergence (2)

Soit $\{\mathfrak{D}_k\}$ à descente (p, κ) et $Z_k = \mathbf{1}\left(\mathrm{cm}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) \geq \kappa\right)$.

Proposition

Soit

$$S_k = \sum_{i=0}^{k-1} \left[Z_i - p_0\right], \quad p_0 = \frac{\ln \theta}{\ln(\theta/\gamma)}.$$

1. $\{\liminf_k \|\nabla f(X_k)\| > 0\} \subset \{S_k \to -\infty\}$.
2. Si $p > p_0$, $\{S_k\}$ est une sous-martingale avec $\mathbb{P}\left(\limsup S_k = \infty\right) = 1$.

Théorème : convergence presque sûre

Si $\{\mathfrak{D}_k\}$ est à descente (p, κ) avec $p > p_0$, on a

$$\mathbb{P}\left(\liminf_{k \to \infty} \|\nabla f(X_k)\| = 0 \right) = 1.$$

Idée intuitive

Soient $G_k = \nabla f(X_k)$ et $Z_k = \mathbf{1}\left(\mathrm{cm}(\mathfrak{D}_k, -G_k) \geq \kappa\right)$.

- Si $Z_k = 1$ et k réussie, on a $\kappa \left\| G_k \right\| < \mathcal{O}(A_k)$...

Complexité et descente probabiliste

Idée intuitive

Soient $G_k = \nabla f(X_k)$ et $Z_k = \mathbf{1}\left(\mathrm{cm}(\mathfrak{D}_k, -G_k) \geq \kappa\right)$.

- Si $Z_k = 1$ et k réussie, on a $\kappa \left\| G_k \right\| < \mathcal{O}(A_k)$...
- ...A_k tend vers 0...
- ...donc si $\inf_{0 \leq l \leq k} \left\| G_l \right\|$ est grand, $\sum_{l=0}^{k} Z_l$ doit être faible.

Idée intuitive

Soient $G_k = \nabla f(X_k)$ et $Z_k = \mathbf{1}\left(\mathrm{cm}(\mathfrak{D}_k, -G_k) \geq \kappa\right)$.

- Si $Z_k = 1$ et k réussie, on a $\kappa \|G_k\| < \mathcal{O}(A_k)$...
- ...A_k tend vers 0...
- ...donc si $\inf_{0 \leq l \leq k} \|G_l\|$ est grand, $\sum_{l=0}^{k} Z_l$ doit être faible.

Une borne utile

Pour chaque réalisation de l'algorithme,

$$\sum_{l=0}^{k} z_l \;\leq\; \mathcal{O}\left(\frac{1}{\kappa^2 \, \|\tilde{g}_k\|^2}\right) + p_0 \, k,$$

où $\|\tilde{g}_k\| = \inf_{0 \leq l \leq k} \|g_l\|$.

Complexité et descente probabiliste (2)

Rappel : $Z_l = \mathbf{1}\left(\mathrm{cm}(\mathfrak{D}_l, -\nabla f(X_l) \geq \kappa\right)$.

Argument d'inclusion

$$\left\{ \inf_{0 \leq l \leq k} \|\nabla f(X_k)\| \geq \epsilon \right\} \subset \left\{ \sum_{l=0}^{k} Z_l \leq \lambda\, k \right\}$$

avec $\lambda = \mathcal{O}\left(\frac{1}{k\,\kappa^2\,\epsilon^{-2}}\right) + p_0$.

Borne de type Chernoff

Pour tout $\lambda \in (0, p)$,

$$\mathbb{P}\left(\sum_{l=0}^{k-1} Z_l \leq \lambda\, k \right) \leq \exp\left[-\frac{(p-\lambda)^2}{2\,p} k \right].$$

Complexité et descent probabiliste (3)

Théorème : Complexité probabiliste

Soient $\{\mathfrak{D}_k\}$ à descente (p, κ), $\epsilon \in (0,1)$ et N_ϵ le nombre d'appels à f nécessaires pour obtenir $\inf_{0 \leq l \leq k} \|\nabla f(X_l)\| \leq \epsilon$. Alors

$$\mathbb{P}\left(N_\epsilon \leq \mathcal{O}\left(\frac{r\,(\kappa\epsilon)^{-2}}{p - p_0}\right)\right) \geq 1 - \exp\left(-\mathcal{O}\left(\frac{p - p_0}{p}(\kappa\,\epsilon)^{-2}\right)\right).$$

Complexité et descent probabiliste (3)

Théorème : Complexité probabiliste

Soient $\{\mathfrak{D}_k\}$ à descente (p, κ), $\epsilon \in (0,1)$ et N_ϵ le nombre d'appels à f nécessaires pour obtenir $\inf_{0 \leq l \leq k} \|\nabla f(X_l)\| \leq \epsilon$. Alors

$$\mathbb{P}\left(N_\epsilon \leq \mathcal{O}\left(\frac{r\,(\kappa\epsilon)^{-2}}{p - p_0}\right)\right) \geq 1 - \exp\left(-\mathcal{O}\left(\frac{p - p_0}{p}(\kappa\,\epsilon)^{-2}\right)\right).$$

- Déterministe : $\mathcal{O}(n^2\,\epsilon^{-2})$.
- Probabiliste : $\mathcal{O}(r\,n\,\epsilon^{-2})$ en probabilité
  $\Rightarrow \mathcal{O}(n\,\epsilon^{-2})$ lorsque $r = 2$!
- Serait-on meilleur (en probabilité) avec moins de directions ?

Outline

Construire une suite à descente (p, κ)

On cherche à avoir

$$p > p_0 = \frac{\ln(\theta)}{\ln(\theta/\gamma)}$$

avec le plus petit $r = |\mathfrak{D}_k|$ possible.

Une technique : génération uniforme sur la sphère unité

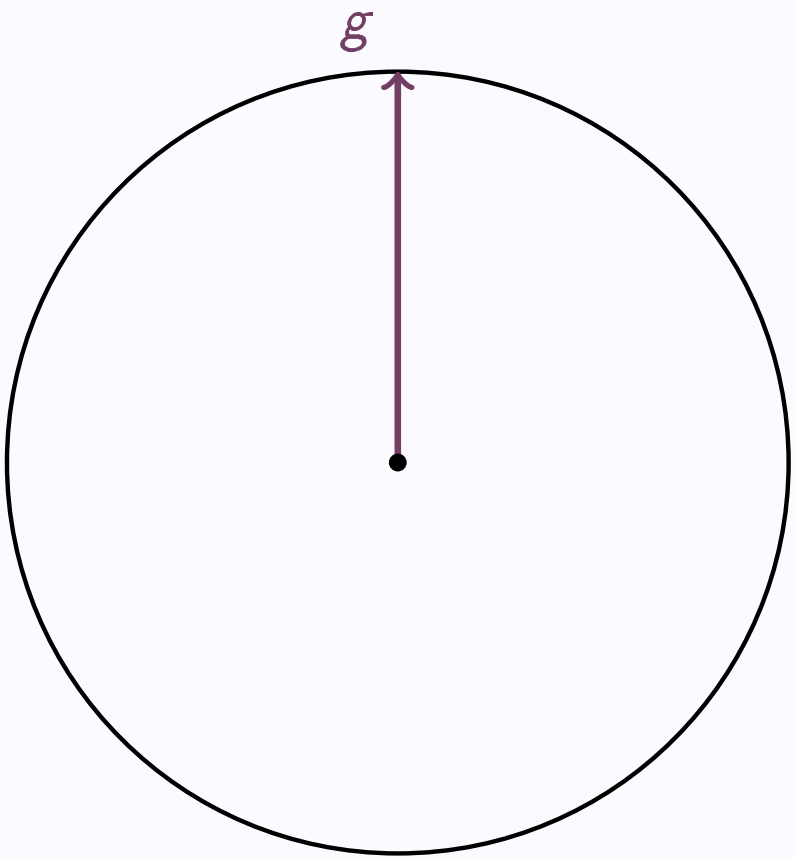
Si

$$r > \log_2 \left(1 - \frac{\ln \theta}{\ln \gamma}\right),$$

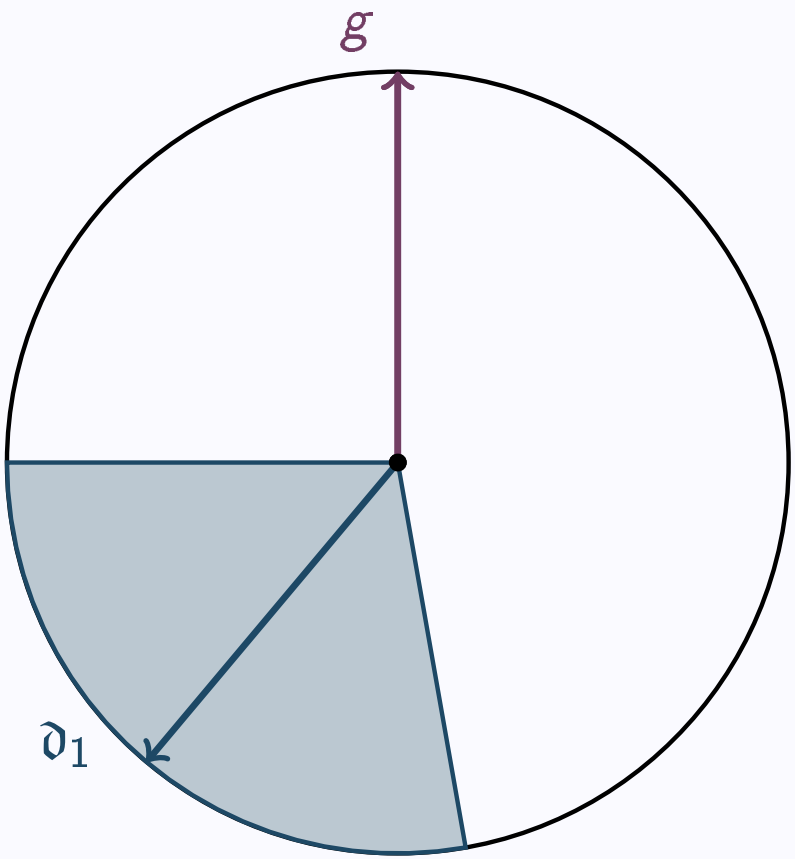
il existe (p, τ) indépendants de n tels que la suite $\mathfrak{D}_k$ est à descente $(p, \tau/\sqrt{n})$ et $p > p_0$.

Si $\gamma = \theta^{-1} = 2$, choisir $r \geq 2$ suffit à garantir $p > \frac{1}{2}$.

Deux directions uniformes suffisent, pas une

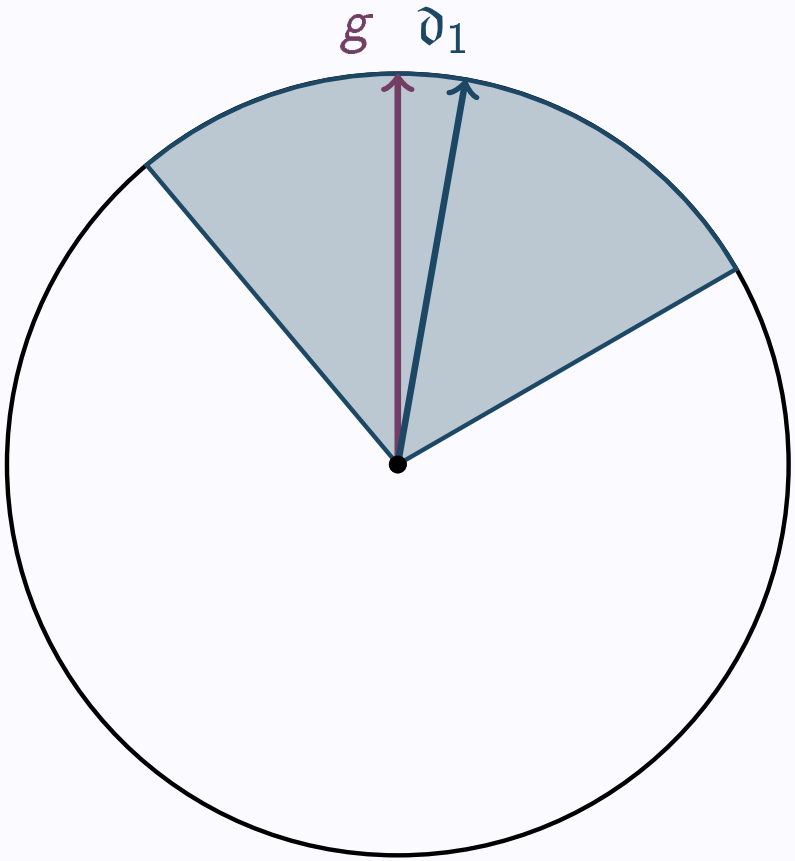


Deux directions uniformes suffisent, pas une



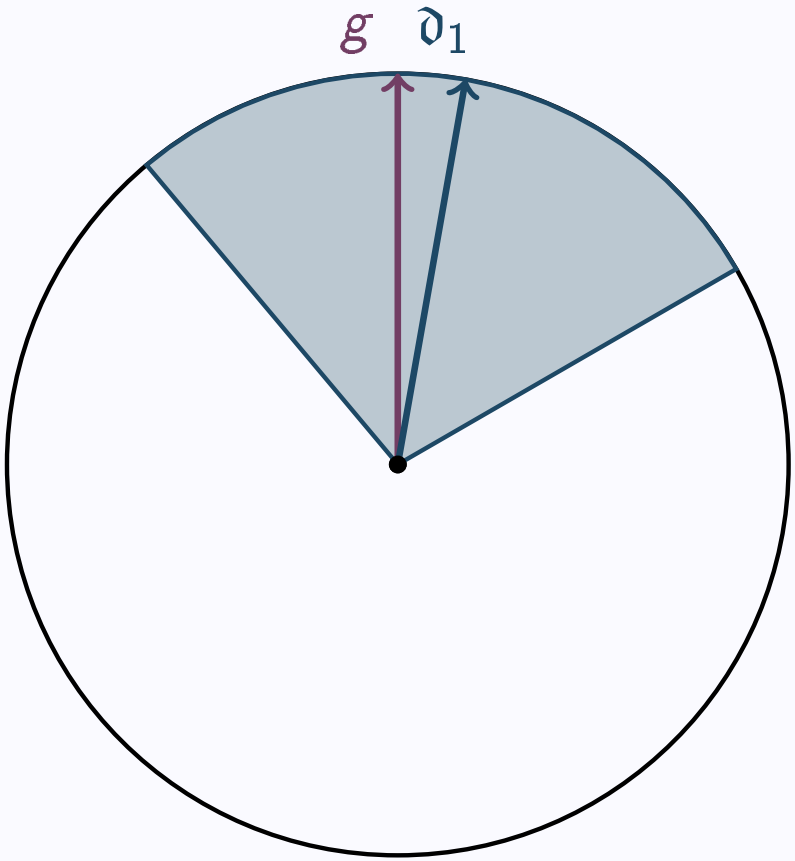
$$\mathfrak{d}_1 \sim \mathcal{U}(\mathbb{S}^1) \Rightarrow \forall \kappa \in (0,1), \quad \mathbb{P}\left(\mathsf{cm}\left(\mathfrak{d}_1, g\right) = \mathfrak{d}_1^\top g \geq \kappa\right) < 1/2.$$

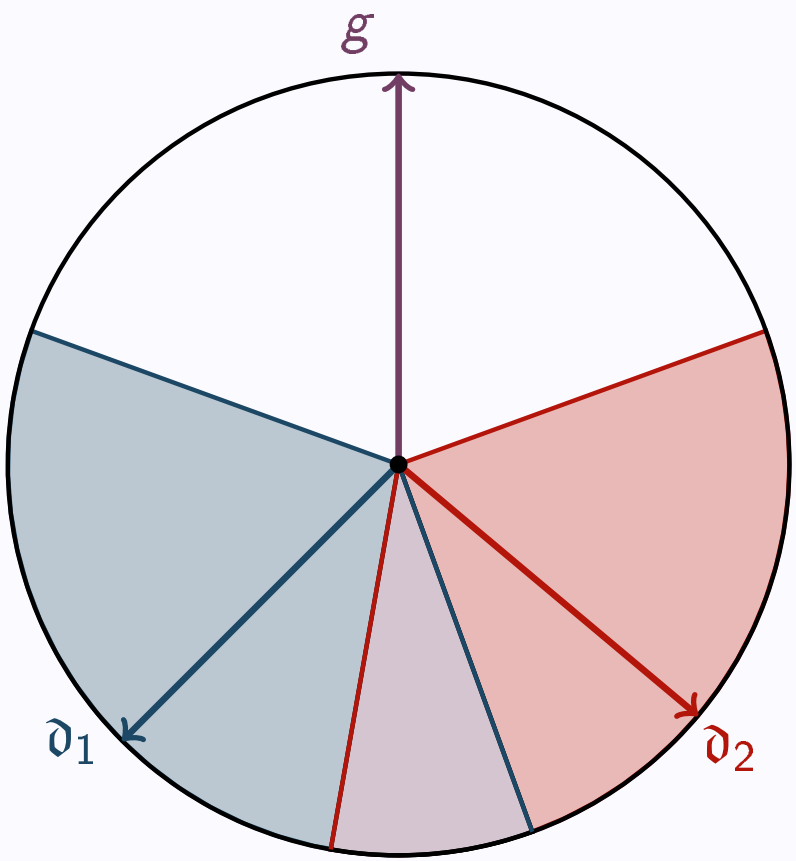
Deux directions uniformes suffisent, pas une



$$\mathfrak{d}_1 \sim \mathcal{U}(\mathbb{S}^1) \Rightarrow \forall \kappa \in (0,1), \quad \mathbb{P}\left(\mathsf{cm}\left(\mathfrak{d}_1, g\right) = \mathfrak{d}_1^\top g \geq \kappa\right) < 1/2.$$

Deux directions uniformes suffisent, pas une





$$\mathfrak{d}_1 \sim \mathcal{U}(\mathbb{S}^1) \Rightarrow \forall \kappa \in (0,1), \quad \mathbb{P}\left(\mathrm{cm}\left(\mathfrak{d}_1, g\right) = \mathfrak{d}_1^\top g \geq \kappa\right) < 1/2.$$

$$\mathfrak{d}_1, \mathfrak{d}_2 \sim \mathcal{U}(\mathbb{S}^1) \Rightarrow \exists \kappa^* \in (0,1), \quad \mathbb{P}\left(\mathrm{cm}\left(\{\mathfrak{d}_1, \mathfrak{d}_2\}, g\right) \geq \kappa^*\right) > 1/2.$$

Outline

Contraintes linéaires d'égalité

$$\begin{cases} \min_{x \in \mathbb{R}^n} & f(x) \\ \text{s.t.} & Ax = b. \end{cases}$$

- Equivalent au problème sans contraintes $\min_{\tilde{x} \in \mathbb{R}^{n-m}} f(x_0 + W\tilde{x})$ où $W \in \mathbb{R}^{n \times (n-m)}$ est une base pour null(A) et $Ax_0 = b$.
- Les résultats déterministes et probabilistes restent valables !

Deux problèmes d'optimisation avec contraintes linéaires

Contraintes linéaires d'égalité

$$\begin{cases} \min_{x \in \mathbb{R}^n} & f(x) \\ \text{s.t.} & Ax = b. \end{cases}$$

- Equivalent au problème sans contraintes $\min_{\tilde{x} \in \mathbb{R}^{n-m}} f(x_0 + W\tilde{x})$ où $W \in \mathbb{R}^{n \times (n-m)}$ est une base pour null(A) et $Ax_0 = b$.
- Les résultats déterministes et probabilistes restent valables !

Contraintes d'intervalle

$$\begin{cases} \min_{x \in \mathbb{R}^n} & f(x) \\ \text{s.t.} & l \leq x \leq u. \end{cases}$$

- En pratique (déterministe) : Utiliser $D_\oplus = \{e_1, \ldots, e_n, -e_1, \ldots, -e_n\}$ permet de converger et de se déplacer parallèlement aux contraintes.

Algorithme (version déterministe)

1. **Initialisation:** $x_0 \in \mathbb{R}^n, \alpha_0 > 0, 0 < \theta < 1 \leq \gamma$.
2. **Pour** $k = 0, 1, 2, \ldots$
   - Choisir un ensemble D_k d'au plus r vecteurs.
   - Si il existe $d_k \in D_k$ tel que $x_k + \alpha_k d_k$ **est admissible** and

$$f(x_k + \alpha_k\, d_k) < f(x_k) - \alpha_k^2,$$

     alors *(k réussie)* poser $x_{k+1} := x_k + \alpha_k\, d_k$ et $\alpha_{k+1} := \gamma\, \alpha_k$.
   - Sinon *(k non réussie)* poser $x_{k+1} := x_k$ et $\alpha_{k+1} := \theta\, \alpha_k$.

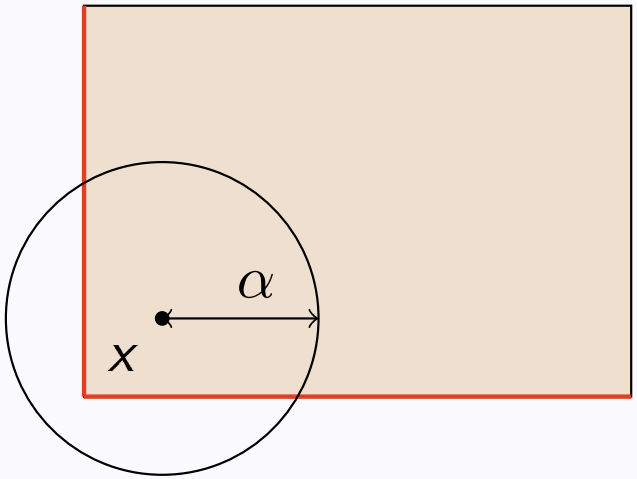
- Domaine admissible : $\mathcal{F} = \{l \leq x \leq u\}$.

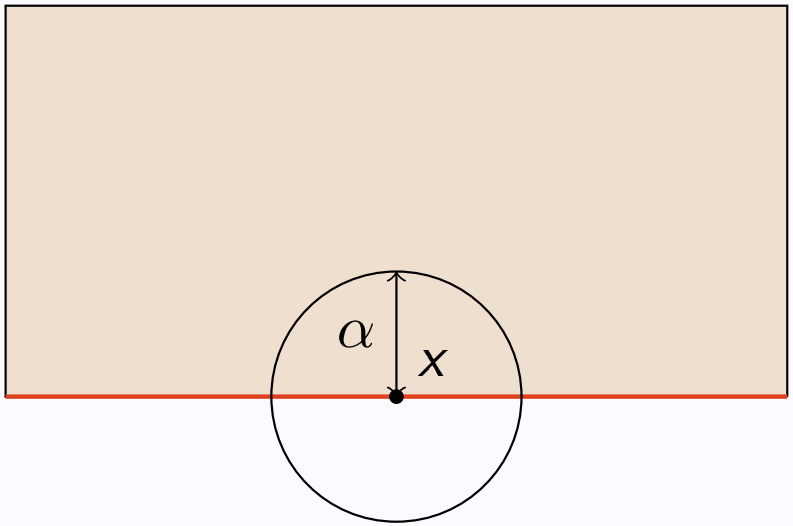
Contraintes proches

Les ensembles

$$\begin{aligned} I_u(x,\alpha) &= \{i : |u_i - [x]_i| \leq \alpha\} \\ I_l(x,\alpha) &= \{i : |l_i - [x]_i| \leq \alpha\} \end{aligned}$$

définissent les contraintes proches en $x \in \mathcal{F}$ pour $\alpha > 0$.



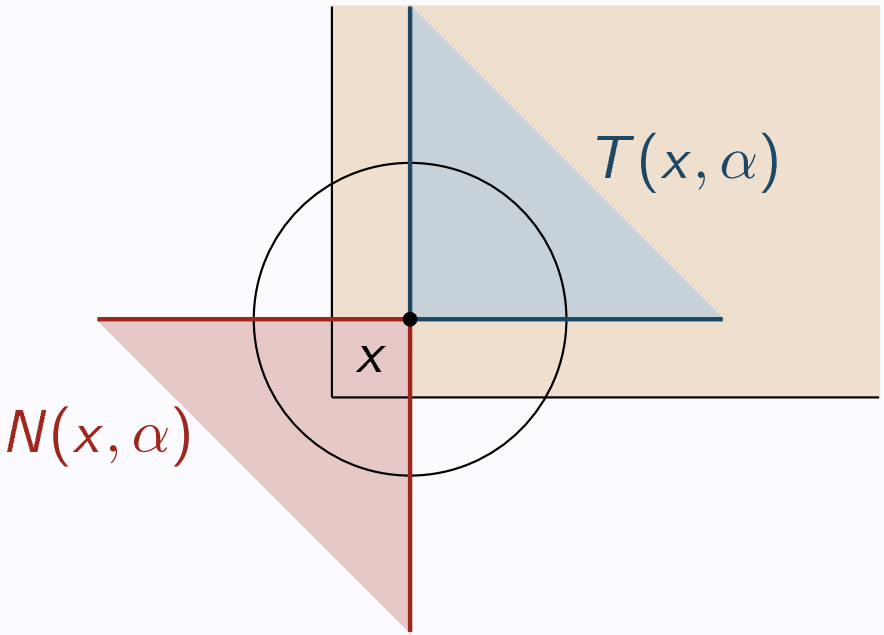


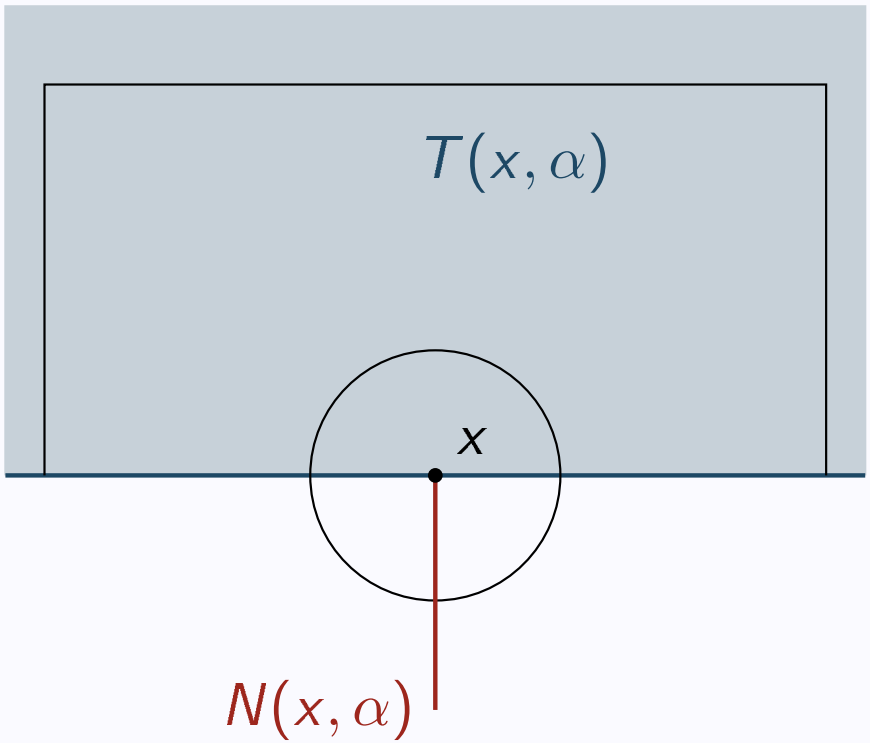
Contraintes d'intervalle (2)

- **Cône normal approché** $N(x,\alpha)$ **:** généré positivement par

$$\{e_i\}_{i\in I_u(x,\alpha)} \cup \{-e_i\}_{i\in I_l(x,\alpha)}.$$

- **Cône tangent** $T(x,\alpha)$ **:** cône polaire de $N(x,\alpha)$.





Propriété de descente réalisable

- Rappel : la mesure cosinus permet d'identifier les directions de descente

$$\mathrm{cm}(D, -\nabla f(x)) = \max_{d \in D} \frac{d^\top [-\nabla f(x)]}{\|d\| \|-\nabla f(x)\|}.$$

Descente réalisable

D est un ensemble à descente κ-admissible pour $T(x, \alpha)$ si $D \subset T(x, \alpha)$ et

$$\mathrm{cm}_{T(x,\alpha)}(D, -\nabla f(x)) = \max_{d \in D} \frac{d^\top [-\nabla f(x)]}{\|d\| \|P_{T(x,\alpha)}[-\nabla f(x)]\|} \geq \kappa.$$

Propriété de descente réalisable

- Rappel : la mesure cosinus permet d'identifier les directions de descente

$$\mathrm{cm}(D, -\nabla f(x)) = \max_{d \in D} \frac{d^\top [-\nabla f(x)]}{\|d\| \|-\nabla f(x)\|}.$$

Descente réalisable

D est un ensemble à descente κ-admissible pour $T(x, \alpha)$ si $D \subset T(x, \alpha)$ et

$$\mathrm{cm}_{T(x,\alpha)}(D, -\nabla f(x)) = \max_{d \in D} \frac{d^\top [-\nabla f(x)]}{\|d\| \|P_{T(x,\alpha)}[-\nabla f(x)]\|} \geq \kappa.$$

- Avec des ensembles à descente κ-admissible : convergence + bornes sur la complexité (analyse similaire au cas sans contraintes).
- $D_\oplus \cap T(x, \alpha)$ est toujours à descente $\frac{1}{\sqrt{n}}$-admissible.

Descente admissible probabiliste

Définition

Une suite $\{\mathfrak{D}_k\}$ est à descente (probabiliste) (p, κ)-admissible si

$$\begin{aligned} \mathbb{P}\left(\mathrm{cm}_{T_0}\left(\mathfrak{D}_0, -\nabla f(x_0)\right) \geq \kappa\right) &\geq p \\ \forall k \geq 1, \quad \mathbb{P}\left(\mathrm{cm}_{T_k}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) \geq \kappa \,\middle|\, \mathfrak{D}_0, \ldots, \mathfrak{D}_{k-1}\right) &\geq p. \end{aligned}$$

avec $T_k = T(X_k, A_k)$.

Garanties théoriques

Su $\{\mathfrak{D}_k\}$ est à descente (p, κ)-admissible avec $p > p_0$,

- **Convergence presque sûre vers un point stationnaire**;
- **Borne de complexité en probabilité :**

$$\mathbb{P}\left(N_\epsilon \leq \mathcal{O}\left(\frac{r(\kappa\,\epsilon)^{-2}}{p - p_0}\right)\right) \geq 1 - \exp\left(-\mathcal{O}\left(\frac{p - p_0}{p}(\kappa\epsilon)^{-2}\right)\right).$$

Directions aléatoires

Principales difficultés

- Définir des ensembles à descente probabiliste admissible.
- Estimer r et κ.
- Utiliser **moins de directions** que dans le cas déterministe ?

Directions aléatoires

Principales difficultés

- Définir des ensembles à descente probabiliste admissible.
- Estimer r et κ.
- Utiliser **moins de directions** que dans le cas déterministe ?

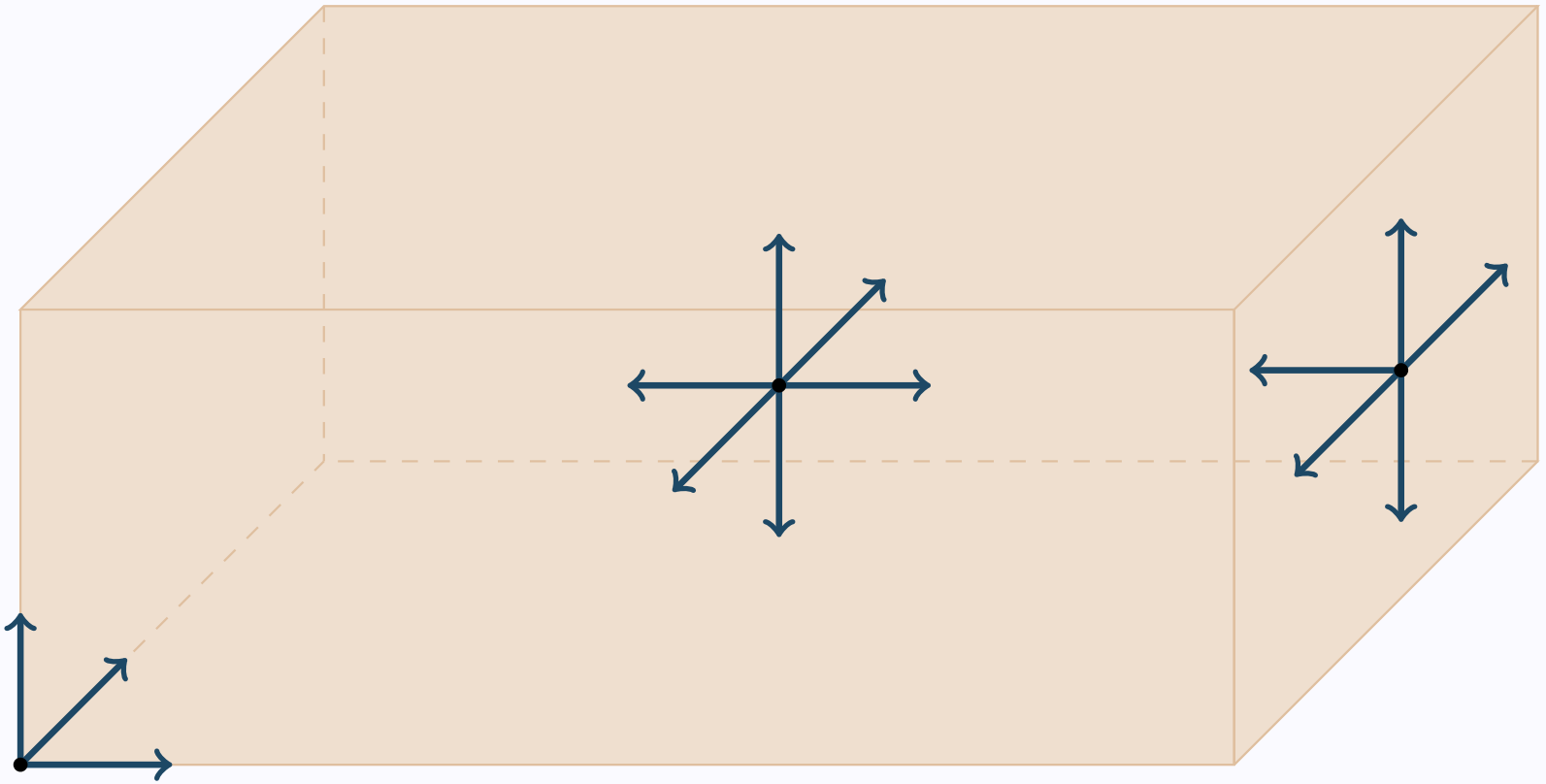
Nos techniques

- Basées sur les générateurs du cône tangent (le choix déterministe);
- Directions aléatoires mais admissibles;
- **Au pire aussi coûteuses que la stratégie déterministe**.

Premier choix de directions

Echantillonnage aléatoire parmi les générateurs

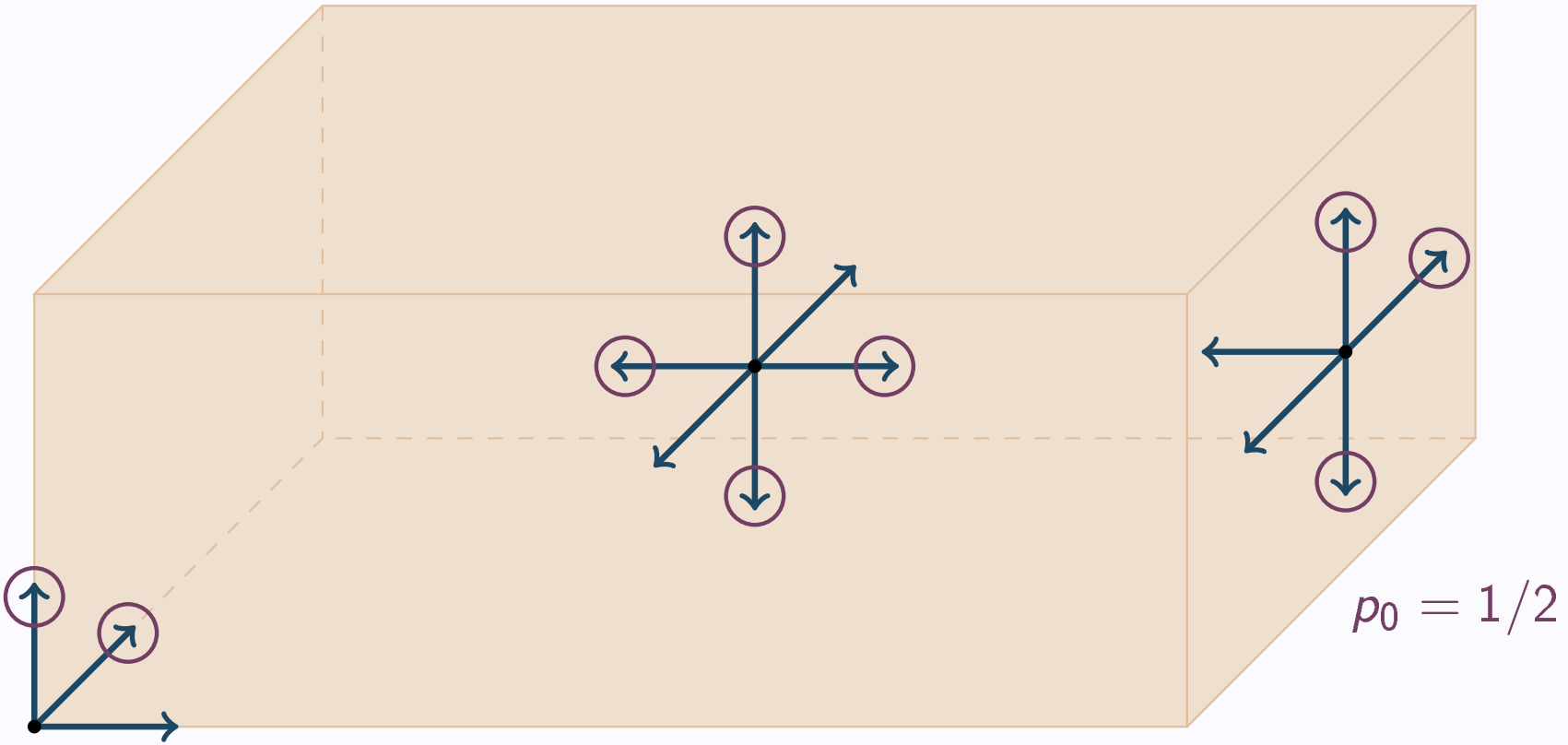
1. Calcul d'un ensemble déterministe V_k générant T_k;



Premier choix de directions

Echantillonnage aléatoire parmi les générateurs

1. Calcul d'un ensemble déterministe V_k générant T_k;
2. Tirer au hasard $\mathfrak{D}_k \subset V_k$ de taille $> |V_k| p_0$;
3. $\{\mathfrak{D}_k\}$ est à descente (p, κ)-admissible avec $p > p_0$.



Utiliser des sous-espaces et (donc) moins de directions

Principe

- Cas sans contraintes : besoin de peu de directions;
- Idem avec contraintes d'égalité : problème sans contraintes dans le noyau de A;
- Bénéfique d'exploiter les sous-espaces non contraints ?

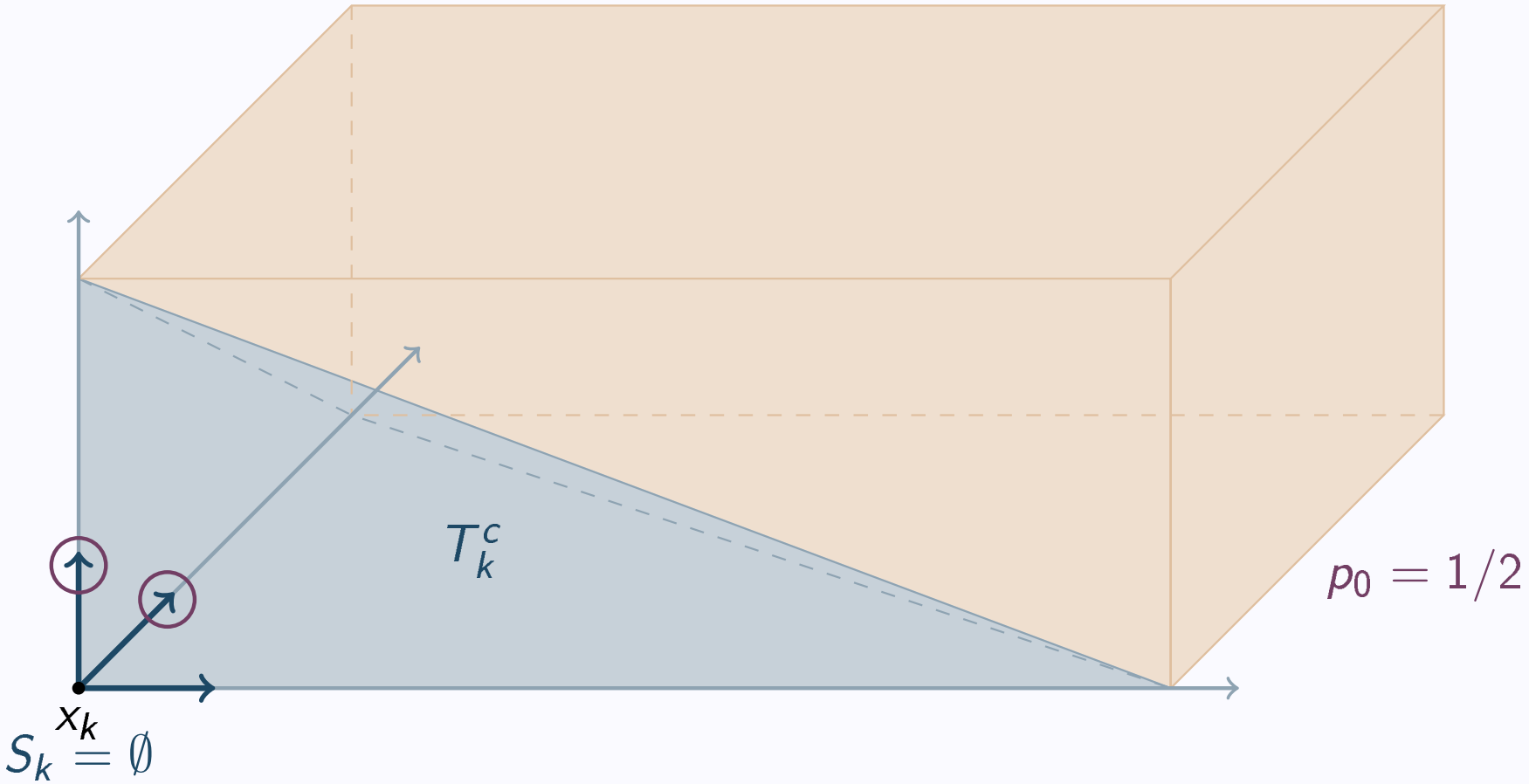
Lemme

Soit S_k un sous-espace inclus dans un cône T_k. Alors on a $T_k = S_k + T_k^c$, où T_k^c est un cône inclus dans $S_k^\perp$.

Second choix de directions

Deux types de vecteurs

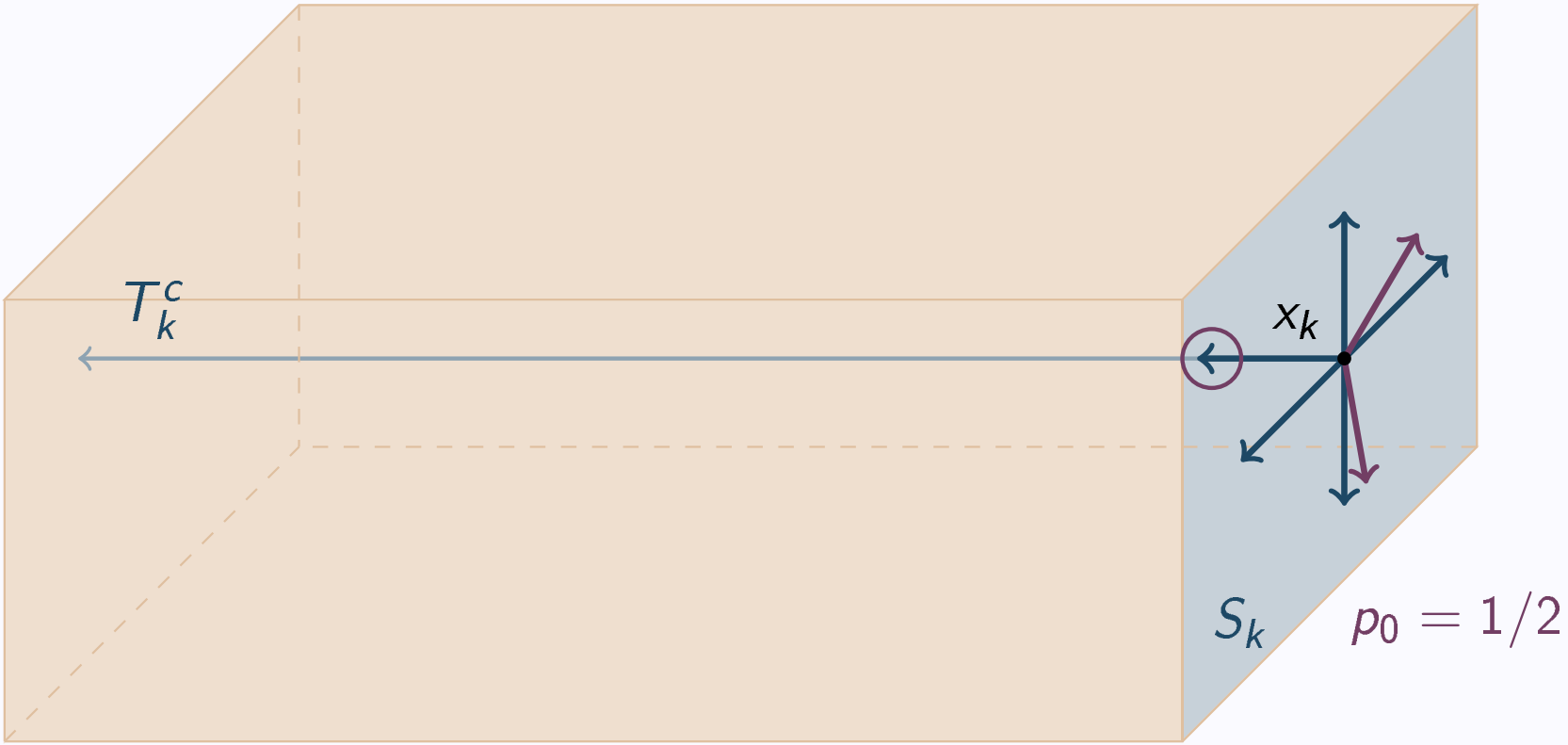
- Sous-espace S_k : Directions aléatoires;
- Complément T_k^c : Sous-ensemble aléatoire des générateurs.



Second choix de directions

Deux types de vecteurs

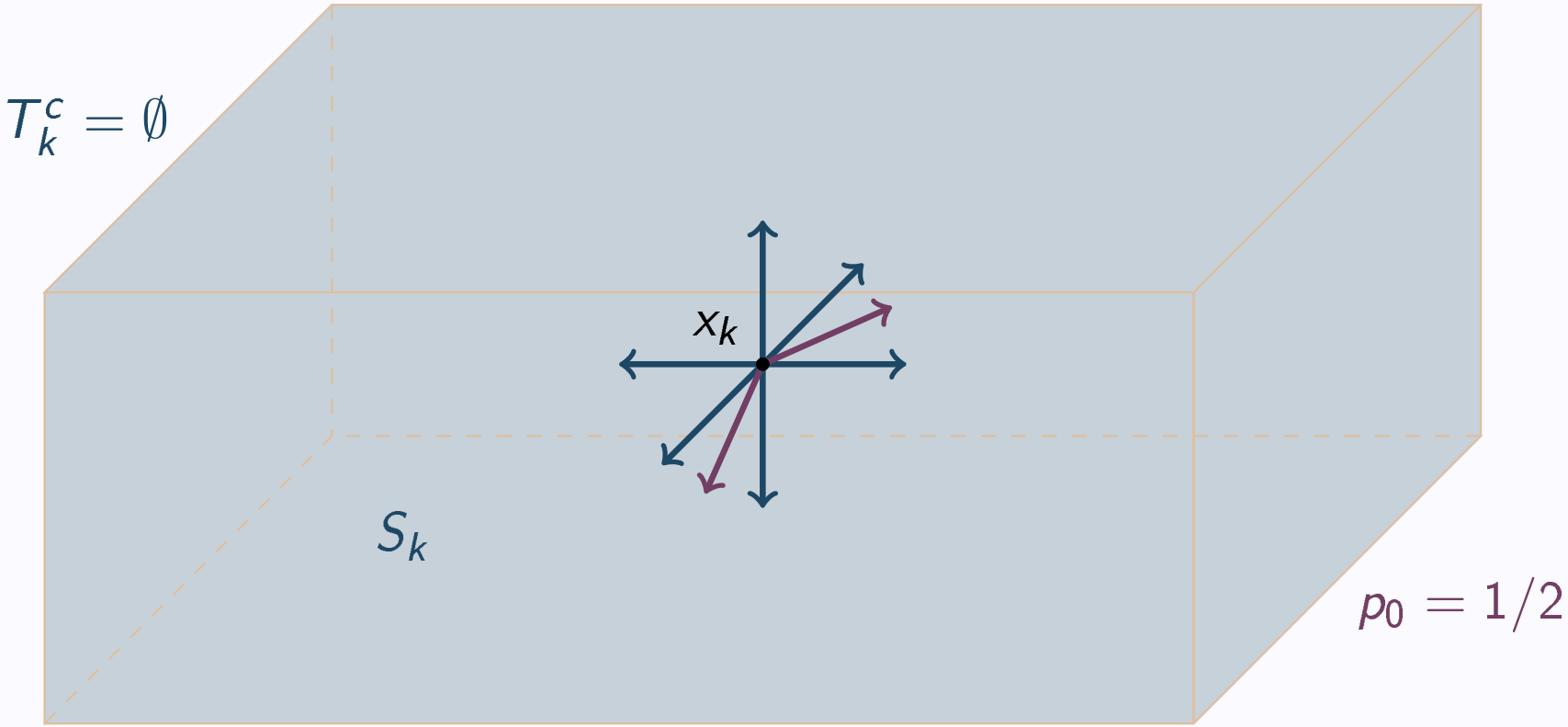
- Sous-espace S_k : Directions aléatoires;
- Complément T_k^c : Sous-ensemble aléatoire des générateurs.



Second choix de directions

Deux types de vecteurs

- Sous-espace S_k : Directions aléatoires;
- Complément T_k^c : Sous-ensemble aléatoire des générateurs.



Impact sur la complexité

- Borne générale : $\mathcal{O}\left(r\kappa^{-2}\epsilon^{-2}\right)$.

Comparaison - Egalités linéaires

Méthode	r	κ	Borne
Déterm.	$2(n-m)$	$\frac{1}{\sqrt{n-m}}$	$\mathcal{O}\left((n-m)^2\epsilon^{-2}\right)$
Proba. 1	$\mathcal{O}(2(n-m)p_0)$	$\frac{1}{\sqrt{n-m}}$	$\mathcal{O}\left((n-m)^2\epsilon^{-2}\right)$
Proba. 2 (subspace)	$\mathcal{O}(1)$	$\frac{\tau}{\sqrt{n-m}}$	$\mathcal{O}\left((n-m)\epsilon^{-2}\right)$

Comparaison - Contraintes d'intervalle sur $n_b < n$ variables seulement

Méthode	r	κ	Borne
Déterm.	$2n$	$\frac{1}{\sqrt{n}}$	$\mathcal{O}\left(n^2\epsilon^{-2}\right)$
Proba. 1	$\mathcal{O}(2np_0)$	$\frac{1}{\sqrt{n}}$	$\mathcal{O}\left(n^2\epsilon^{-2}\right)$
Proba. 2 (subspace)	$\mathcal{O}(1)+\mathcal{O}\left(n_b\, p_0\right)$	$\frac{1}{\sqrt{n}}$	$\mathcal{O}\left(n\, n_b\epsilon^{-2}\right)$

Résultats numériques - contraintes d'intervalle

- Comparaison avec le solveur `patternsearch` de MATLAB.

Quatre méthodes

Nom	Directions dans $T(x_k, \alpha_k) = T_k = S_k + T_k^c$	Garanties
`dspfd-0`	$D_\oplus \cap T_k$, ordre aléatoire	Déterm.
`dspfd-1`	Tirage aléatoire dans $D_\oplus \cap T_k$	Proba.
`dspfd-2`	Vecteurs dans S_k/tirage dans $D_\oplus \cap T_k^c$	Proba.
`matlab`	$D_\oplus \cap T(x_k, t\alpha_k), t \in (0,1)$	Déterm.

Profils de performance

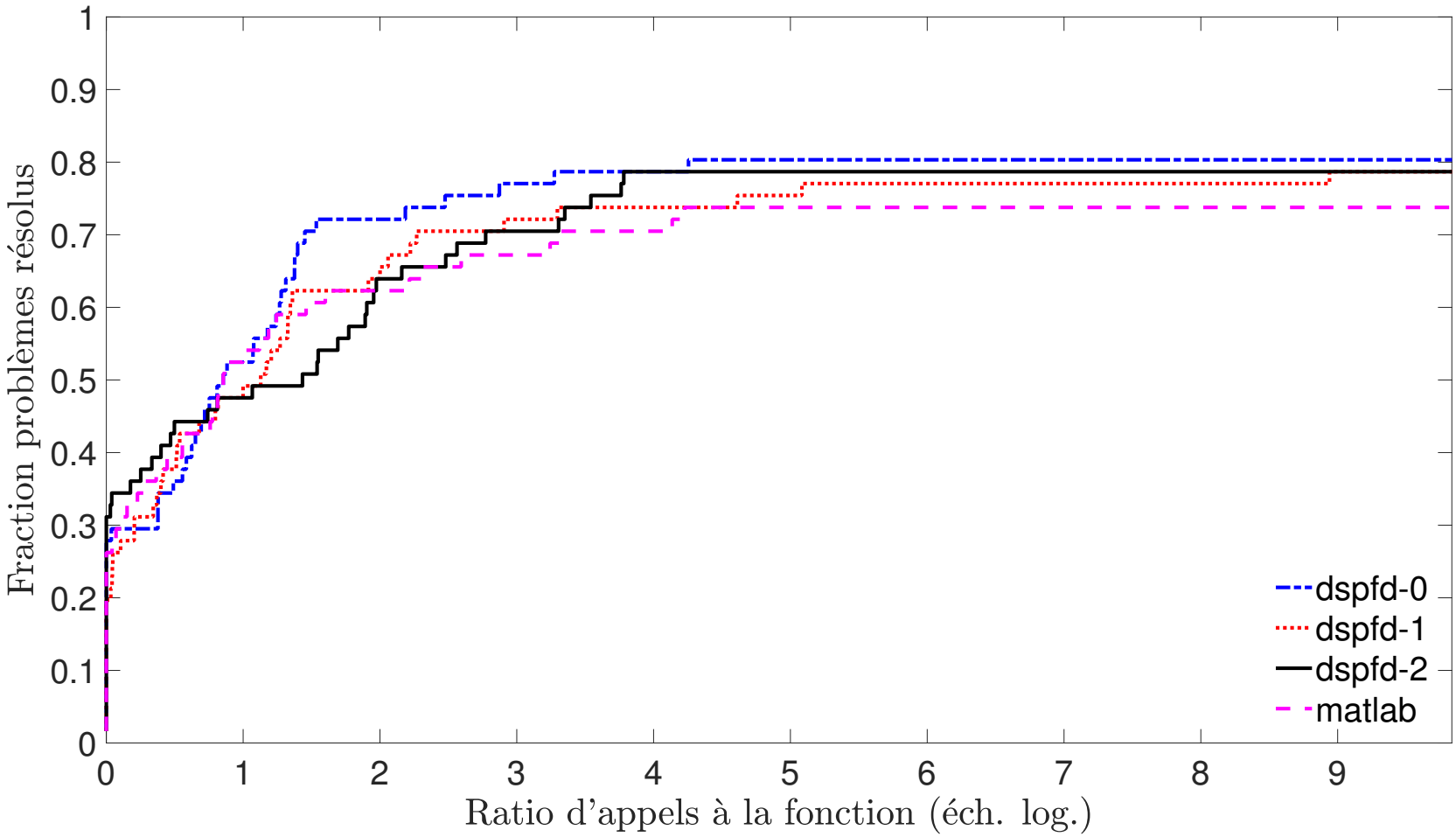
- Critère : # d'appels à f (budget de $2000n$) pour satisfaire

$$f(x_k) - f_{best} < 10^{-3}(f(x_0) - f_{best}).$$

- Problèmes de la bibliothèque CUTEst.

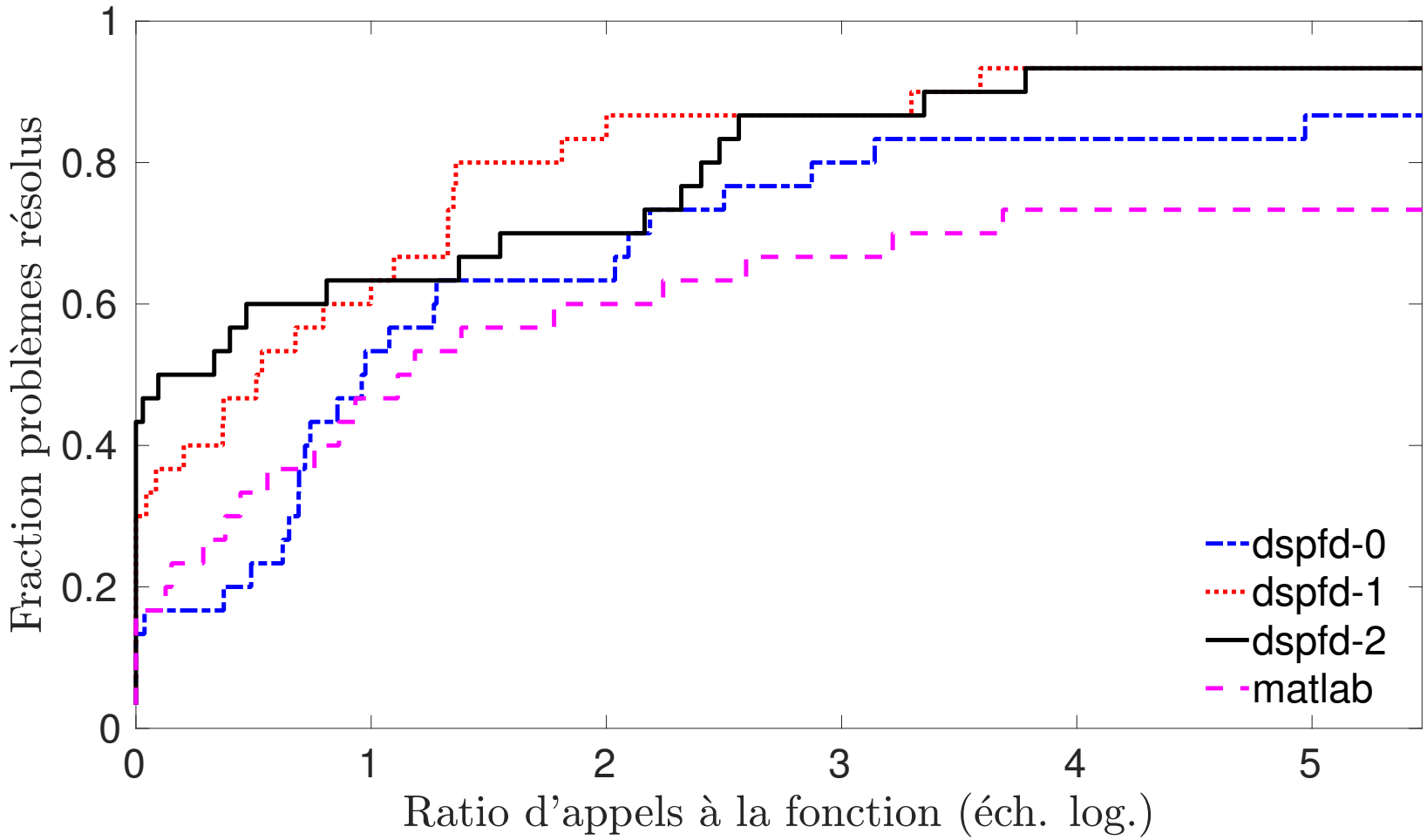
Profils avec contraintes d'intervalles

- 63 problèmes avec contraintes d'intervalles, de petites tailles : $2 \leq n \leq 20$.



Profils avec contraintes d'intervalles (2)

- 31 problèmes avec contraintes d'intervalles, **de tailles plus grandes :** $20 \leq n \leq 52$.



Outline

Contexte

$$\min_{x \in \mathbb{R}^n} f(x)$$

- f de classe $\mathcal{C}^2$,
- f en général non convexe.

Du premier ordre au second dans les algorithmes

- L'accès à la matrice Hessienne est coûteux...
- ...tout comme l'algèbre linéaire associée:
  - Vecteurs propres;
  - Systèmes linéaires.

Peut-on s'aider de propriétés probabilistes ?

Un thème de recherche actif

Analyse probabiliste de méthodes d'algèbre linéaire

Avec un point initial aléatoire...

- La puissance itérée trouve un vecteur propre à ϵ près en $\mathcal{O}\left(\epsilon^{-1}\right)$ itérations;
- La méthode de Lanczos trouve un vecteur propre à ϵ près en $\mathcal{O}\left(\epsilon^{-1/2}\right)$ itérations.

...**avec une forte probabilité**.

Utilité

- En lien avec les méthodes de premier ordre;
- Pour problèmes convexes et non convexes.

Convergence d'ordre deux

Points clés

- Echapper aux points selles;
- Détecter les valeur propres négatives de la matrice Hessienne;
- Utiliser des **directions de courbure négative**;

Les meilleures méthodes en terme de complexité garantissent la convergence à l'ordre deux !

Notre approche

- Revisiter les méthodes efficaces en pratique;
- Via leur analyse de complexité;
- Incorporer des progrès récents.

Blocs de base

- Méthodes de Newton-Krylov (ex : gradient conjugué):
  Basées sur des produits matrice/vecteur.
- Recherches linéaires;
- Analyse probabiliste.

Conclusions

Problèmes sans contraintes

- Convergence via propriétés probabilistes.
- Moins d'évaluations en théorie et en pratique.

Direct search based on probabilistic descent. Gratton, Royer, Vicente and Zhang, *SIAM J. Optim.*, 2015.

Problèmes avec contraintes linéaires

- Descente probabiliste admissible.
- Se ramener à des sous-espaces "non contraints".
- Méthode efficace en pratique.

Direct search based on probabilistic feasible descent for bound and linearly constrained problems. Gratton, Royer, Vicente and Zhang, *Submitted*, 2017.

A suivre

- Cas sans dérivées
  - Contraintes non linéaires;
  - Parallélisation.
- Contexte général
  - Etude de complexité;
  - **Courbure négative**.

A suivre

- Cas sans dérivées
  - Contraintes non linéaires;
  - Parallélisation.
- Contexte général
  - Etude de complexité;
  - **Courbure négative**.

Merci de votre attention !

croyer2@wisc.edu