Opérateurs monotones aléatoires et application à l'optimisation stochastique

Adil SALIM

Télécom ParisTech

26 novembre 2018

Sommaire

Contexte : approximation stochastique

$$\min_{x \in \mathsf{X}} F(x)$$

$F : \mathsf{X} \to \mathbb{R}$ convexe lisse, X euclidien.

- Traitement du signal/Apprentissage en ligne :

$$F(x) = \mathbb{E}_{\xi}(f(\xi, x))$$

  V.a. ξ : donnée aléatoire.

- Sommes finies : $N > 0$ grand

$$F(x) = \frac{1}{N} \sum_{i=1}^{N} f(i, x).$$

  $F(x) = \mathbb{E}_{\xi}(f(\xi, x))$ où $\xi \sim \mathrm{Unif}(1, \ldots, N)$.

- Optimisation distribuée asynchrone
- ...

Algorithme du gradient stochastique

Algorithme [Robbins, Monro'51] : (ξ_n) suite de v.a,

$$x_{n+1} = x_n - \gamma_{n+1} \nabla_x f(\xi_{n+1}, x_n)$$

où $\gamma_n > 0$.

Méthode de l'EDO [Ljung'77, Kushner'77] :
Trajectoire interpolée proche de l'EDO

$$\dot{x}(t) = -\nabla F(x(t)), \quad t \geq 0, \quad x(0) = x_0.$$

Cas non lisse?

Algorithme du gradient proximal

Présence de contraintes ou de régularisations.

$$\min_{x \in \mathsf{X}} F(x) + G(x)$$

où G convexe, s.c.i (non lisse).

Algorithme du gradient proximal : $\gamma > 0$,

$$x_{n+1} = \mathrm{prox}_{\gamma G}(x_n - \gamma \nabla F(x_n))$$

où [Moreau'62]

$$\mathrm{prox}_G(x) := \arg\min_{y \in \mathsf{X}} \frac{1}{2}\|y - x\|^2 + G(y).$$

Parfois difficile à évaluer.

Algorithme du gradient proximal stochastique

$$F(x) = \mathbb{E}_\xi(f(\xi, x)), \quad G(x) = \mathbb{E}_\xi(g(\xi, x))$$

Algorithme du **gradient proximal stochastique** :

$$x_{n+1} = \mathrm{prox}_{\gamma_{n+1} g(\xi_{n+1}, \cdot)}(x_n - \gamma_{n+1} \nabla_x f(\xi_{n+1}, x_n))$$

$\gamma_n > 0$, (ξ_n) i.i.d.

Exemples : $G(x) = \mathbb{E}_\xi(g(\xi, x))$.

- Contraintes : $G = \iota_{\mathcal{C}}$.
  Si $\mathcal{C} = \bigcap_{s=1}^{p} C_s$, alors $G(x) = \mathbb{E}_\xi(\iota_{C_\xi}(x))$
- Régularisations : Lasso structuré (groupes, graphes)
- Fonctions de coût (logistique, SVM)

Retour à la méthode de l'EDO

Inclusion Differentielle (ID)

$$\dot{x}(t) \in -(\nabla F + \partial G)(x(t)), \quad t \geq 0.$$

L'EDO devient cette ID.

Autres problèmes et algorithmes

1. Problèmes : minimiseurs, points selles, solutions d'inégalités variationnelles
2. Variété d'algorithmes déterministes : gradient proximal, Douglas-Rachford, ADMM, Chambolle-Pock, Vu-Condat.

Cadre théorique

1. Opérateur monotone qui généralise sous-différentiel convexe
2. Algorithme Forward-Backward qui généralise gradient proximal

Version stochastique?

Idée :

- Opérateurs monotones aléatoires
- Forward-Backward stochastique
- ID monotone

Opérateurs monotones aléatoires et application à l'optimisation stochastique

Sommaire

Opérateurs monotones maximaux

Opérateur monotone sur X. $A : X \to 2^X$

$$\forall x, y, \quad \forall u \in A(x), v \in A(y), \quad \langle u - v, x - y \rangle \geq 0.$$

Exemple :

- $A = \partial G$, G convexe.
- Si $X = \mathbb{R}$, monotone $\Longleftrightarrow$ croissant.

$\mathscr{M}(X) = \{\text{Opérateurs monotones maximaux}\}$: Graphe maximal.

Opérateurs monotones maximaux

Problème : Trouver

$$Z(\mathrm{A}+\mathrm{B}) := \{x \in \mathrm{X} \text{ tel que } 0 \in \mathrm{A}(x) + \mathrm{B}(x)\}$$

où $\mathrm{A}, \mathrm{B} \in \mathscr{M}(\mathrm{X})$.

Exemples : $\mathrm{A} = \partial G$, $\mathrm{B} = \nabla F$, $Z(\mathrm{A}+\mathrm{B}) = \arg\min F + G$

Domaine : $\operatorname{dom} \mathrm{A} := \{x \in \mathrm{X}, A(x) \neq \emptyset\}$
Résolvante : $\gamma > 0$, [Minty'62]

$$J_{\gamma \mathrm{A}} := (I + \gamma \mathrm{A})^{-1} \quad \text{monovalué.}$$

Exemple : $\mathrm{A} = \partial G$, $J_{\gamma \mathrm{A}} = \operatorname{prox}_{\gamma G}$.

Algorithme Forward-Backward (FB)

B monovalué, continue sur X.

Algorithme **Forward Backward** :

$$x_{n+1} = J_{\gamma \mathrm{A}}(x_n - \gamma \mathrm{B}(x_n))$$

Exemples :

- $\mathrm{A} = \partial G$, $\mathrm{B} = \nabla F$, algorithme du gradient proximal
- Douglas-Rachford, ADMM, Chambolle-Pock, Vu-Condat.

Version stochastique?

Opérateurs monotones aléatoires

Opérateurs monotones aléatoires : V.a. à valeurs $\mathscr{M}(\mathrm{X})$.
$A(\xi), B(\xi)$ opérateurs monotones aléatoires, B monovalué.

Algorithme **Forward Backward stochastique** à **pas constant** :

$$x_{n+1}^{\gamma} = J_{\gamma A(\xi_{n+1})}(x_n^{\gamma} - \gamma B(\xi_{n+1}, x_n^{\gamma}))$$

- Notation : $B(\xi, x) := B(\xi)(x)$
- $(\xi_n)_n$ copies i.i.d de ξ
- **Pas** $\gamma > 0$ **constant** (pas décroissant [Bianchi, Hachem'16])
- dom $A(\xi)$ aléatoire.

Littérature

Gradient proximal stochastique

- Prox déterministe : [Nemirovski *et al.*'09], [Atchadé *et al.*'14,'17], [Rosasco *et al.*'14], [Defazio *et al.*'14]
- Prox stochastique : [Wang, Bertsekas'13,'15], [Ryu, Boyd'14], [Toulis *et al.*'15], [Patrascu, Necoara'17]

Forward Backward stochastique

- Backward déterministe : [Ouyang *et al.*'13], [Rosasco *et al.*'16], [Yurtsever *et al.*'16]
- Backward stochastique : [Rockafellar'76], [Combettes, Pesquet'16], [Bianchi'16]

Approximation stochastique avec inclusions différentielles

- Pas décroissant : [Benaïm *et al.*'05], [Faure, Roth'10], [Bianchi, Hachem'16], [Majewski *et al.*'18]
- Pas constant : [Roth, Sandholm'13]

Démarche

Pas constant : Pas de convergence p.s du processus $x^\gamma = (x_n^\gamma)_n$

- $\{x^\gamma\}_\gamma$ Famille de processus indexée par γ petit.
- Adaptation de la méthode de l'EDO
- Double régime asymptotique $n \to \infty$ puis $\gamma \to 0$

EDO devient Inclusion différentielle

Opérateur moyen : Intégrale de sélection

$$\mathcal{B}(x) := \mathbb{E}_\xi(B(\xi, x))$$
$$\mathcal{A}(x) := \overline{\{\mathbb{E}(\varphi) \,:\, \varphi \text{ integrable}, \ \varphi \in A(\xi, x) \text{ p.s.}\}}.$$

Exemple [Rockafellar, Wets'82] :
$A(\xi) = \partial g(\xi, \cdot)$, $\mathcal{A} = \partial G$ où $G(x) = \mathbb{E}_\xi(g(\xi, x))$.

Inclusion Différentielle (ID)

$$\dot{x}(t) \in -(\mathcal{A} + \mathcal{B})(x(t)), \quad t \geq 0.$$

ID monotones [Komura'67], [Brezis'73] :
Solution unique si x(0) $\in$ dom $\mathcal{A}$.

1er résultat : comportement dynamique

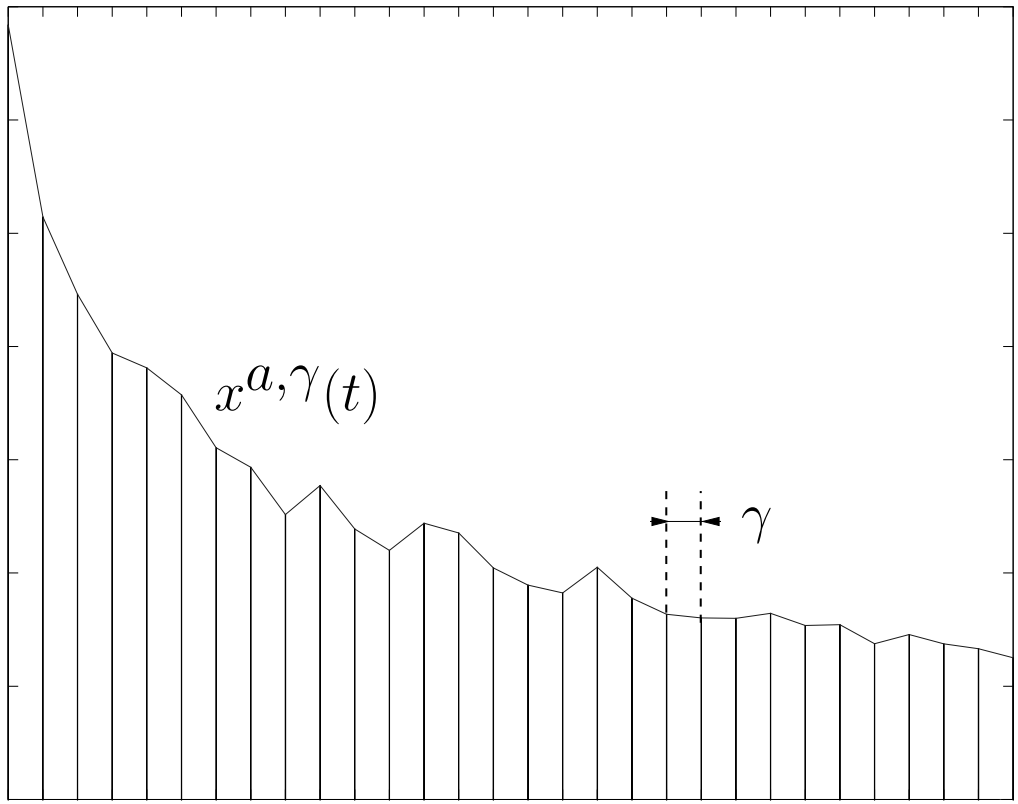


Figure 1: Processus continu interpolé : $x^{a,\gamma}(t)$ démarrant en $x^{a,\gamma}(0) = a$.

Famille $\{x^{a,\gamma}\}_\gamma$ de $C(\mathbb{R}_+, \mathrm{X})$–variables aléatoires, topologie cvg. unif. sur compacts

1er résultat : comportement dynamique

Soit $a \in \operatorname{dom} \mathcal{A}$, $\Phi(a, \cdot)$ la solution de ID issue de a, Φ flot.

Convergence de processus stochastiques : $x^{a,\gamma} \xrightarrow[\gamma \to 0]{\text{loi}} \Phi(a, \cdot)$.

1. Tension de $\{x^{a,\gamma}\}_\gamma$ (sous des hypothèses légères de **moments** et de **régularité des domaines** $\operatorname{dom} A(\xi)$).
2. Identification des valeurs d'adhérence.

Théorème [BHS'19] : **Comportement dynamique**
$\forall \varepsilon > 0,\ T > 0,\ K$ compact de $\operatorname{dom} \mathcal{A}$,

$$\sup_{a \in K} \mathbb{P}\left[\sup_{t \in [0,T]} \|x^{a,\gamma}(t) - \Phi(a,t)\| > \varepsilon \right] \xrightarrow[\gamma \to 0]{} 0.$$

2e résultat : Mesures invariantes

- $(x_n^\gamma)_n$ chaîne de Markov Feller de noyau P_γ
- $I_\gamma = \{\text{lois invariantes pour } P_\gamma\}$
- $\mathrm{Inv} = \bigcup_{\gamma\in(0,\gamma_0]} I_\gamma$

Corollaire [Haz'minskii'63] [1] : **sous réserve d'existence.**
Soit π valeur d'adhérence de Inv, $\gamma \to 0$.
Alors π invariante pour le flot :

$$\pi = \pi\Phi(\cdot, t)^{-1}, \quad t \geq 0$$

Conséquence.

1. Moyenne de π est un zéro
2. Demipositivité : π supportée par les zéros.

[1] [BHS'19] : Cas du domaine dom $A(\xi)$ aléatoire

Existence de mesures invariantes

Lemme [BHS'18] : **Stabilité**.
Si $\exists V \geq 0, \psi$ coercive,

$$P_\gamma V \leq V - \gamma\psi + \gamma^2 C. \qquad \text{(PH)}$$

Alors

1. $I_\gamma \neq \emptyset$ (classique)
2. Tension de Inv

   De plus,
3. Tension de $\{\frac{1}{n}\sum_{k=0}^{n-1} \delta_{x_k^\gamma}, n \in \mathbb{N}\}$ (en tant que v.a.)

Cas sous-différentielles, affine, etc.

Comportement asymptotique des itérées

Théorème [BHS'19] : **Comportement asymptotique**

1. $\bar{x}_n^\gamma = \frac{1}{n}\sum_{k=0}^{n-1} x_k^\gamma$

$$\forall \varepsilon > 0, \quad \limsup_{n\to\infty} \mathbb{P}\left[d\left(\bar{x}_n^\gamma, Z(\mathcal{A}+\mathcal{B})\right) \geq \varepsilon\right] \xrightarrow[\gamma\to 0]{} 0.$$

$$\limsup_{n\to\infty} d\left(\mathbb{E}(\bar{x}_n^\gamma), Z(\mathcal{A}+\mathcal{B})\right) \xrightarrow[\gamma\to 0]{} 0.$$

2. $\mathcal{A}+\mathcal{B}$ demipositif :

$$\forall \varepsilon > 0, \quad \limsup_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{P}\left[d\left(x_k^\gamma, Z(\mathcal{A}+\mathcal{B})\right) \geq \varepsilon\right] \xrightarrow[\gamma\to 0]{} 0\,.$$

Sommaire

Approximation stochastique générique

Algorithme d'approximation stochastique générique :

$$x_{n+1}^{\gamma} = x_n^{\gamma} + \gamma h_{\gamma}(\xi_{n+1}, x_n^{\gamma})$$

(ξ_n) i.i.d.

Applications :

- Modèle de files d'attentes parallèles [Gast, Gaujal'12]
- Gradient stochastique non convexe proximal [Bolte'15].

Inclusion différentielle

Méthode de l'Inclusion Différentielle [BHS'18] :
Trajectoire interpolée proche de l'ID

$$\dot{x}(t) \in H(x(t)), \quad t \geq 0.$$

$-H$ non monotone.

Principales différences :

- Multiples solutions issues de $a \in X$
- Notion de mesure invariante plus complexe [Roth, Sandholm'13].

Sommaire

Sommaire

Problème d'optimisation sous contraintes stochastiques

$$\min_{x \in \mathsf{X}} F(x) + G(x) \quad \text{s.c.} \quad \mathbf{M}x = \mathbf{p} \tag{1}$$

où

- $F(x) = \mathbb{E}_\xi(f(\xi, x))$
- $G(x) = \mathbb{E}_\xi(g(\xi, x))$
- $\mathbf{M} = \mathbb{E}_\xi(M(\xi))$
- $\mathbf{p} = \mathbb{E}_\xi(p(\xi))$

Approche

1. Points selles du Lagrangien
2. Zéros d'une somme d'opérateurs monotones
3. Opérateurs monotones comme opérateurs moyens
4. Forward-Backward stochastique à pas décroissants

Algorithme proposé

Généralisation du gradient proximal stochastique
[Bianchi, Hachem'16].

$$x_{n+1} = \mathrm{prox}_{\gamma_{n+1} g(\xi_{n+1},\cdot)} \left(x_n - \gamma_{n+1} (\widetilde{\nabla} f(\xi_{n+1}, x_n) + M(\xi_{n+1})^T \lambda_n) \right)$$
$$\lambda_{n+1} = \lambda_n + \gamma_{n+1} \left(M(\xi_{n+1}) x_n - p(\xi_{n+1}) \right) .$$

où

- $\widetilde{\nabla} f(s,x)$ sous-gradient de $f(s,\cdot)$ en x.
- $\gamma_n \downarrow 0$

Convergence de l'algorithme

Soit $\bar{x}_n = \frac{\sum_{k=1}^n \gamma_k x_k}{\sum_{k=1}^n \gamma_k}$, idem pour $\bar{\lambda}_n$.

Théorème [SBH'18]: $(\bar{x}_n, \bar{\lambda}_n) \longrightarrow_{n\to+\infty} (x_\star, \lambda_\star)$ p.s. où $x_\star$ solution de (1) et $\lambda_\star$ solution duale.

Sommaire

Problème d'optimisation sur graphe

- Graphe $G = (V, E)$
- $x \in \mathbb{R}^V$
- Variation Totale

$$\mathrm{TV}(x, G) = \sum_{\{i,j\} \in E} |x(i) - x(j)|.$$

Problème:

$$\min_{x \in \mathbb{R}^V} F(x) + \mathrm{TV}(x, G) \tag{2}$$

$F : \mathbb{R}^V \to \mathbb{R}$ convexe, lisse.

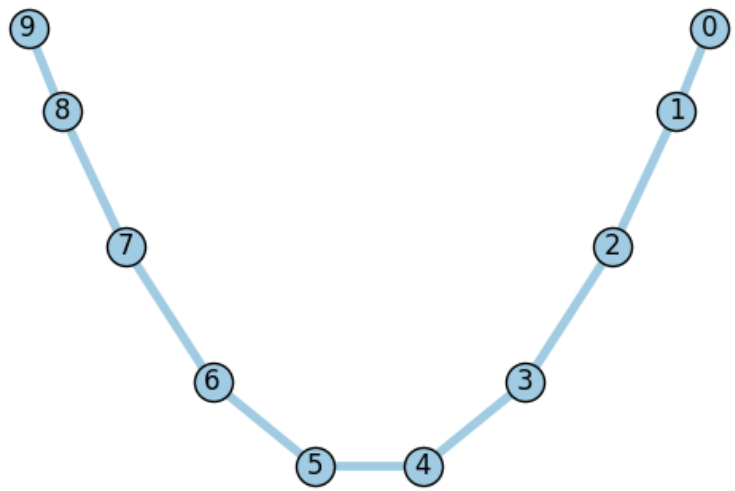
Problème

Algorithme du gradient proximal

$$x_{n+1} = \mathrm{prox}_{\gamma \mathrm{TV}(\cdot, G)}(x_n - \gamma \nabla F(x_n))$$

Evaluation de $\mathrm{prox}_{\mathrm{TV}(\cdot, G)}(y)$

- Efficace si G est une chaîne : **Algorithme Taut String** [Condat'13],[Johnson'13],[Barbero, Sra'14].



- Difficile sur les graphes non structurés

Tirage de marche aléatoires

Marche aléatoire simple et stationnaire sur G
$\xi = (v_0, \dots, v_L)$.

$$\mathbb{E}_\xi \left(\mathrm{TV}(x, \xi) \right) = \frac{L}{|E|} \mathrm{TV}(x, G).$$

Problème équivalent

$$\min_{x \in \mathbb{R}^V} LF(x) + |E| \mathbb{E}_\xi \left(\mathrm{TV}(x, \xi) \right).$$

Algorithme du gradient proximal stochastique:

$$\begin{cases} \text{Tirer la marche } \xi_{n+1}. \\ x_{n+1} = \mathrm{prox}_{\gamma_{n+1}|E|\mathrm{TV}(\cdot, \xi_{n+1})}(x_n - \gamma_{n+1} L \nabla F(x_n)) \end{cases}$$

Algorithme Snake

Principe de l'algorithme :

- Découper la marche ξ en chemins sans cycles
- $\mathrm{prox}_{\gamma \mathrm{TV}(\cdot,\xi)}$ devient une composition de $\mathrm{prox}_{\gamma \mathrm{TV}}$ sur ces chemins (Taut String)
- Preuve de convergence de l'algorithme de composition de gradients-proximaux non indépendants.

Théorème [SBH'17] : Si $\gamma_n \downarrow 0$, $x_n \xrightarrow[n \to +\infty]{} x_\star$ où $x_\star$ solution de (2).

Se généralise à d'autres régularisations sur graphe.

Illustration

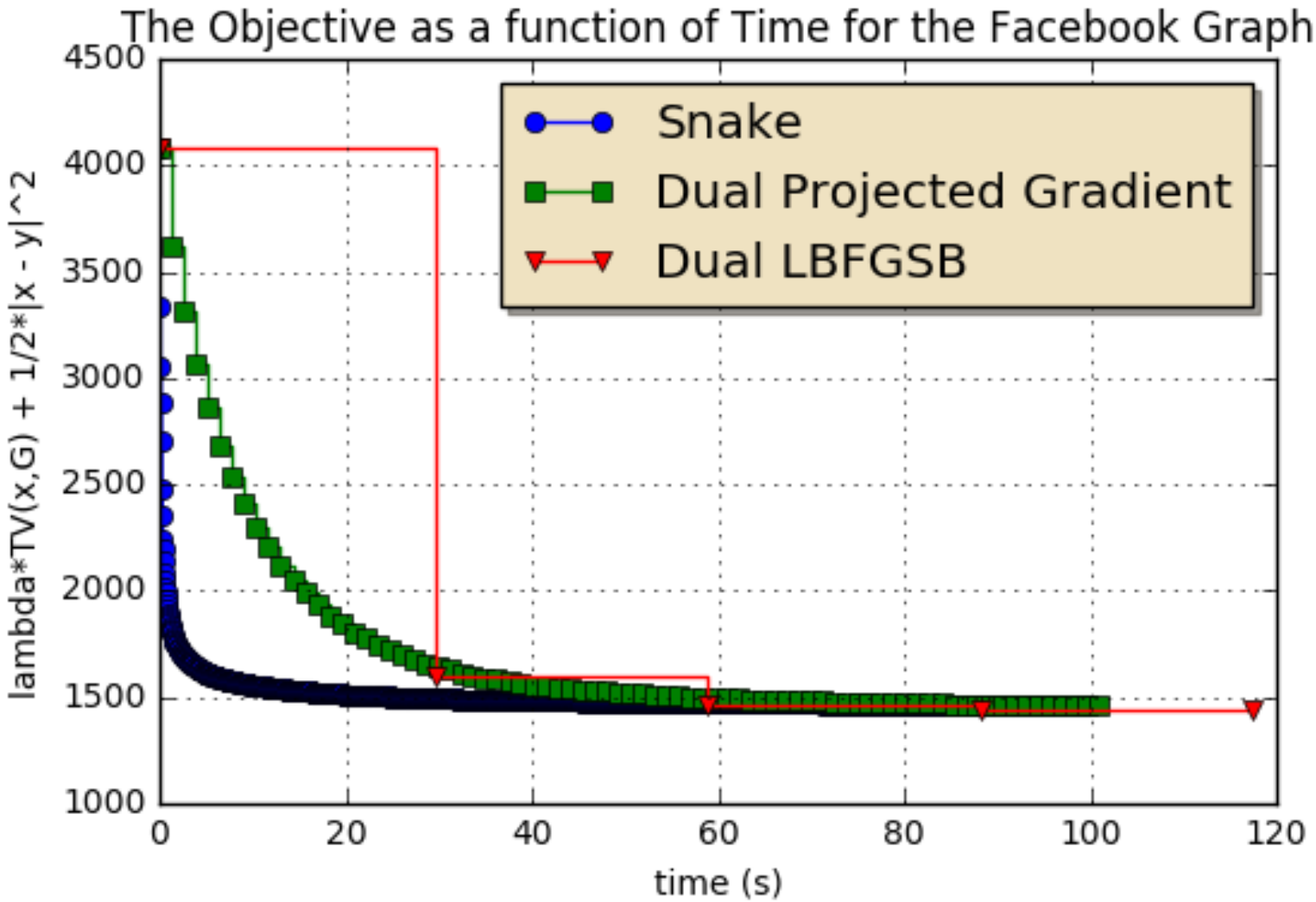


Figure 2: Snake sur le graphe "Facebook" de [Leskovec *et al.*'16].

Conclusion et perspectives

Conclusion

- Approximation stochastique pour des ID monotones et non monotones
- Primal-dual stochastique
- Optimisation sur graphes

Perspectives

- Analyse non asymptotique des algorithmes
- Méthodes de simulations basés sur l'optimisation
- Optimisation non convexe, non lisse, stochastique