# A stochastic Forward Backward algorithm with application to large graphs regularization

Adil Salim
adil-salim.github.io

Telecom ParisTech

March 7, 2018

Joint work with Pascal Bianchi and Walid Hachem

# Table of Contents

Stochastic Forward Backward algorithm
The algorithm
Asymptotic convergence analysis

Regularization over large graphs

# Stochastic Gradient algorithm

**General Problem**:

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

with $F$ smooth over $\mathcal{X}$, Euclidean space.
In ML, $\nabla F$ is often intractable.

**Constant step Stochastic Gradient algorithm** (*e.g* [Dieuleveut *et al.*'17]) :

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

with

- $\gamma > 0$
- $(\xi_n)$ iid
- $\mathbb{E}_\xi(f(x, \xi)) = F(x)$

# Proximal Stochastic Gradient algorithm

**General Problem**:

$$\min_{x \in \mathcal{X}} F(x) + R(x)$$

with $R$ nonsmooth convex over $\mathcal{X}$, $F$ smooth.

**Constant step Proximal Stochastic Gradient algorithm** (*e.g* [Rosasco *et al.*'14],[BHS'16]) :

$$x_{n+1}^{\gamma} = \mathrm{prox}_{\gamma R}(x_n^{\gamma} - \gamma \nabla f(x_n^{\gamma}, \xi_{n+1}))$$

where

$$\mathrm{prox}_{\gamma R}(x) = \arg\min_{y \in \mathcal{X}} \frac{1}{2\gamma} \|x - y\|^2 + R(y).$$

# Asymptotic Convergence: $F$ non convex and $R$ deterministic

Let $\mathcal{Z} = \{x \in E, 0 \in \nabla F(x) + \partial R(x)\}$.

Theorem [BHS'16] : If $f(\cdot, \xi)$ is not convex but $f(\cdot, \xi), R$ satisfy the Proximal-P-L condition, then,

$$\limsup_{n \to +\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{P}(d(x_k^\gamma, \mathcal{Z}) > \varepsilon) \longrightarrow_{\gamma \to 0} 0.$$

# Stochastic Proximal Gradient algorithm

What if both $\mathrm{prox}_{\gamma R}$ and $\nabla F$ are intractable?
Assume now that $F$ is **convex**.

**Stochastic Proximal Gradient algorithm** [Combettes *et al.*'16], [BHS'17] : If $F$ and $R$ are convex,

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

with

- $(\xi_n)$ iid
- $\mathbb{E}_{\xi}(f(x,\xi)) = F(x)$
- $\mathbb{E}_{\xi}(r(x,\xi)) = R(x)$.

# Asymptotic Convergence: $F$ and $R$ random

Theorem [BHS'17] : If $F$ and $R$ are convex,

$$\limsup_{n\to+\infty} \frac{1}{n}\sum_{k=1}^{n} \mathbb{P}(d(x_k^\gamma, \arg\min_{\mathcal{X}} F + R) > \varepsilon) \longrightarrow_{\gamma\to 0} 0.$$

# Proof of the Asymptotic Convergences

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

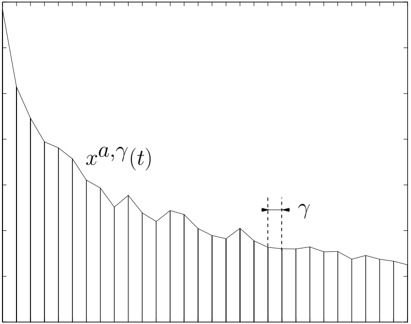



Figure 1: Continuous interpolated process : $x^{a,\gamma}(t)$ starting at $x^{a,\gamma}(0) = a$.

# First step : Dynamical behavior

The Differential Inclusion (DI) over $\mathbf{R}_+$

$$\dot{x_a}(t) \in -(\nabla F + \partial R)(x_a(t)), \quad x_a(0) = a$$

admits an unique solution $x_a$.

We look at $(x^{a,\gamma})_\gamma$ as a family of stochastic processes in $C(\mathbf{R}_+, \mathcal{X})$ in order to apply the ODE method. Under mild assumptions,

$$x^{a,\gamma} \Longrightarrow_{\gamma \to 0} x_a.$$

in the sense of the convergence of stochastic processes.

# Second step : Asymptotic behavior

We look at $(x_n^\gamma)_n$ as a Markov Chain depending on $\gamma$ in order study its stability.

**Stability assumption**:

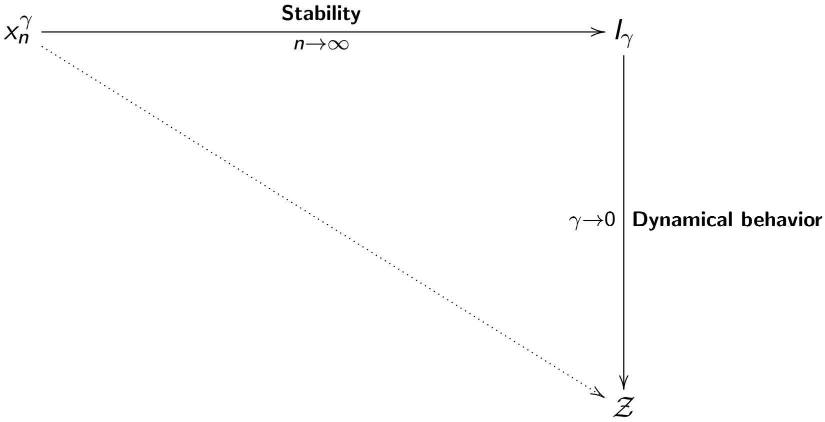
- $F + R \longrightarrow_\infty +\infty$

Then, using the dynamical behavior result,

Invariant measures for $(x_n^\gamma) \Longrightarrow_{\gamma \to 0}$ Invariant measures for the DI.

# End of the proof

Invariant measures for the DI are supported by
$\mathcal{Z} = \{x \in E, 0 \in \nabla F(x) + \partial R(x)\}$.



# Table of Contents

Stochastic Forward Backward algorithm

Regularization over large graphs
- The problem
- Application of Stochastic Proximal Gradient algorithm
- Snake algorithm

# Problem Statement

Consider

- An undirected graph $G = (V, E)$
- A vector of parameters over the nodes $x \in \mathbb{R}^V$
- The **Total Variation** (TV) regularization over $G$

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

**Our problem**:

$$\min_{x \in \mathbb{R}^V} F(x) + \mathrm{TV}(x, G) \tag{1}$$

with $F : \mathbb{R}^V \to \mathbb{R}$ convex, smooth.

# Example: Trend Filtering on Graphs [Wang *et al.*'16]

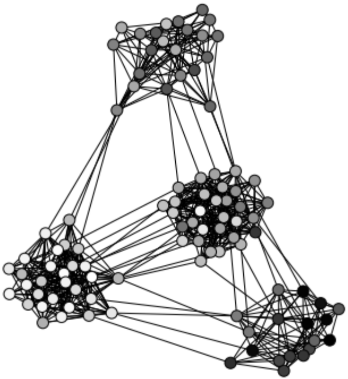


Figure 2: $\min_{x \in \mathbb{R}^V} \frac{1}{2}\|x - y\|^2 + \mathrm{TV}(x, G)$

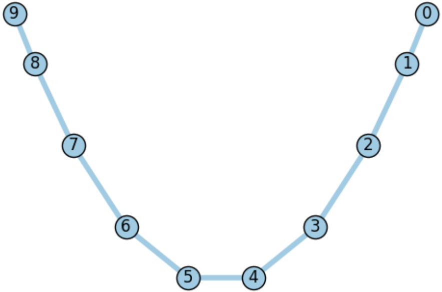
# Problem Statement

Proximal Gradient algorithm

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

The computation of $\mathrm{prox}_{\mathrm{TV}(.,G)}(y)$ is

- Fast when the graph $G$ is a path graph : **Taut String algorithm** [Condat'13],[Johnson'13],[Barbero and Sra'14].




- Difficult over general large graphs

# Sampling Random Walks

Let $L \geq 1$.
Let $\xi$ is a stationary simple random walk over $G$ with length $L+1$

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

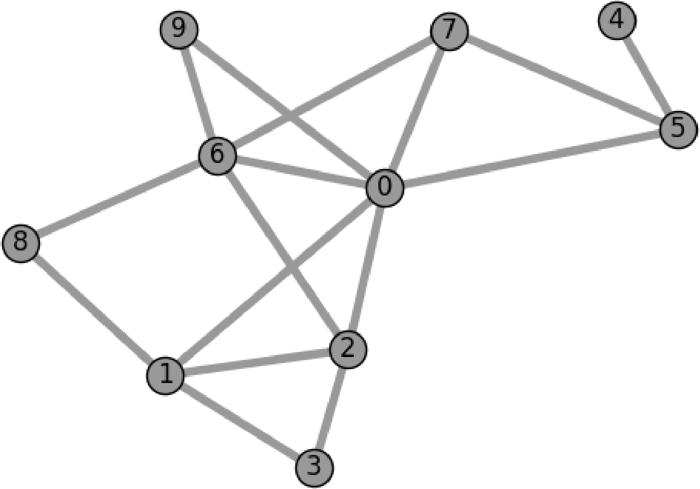
Our problem is equivalent to

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

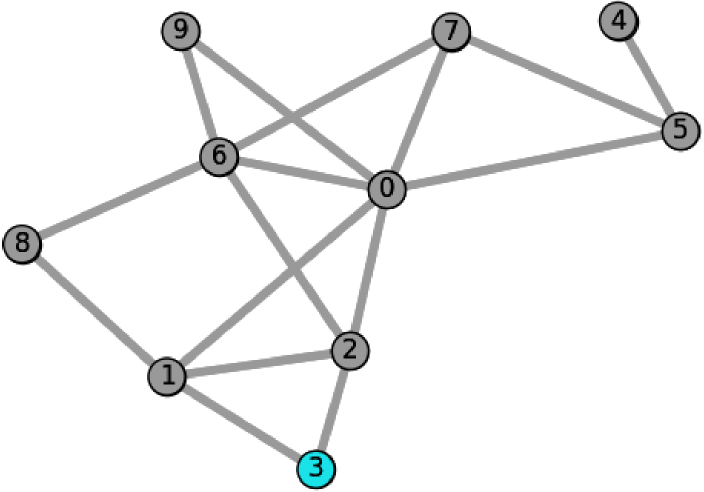
**Stochastic Proximal Gradient algorithm**:

$$\begin{cases} \text{Sample the Stationary Random Walk } \xi_{n+1} \text{ with length } L+1 \\ x_{n+1} = \mathrm{prox}_{\gamma_n |E| \mathrm{TV}(\cdot, \xi_{n+1})}(x_n - \gamma_n L \nabla F(x_n)) \end{cases}$$

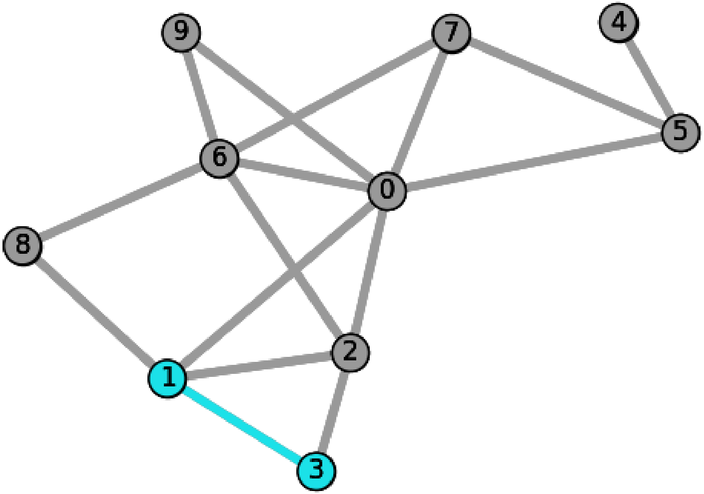
# Example : The Graph $G$



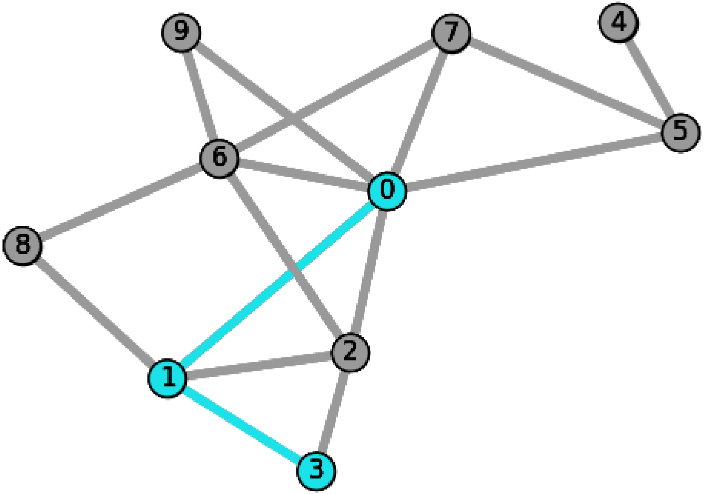
# Example : Sampling the Random Walk $\xi_{n+1}$



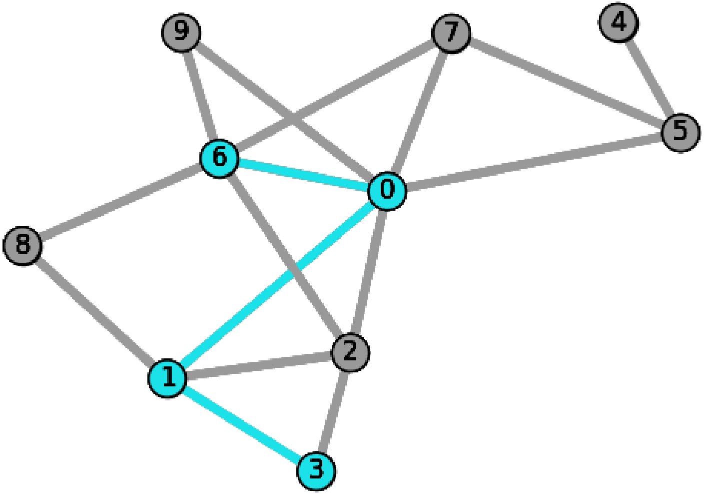
# Example : Sampling the Random Walk $\xi_{n+1}$



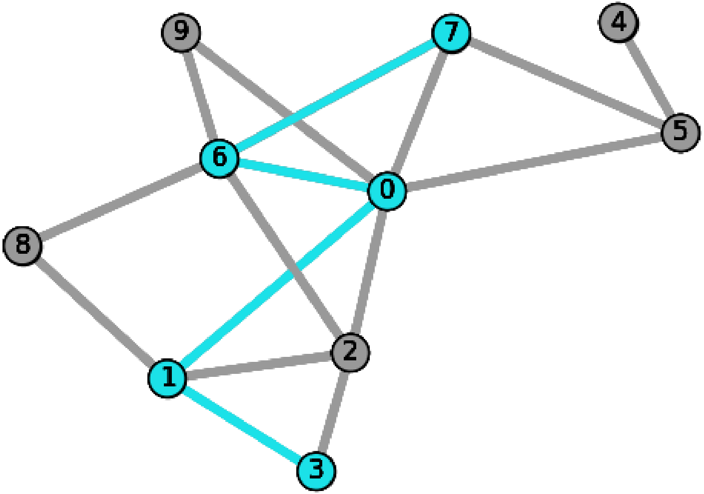
# Example : Sampling the Random Walk $\xi_{n+1}$



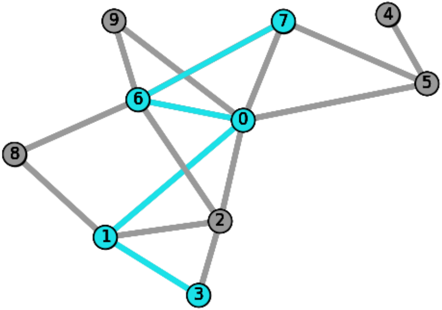
# Example : Sampling the Random Walk $\xi_{n+1}$



# Example : Sampling the Random Walk $\xi_{n+1}$



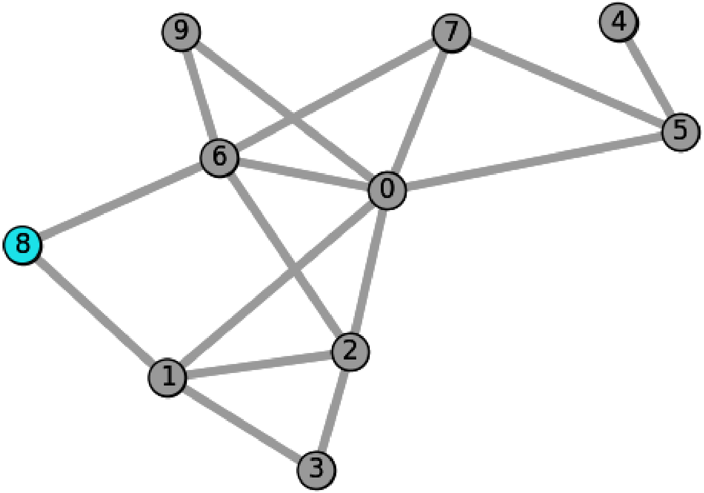
# Example : Stochastic Proximal Gradient step



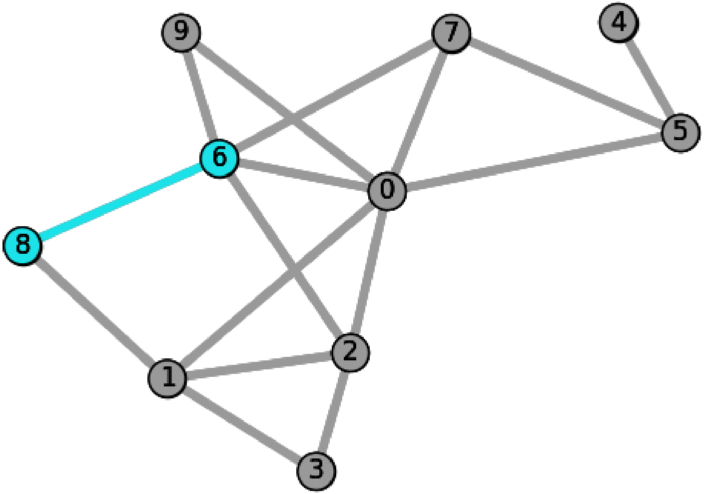

$$\mathrm{TV}(x, \xi_{n+1}) = |x(3) - x(1)| + |x(1) - x(0)| + |x(0) - x(6)| + |x(6) - x(7)|$$

$$x_{n+1} = \mathrm{prox}_{\gamma_n |E| \mathrm{TV}(\cdot, \xi_{n+1})}(x_n - \gamma_n L \nabla F(x_n))$$

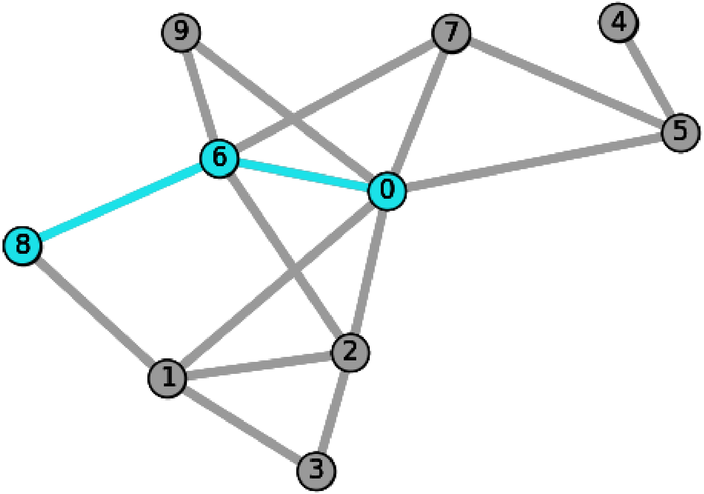
# Example : Sampling the Random Walk $\xi_{n+2}$



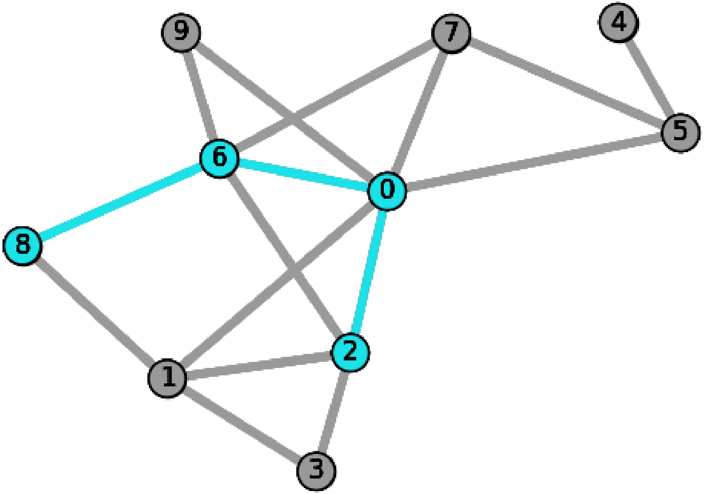
# Example : Sampling the Random Walk $\xi_{n+2}$



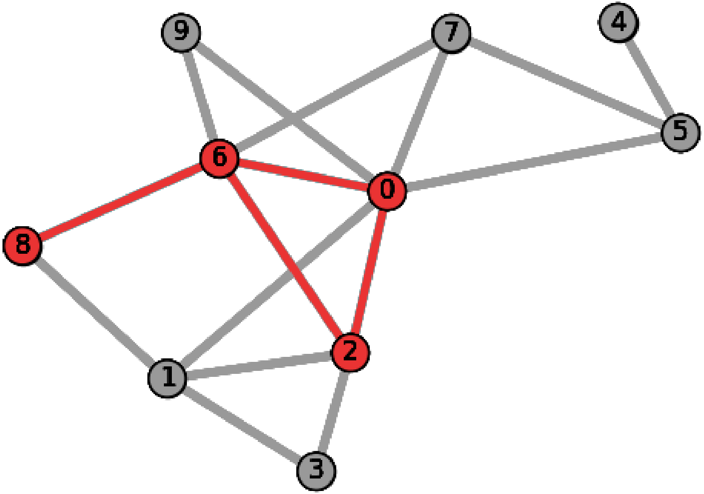
# Example : Sampling the Random Walk $\xi_{n+2}$



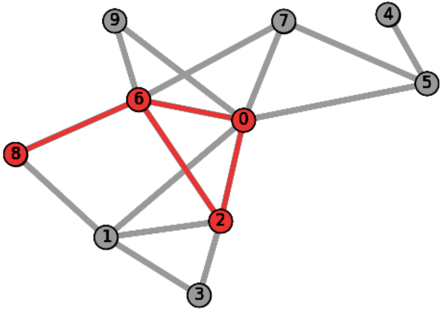
# Example : Sampling the Random Walk $\xi_{n+2}$



# Example : Loop



# Example : Stochastic Proximal Gradient step




$$\mathrm{TV}(x, \xi_{n+2}) = |x(8) - x(6)| + |x(6) - x(0)| + |x(0) - x(2)| + |x(2) - x(6)|$$

$$x_{n+2} = \mathrm{prox}_{\gamma_{n+1}|E|\mathrm{TV}(\cdot,\xi_{n+2})}(x_{n+1} - \gamma_{n+1} L \nabla F(x_{n+1}))$$

**Problem : $\xi_{n+2}$ is not a path graph**

# Snake algorithm

Let $\xi$ is a stationary simple random walk over $G$ with length $L+1$

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

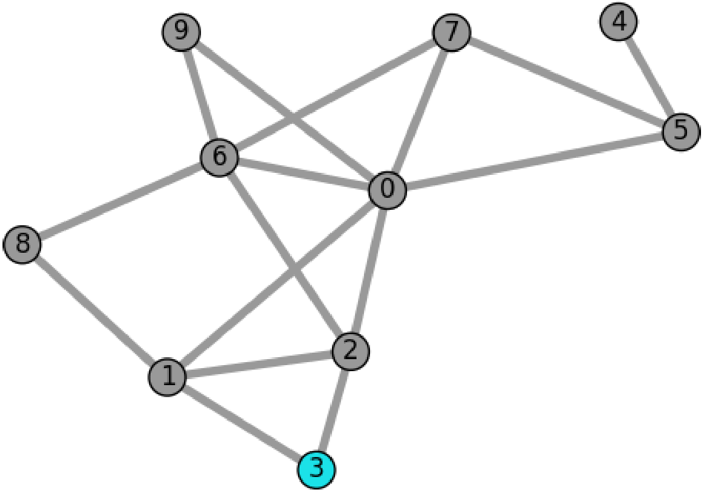
Our problem is equivalent to

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

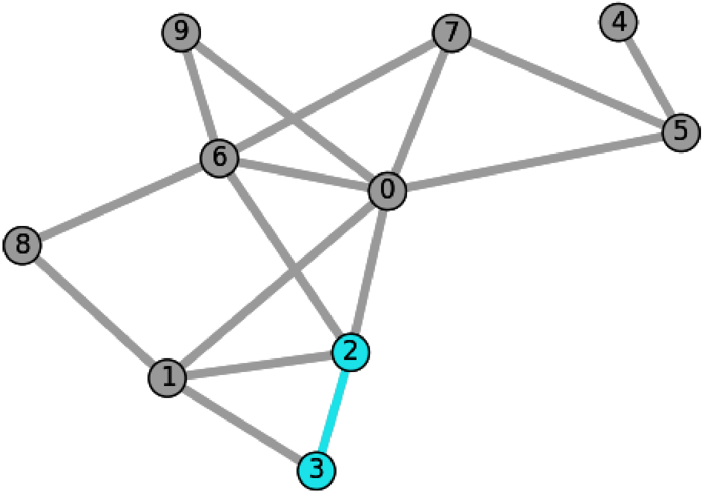
**Snake algorithm**:

$$\begin{cases} \text{Sample the Stationary Random Walk } \xi_{n+1} \textbf{ until Loop} \\ x_{n+1} = \mathrm{prox}_{\gamma_n|E|\mathrm{TV}(\cdot,\xi_{n+1})}(x_n - \gamma_n L(\xi_{n+1})\nabla F(x_n)) \end{cases}$$

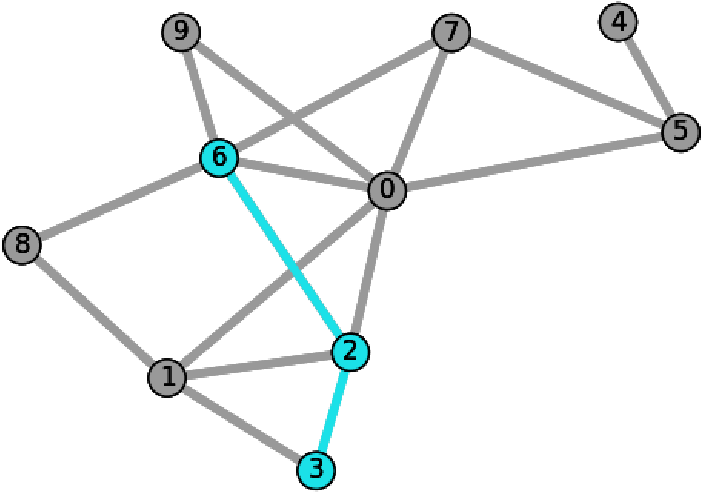
# Example : Snake



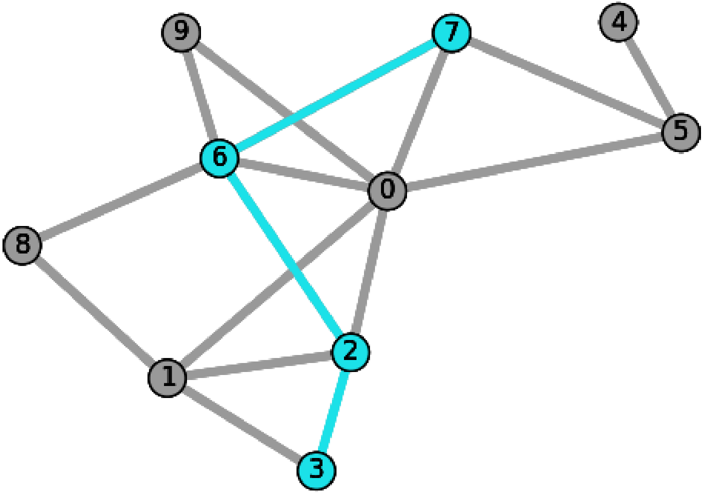
# Example : Snake



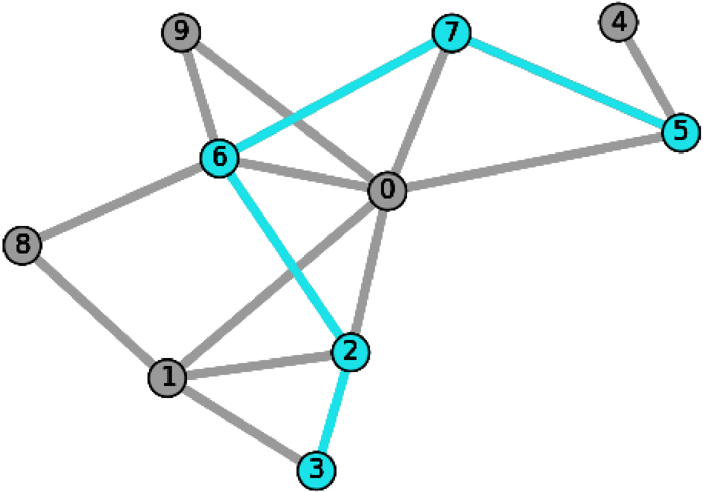
# Example : Snake



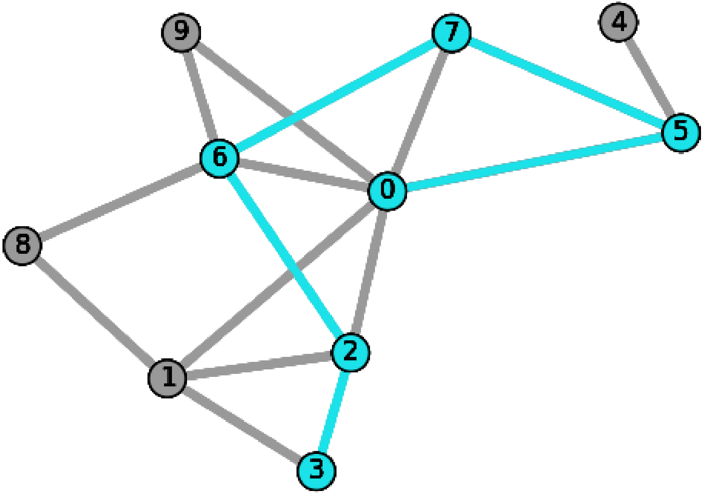
# Example : Snake



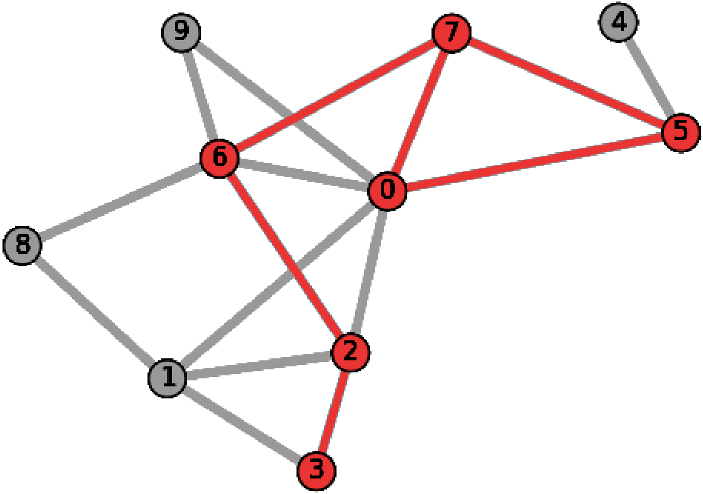
# Example : Snake



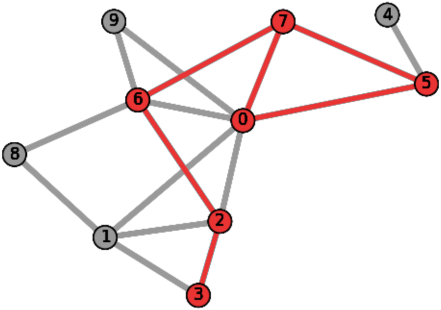
# Example : Snake



# Example : Snake



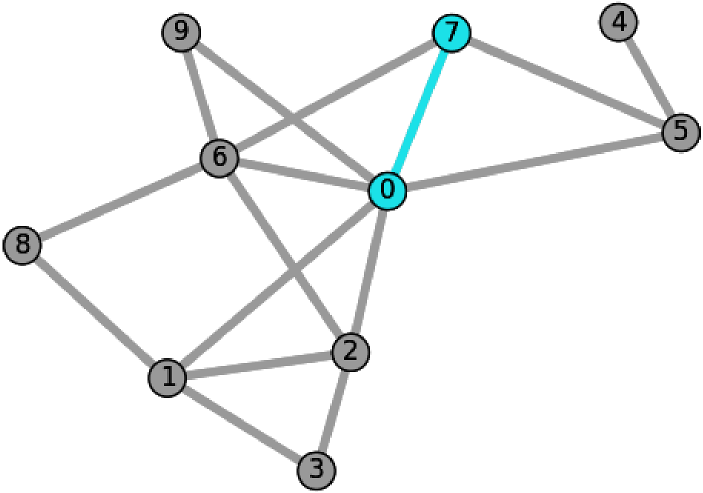
# Example : Snake



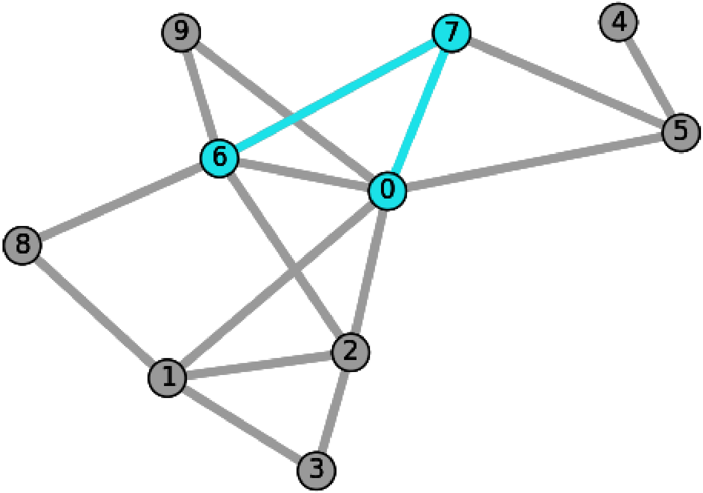

$$\mathrm{TV}(x, \xi_{n+1}) = |x(3) - x(2)| + |x(2) - x(6)| \\ + |x(6) - x(7)| + |x(7) - x(5)| + |x(5) - x(0)|$$

$$x_{n+1} = \mathrm{prox}_{\gamma_n |E| \mathrm{TV}(\cdot, \xi_{n+1})}(x_n - \gamma_n L(\xi_{n+1}) \nabla F(x_n))$$

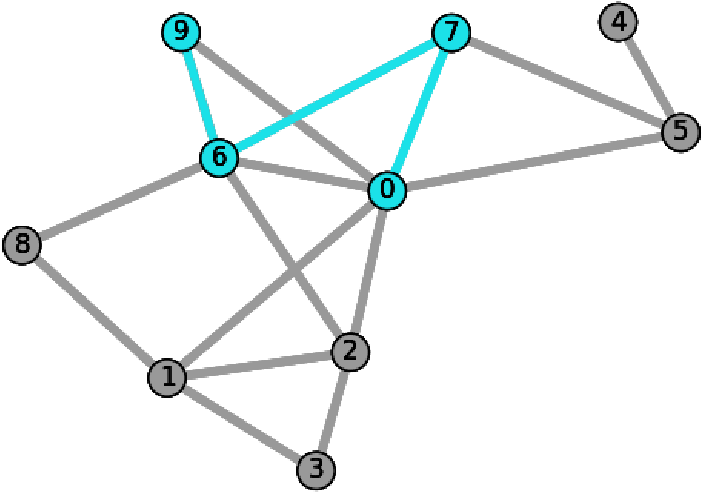
# Example : Snake



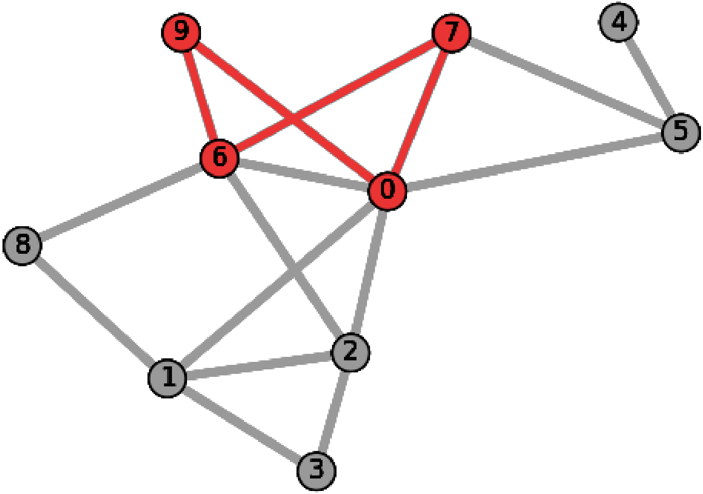
# Example : Snake



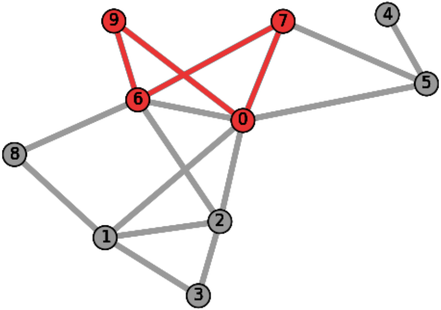
# Example : Snake



# Example : Snake



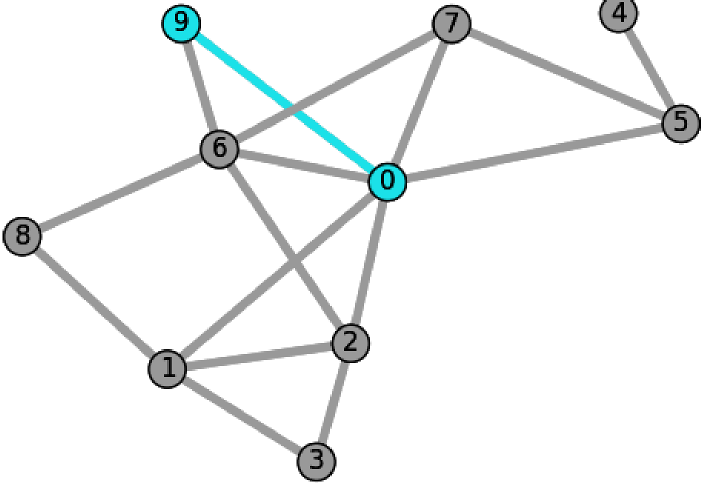
# Example : Snake



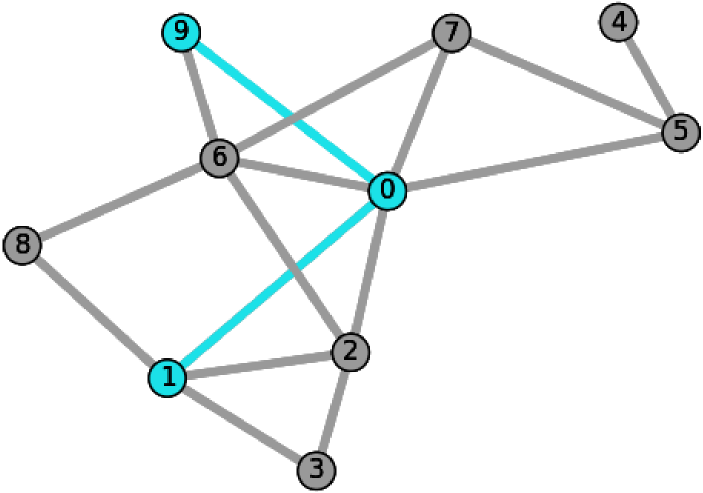

$$\mathrm{TV}(x, \xi_{n+2}) = |x(0) - x(7)| + |x(7) - x(6)| + |x(6) - x(9)|$$

$$x_{n+2} = \mathrm{prox}_{\gamma_{n+1}|E|\mathrm{TV}(\cdot,\xi_{n+2})}(x_{n+1} - \gamma_{n+1} L(\xi_{n+2}) \nabla F(x_{n+1}))$$

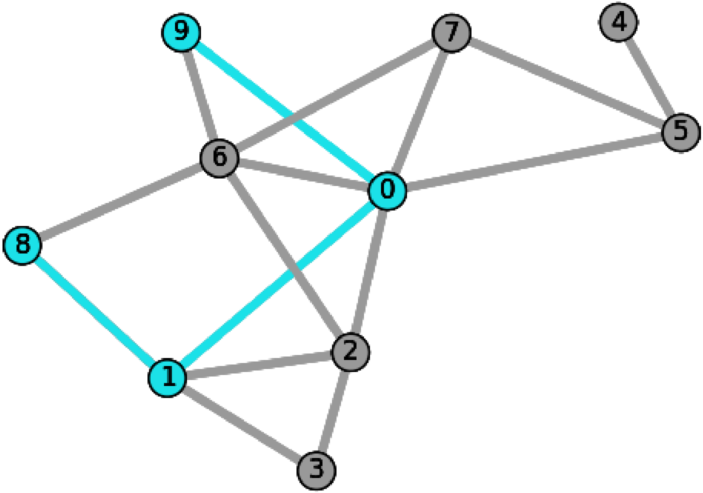
# Example : Snake



# Example : Snake



# Example : Snake



# Convergence of Snake algorithm

Snake is no longer an instance of the stochastic proximal gradient algorithm.

**Theorem** [SBH'17] : If $\gamma_n \downarrow 0$, $x_n \longrightarrow_{n \to +\infty} x_\star$ where $x_\star \in \arg\min_{x \in \mathbb{R}^V} F(x) + \mathrm{TV}(x)$ a.s.

**Proof**:

- $\mathbb{E}_\xi \left(\mathrm{TV}(x, \xi)\right) = \frac{|E|}{L} \mathrm{TV}(x, G)$
- **Convergence of a Generalized Stochastic Proximal Gradient Algorithm**

# Illustration: Online Regularization

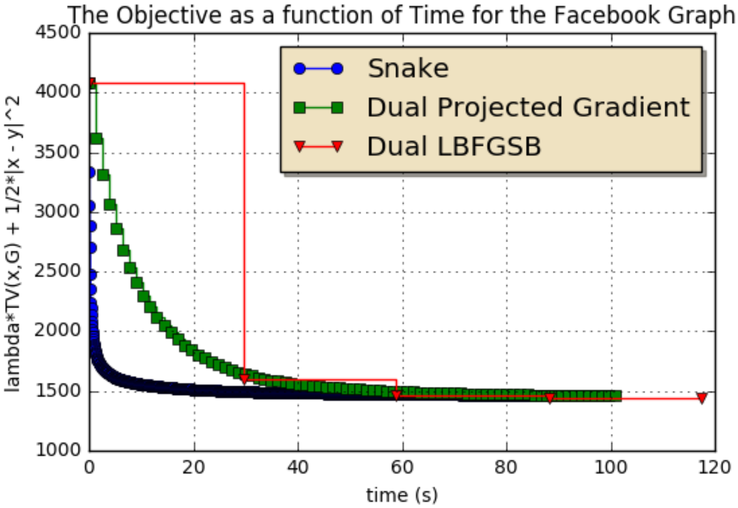



Figure 3: Snake: Trend Filtering over Facebook Graph [Leskovec *et al.*'16]

# Structured Regularizations over Graphs

**Other versions**

$$\min_{x \in \mathbb{R}^V} F(x) + R(x)$$

where

$$R(x) = \sum_{\{i,j\} \in E} \phi_{i,j}(x(i), x(j))$$

with $\phi_{i,j}$ symmetric convex.

**Examples**

- Weighted TV regularization, Laplacian regularization, Weighted/Normalized Laplacian regularization (**DCT**)
- $F(x) = \mathbb{E}_\xi(f(x, \xi))$ or $\sum_{i \in V} f_i(x(i))$

# References


A. Dieuleveut, A. Durmus, and F. Bach.
Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains.
*ArXiv e-prints, 1707.06386*, 2017.

A. Salim, P. Bianchi, and W. Hachem.
Snake: a Stochastic Proximal Gradient Algorithm for Regularized Problems over Large Graphs.
*To appear in Transactions on Automatic Control*, 2017.

P. Bianchi, W. Hachem and A. Salim.
Constant Step Stochastic Approximations Involving Differential Inclusions: Stability, Long-Run Convergence and Applications.
*ArXiv e-prints, arXiv:1612.03831*, 2016.

P. Bianchi, W. Hachem and A. Salim.
A constant step Forward-Backward algorithm involving random maximal monotone operators.
*To appear in Journal of Convex Analysis*, 2017.