Snake: a Stochastic Proximal Gradient Algorithm for Regularized Problems over Large Graphs

Adil Salim
adil-salim.github.io

Telecom ParisTech
GdR ISIS

February 8, 2018

Joint work with Pascal Bianchi and Walid Hachem

Table of Contents

Proximal Gradient algorithm

General Problem:

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

with F, R convex over $\mathcal{X}$, Euclidean space.

If F smooth and R non smooth, Proximal Gradient algorithm:

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

where $\gamma > 0$ and the **proximity operator**

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

Proximal Stochastic Gradient algorithm

In ML, ∇F is often intractable.
Proximal Stochastic Gradient algorithm [Atchadé *et al.*'16] :

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

with

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

Theorem [Atchadé *et al.*'16] : If $\gamma_n \downarrow 0$, then $x_n \longrightarrow_{n\to+\infty} x_\star$ where $x_\star \in \arg\min_{\mathcal{X}} F + R$ a.s.

Constant step - Nonconvex analogous

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

Theorem [BHS'16] : If $\gamma_n \equiv \gamma$ is constant and $f(\cdot, \xi)$ 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 ∇F are intractable?
Stochastic Proximal Gradient algorithm [BH'16] :

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

with

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

Theorem [BH'16] : If $\gamma_n \downarrow 0$, $x_n \longrightarrow_{n\to+\infty} x_\star$ where $x_\star \in \arg\min_{\mathcal{X}} F + R$ a.s.

Constant step analogous

Theorem [BHS'17] : If $\gamma_n \equiv \gamma$ is constant, then

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

Table of Contents

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]

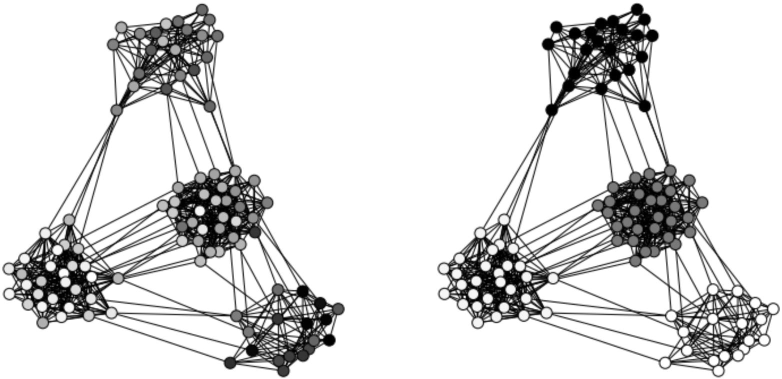


Figure 1: $\min_{x \in \mathbb{R}^V} \frac{1}{2}\|x - y\|^2 + \mathrm{TV}(x, G)$

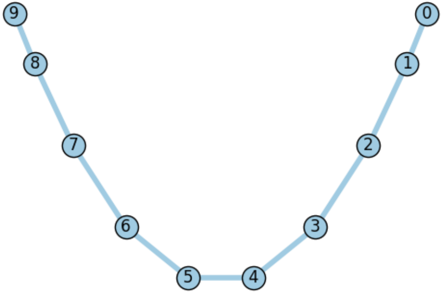
Problem Statement

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

Table of Contents

Sampling Random Walks

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

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

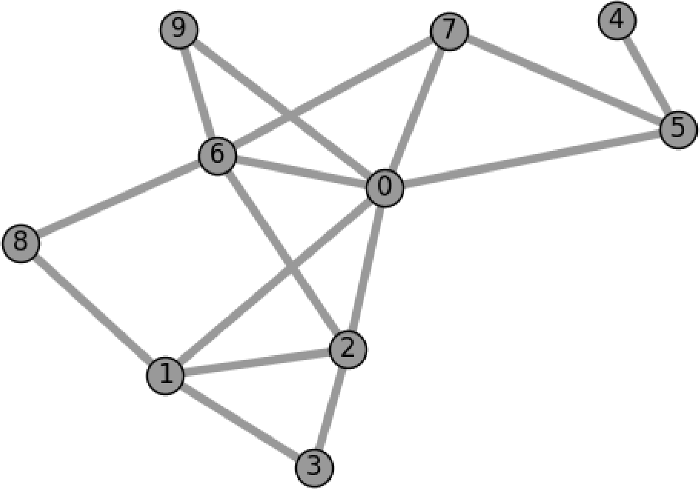
Our problem is equivalent to

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

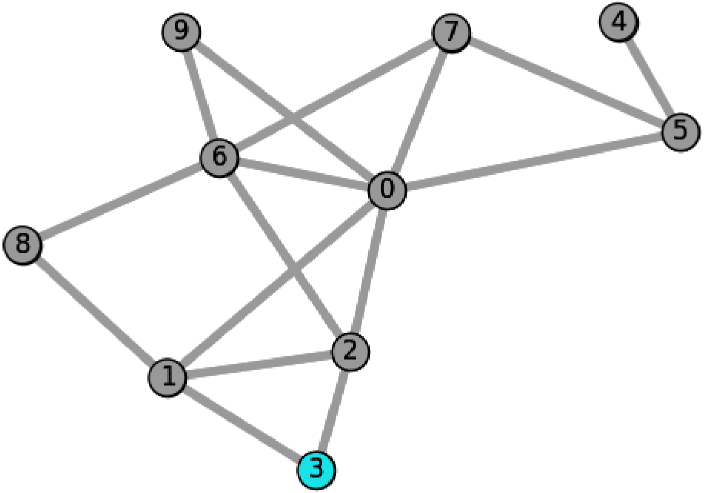
Stochastic Proximal Gradient algorithm:

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

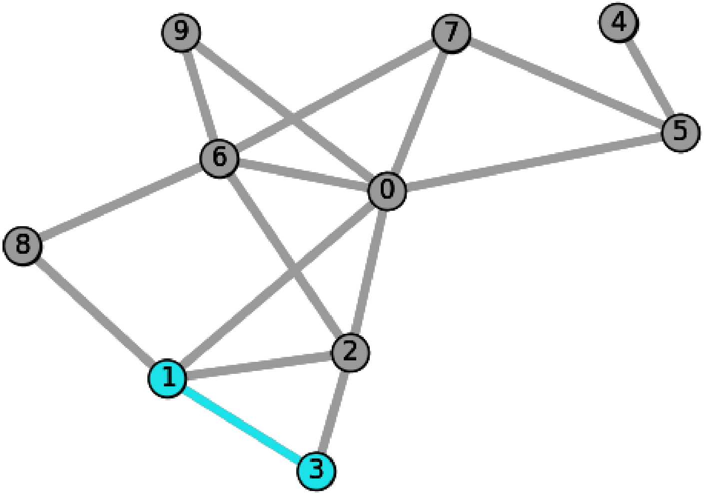
Example : The Graph G



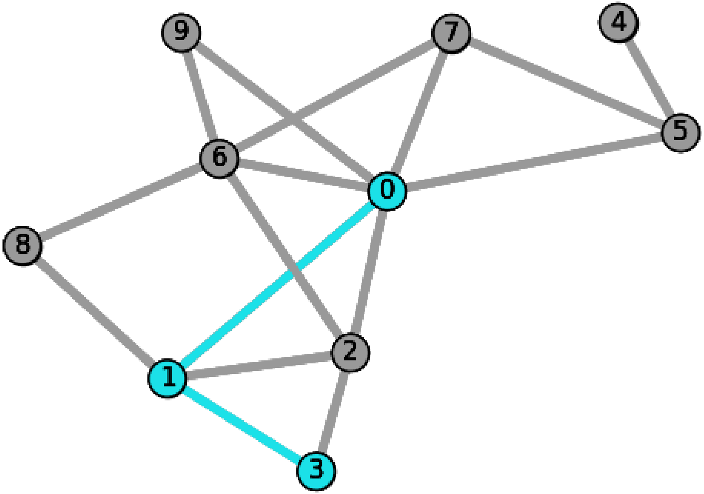
Example : Sampling the Random Walk ξ_{n+1}



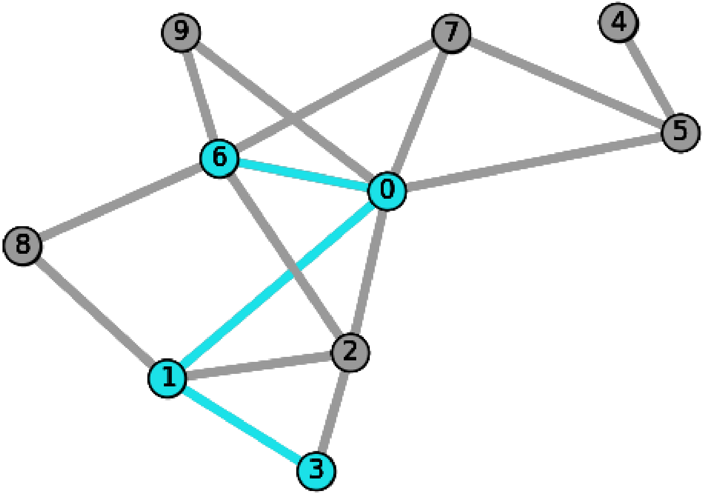
Example : Sampling the Random Walk ξ_{n+1}



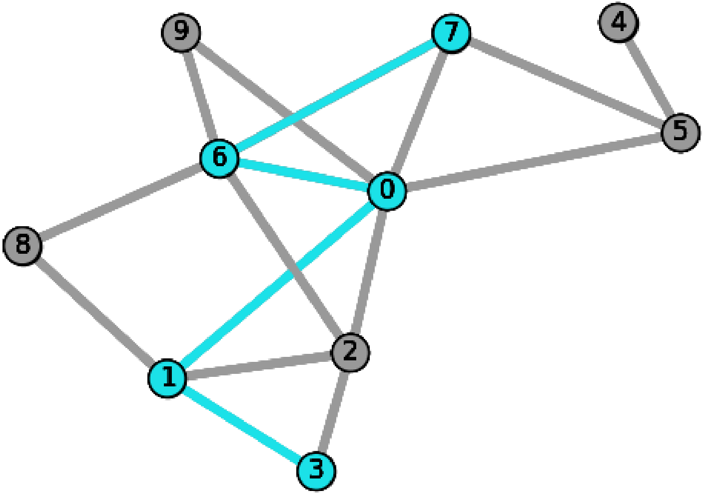
Example : Sampling the Random Walk ξ_{n+1}



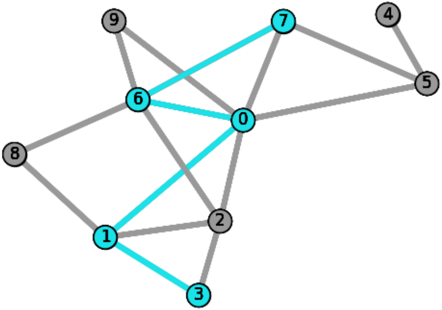
Example : Sampling the Random Walk ξ_{n+1}



Example : Sampling the Random Walk ξ_{n+1}



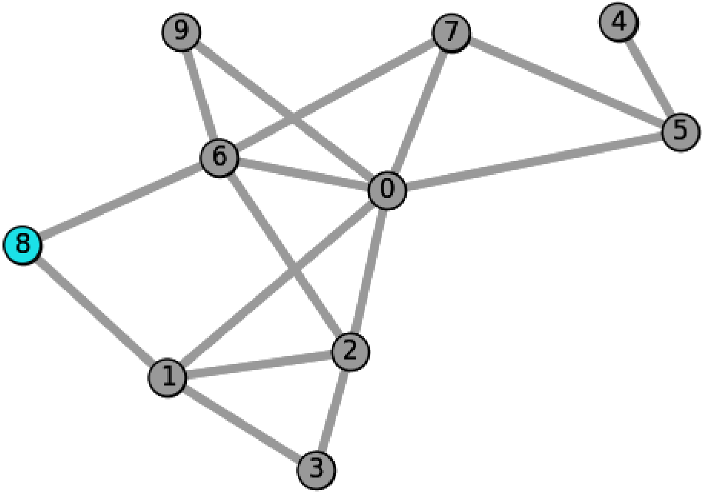
Example : Stochastic Proximal Gradient step



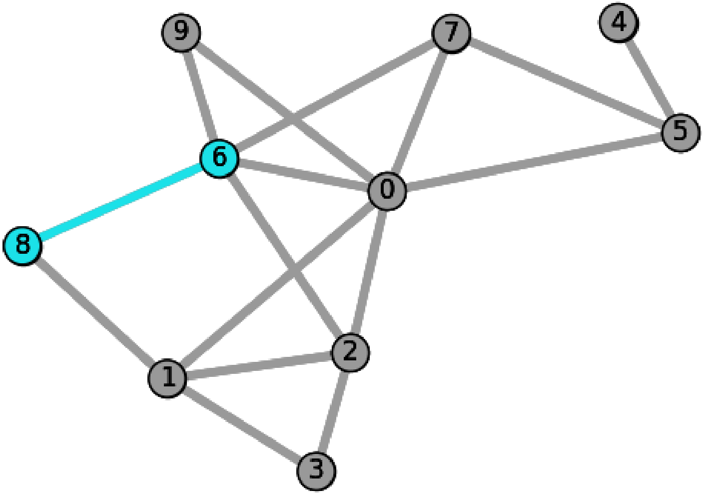
$$\mathrm{TV}(x, \xi_{n+1}) = |x(3) - x(1)| + |x(1) - x(0)| + |x(0) - x(6)| + |x(6) - x(7)|$$

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

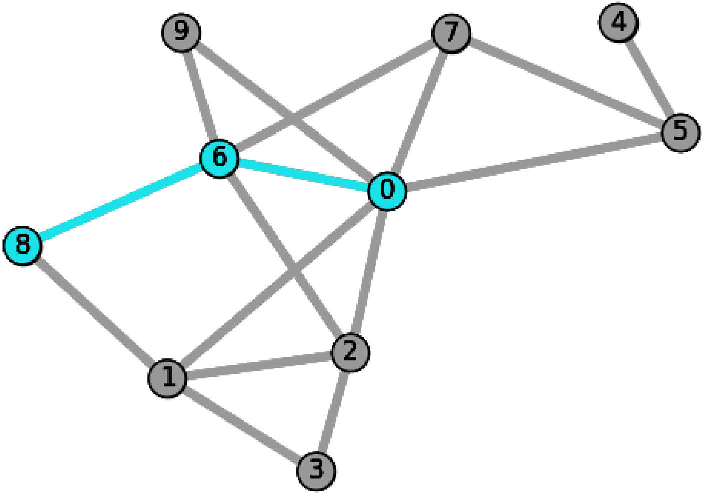
Example : Sampling the Random Walk ξ_{n+2}



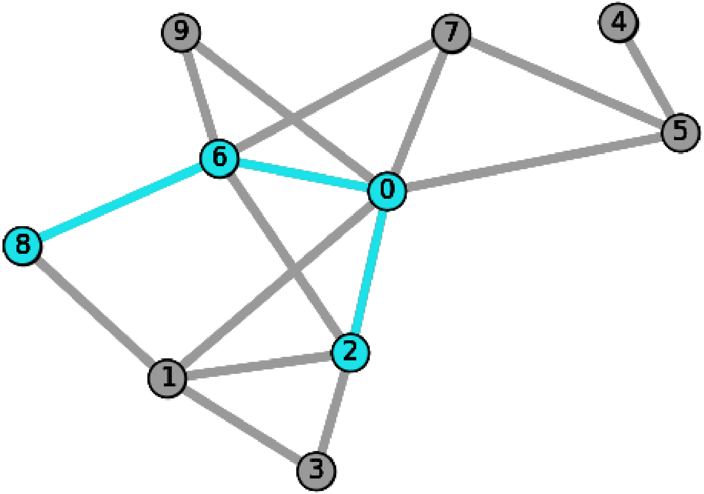
Example : Sampling the Random Walk ξ_{n+2}



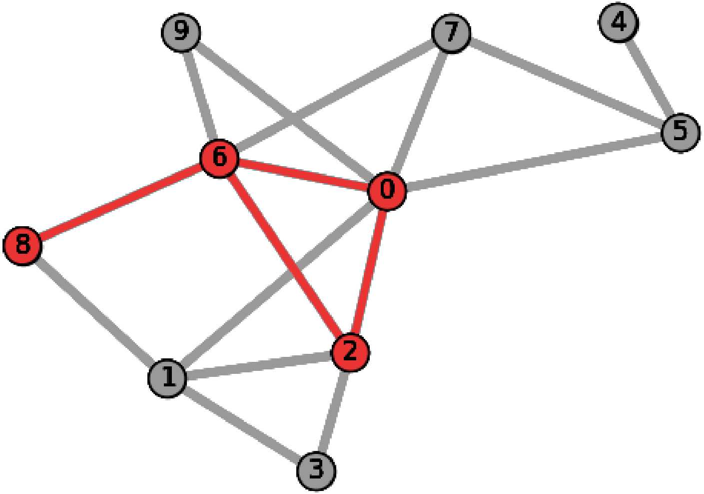
Example : Sampling the Random Walk ξ_{n+2}



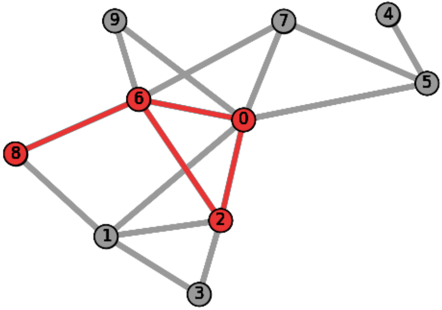
Example : Sampling the Random Walk ξ_{n+2}



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 : ξ_{n+2} is not a path graph

Table of Contents

Snake algorithm

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

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

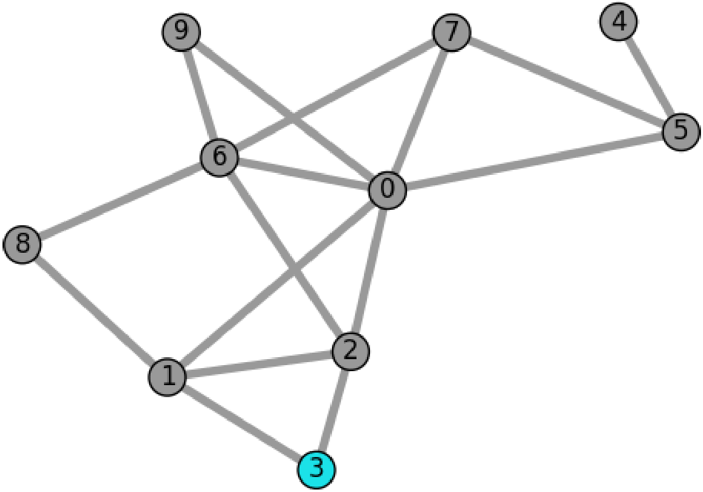
Our problem is equivalent to

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

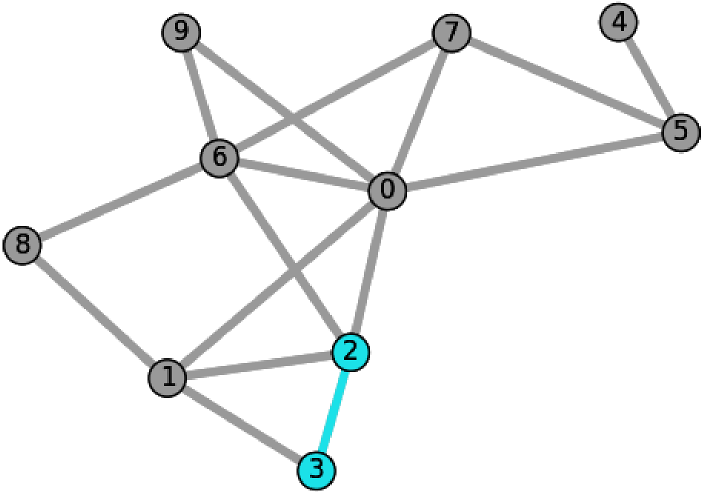
Snake algorithm:

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

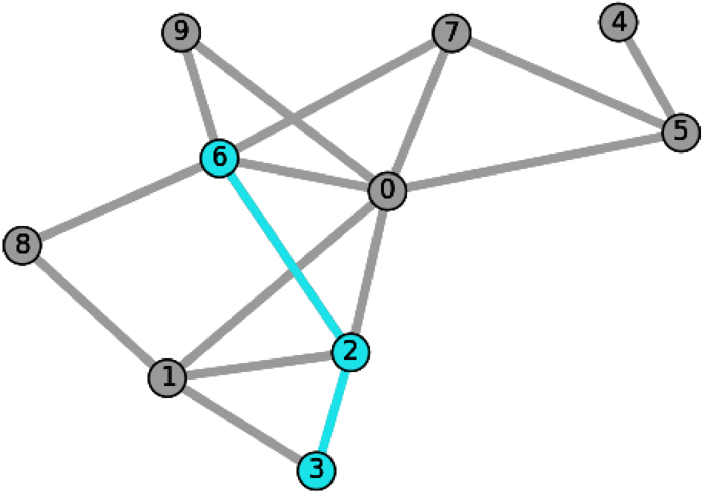
Example : Snake



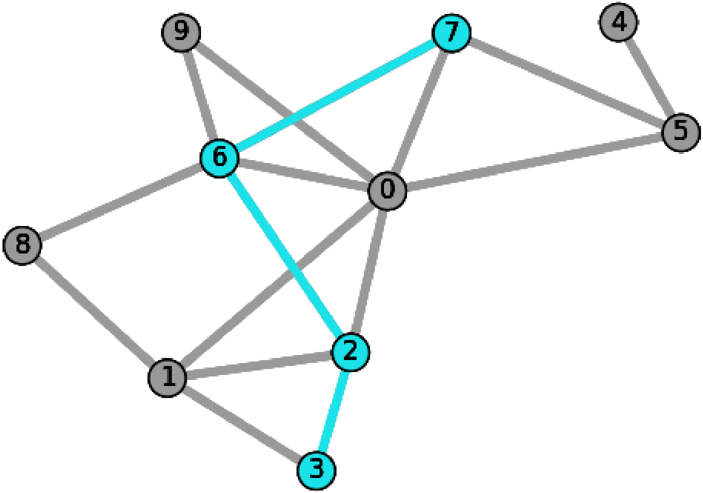
Example : Snake



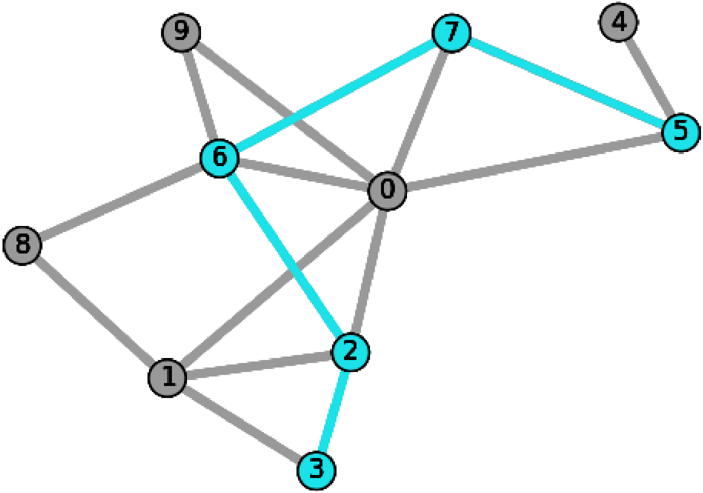
Example : Snake



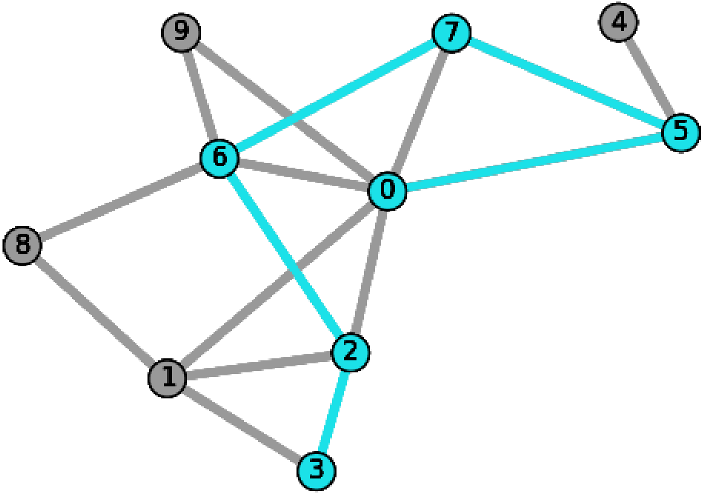
Example : Snake



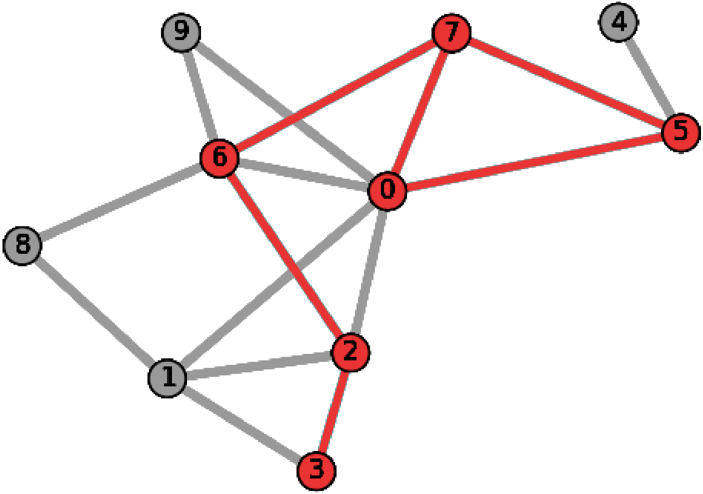
Example : Snake



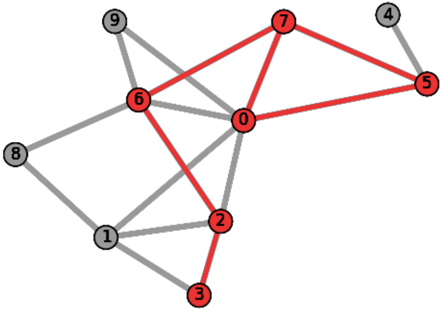
Example : Snake



Example : Snake



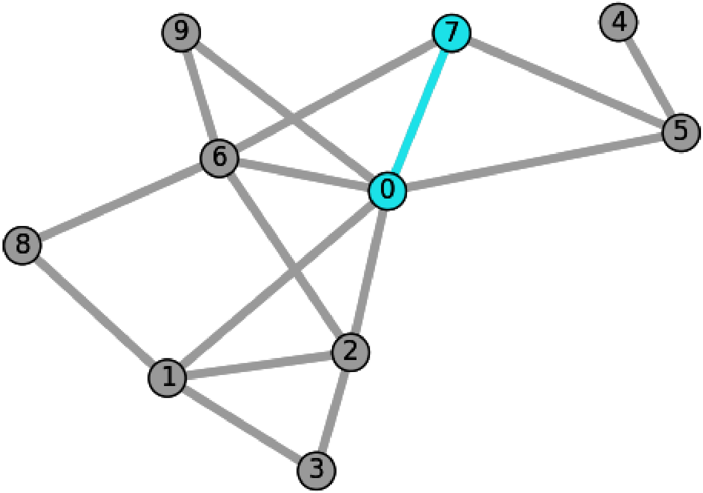
Example : Snake



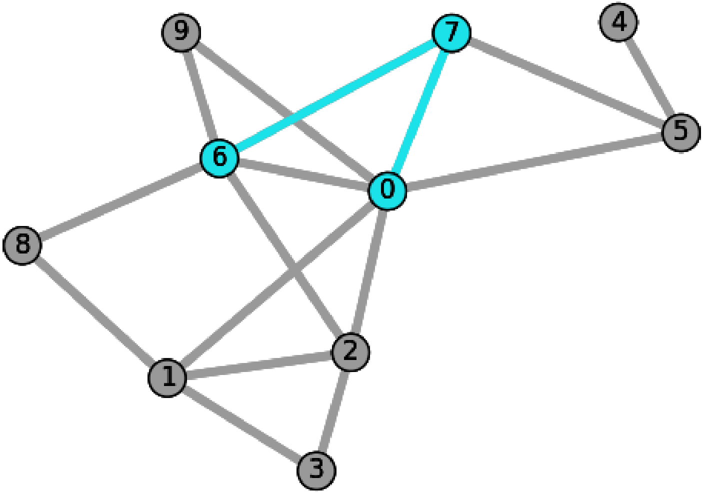
$$\mathrm{TV}(x, \xi_{n+1}) = |x(3) - x(2)| + |x(2) - x(6)| \\ + |x(6) - x(7)| + |x(7) - x(5)| + |x(5) - x(0)|$$

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

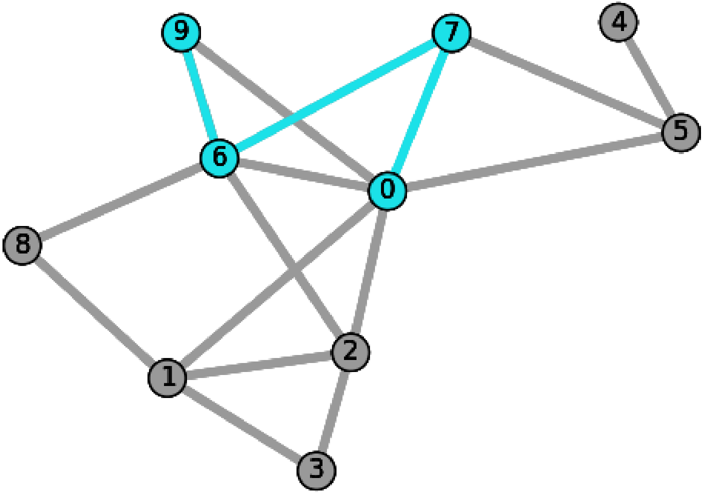
Example : Snake



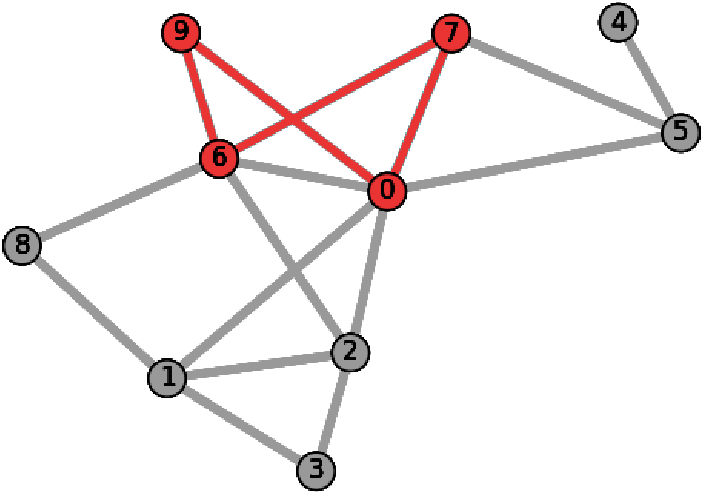
Example : Snake



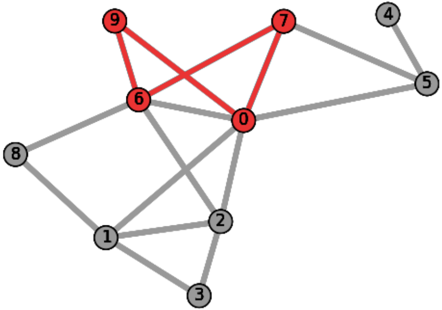
Example : Snake



Example : Snake



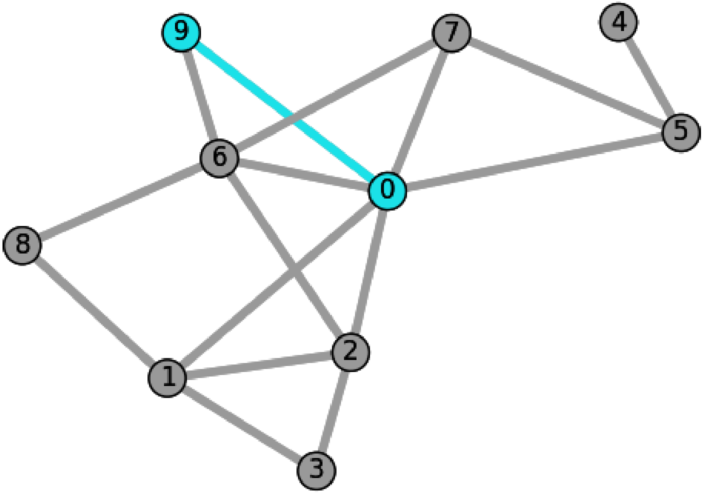
Example : Snake



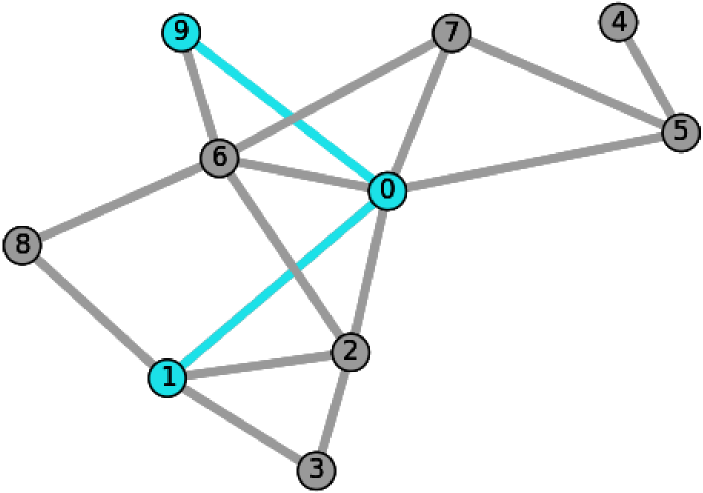
$$\mathrm{TV}(x, \xi_{n+2}) = |x(0) - x(7)| + |x(7) - x(6)| + |x(6) - x(9)|$$

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

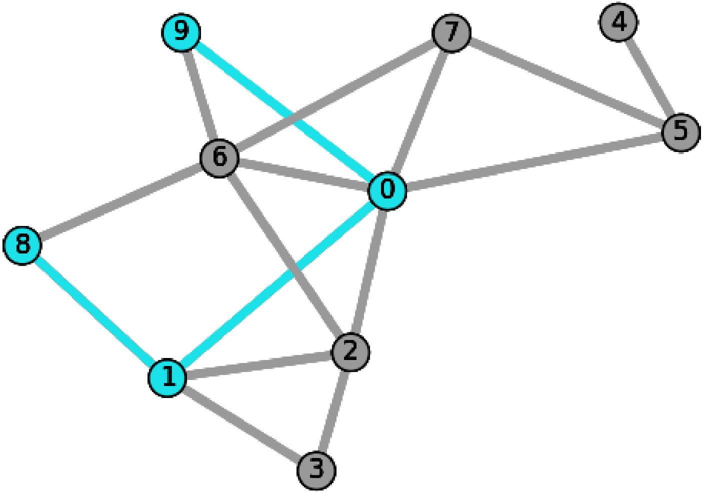
Example : Snake



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

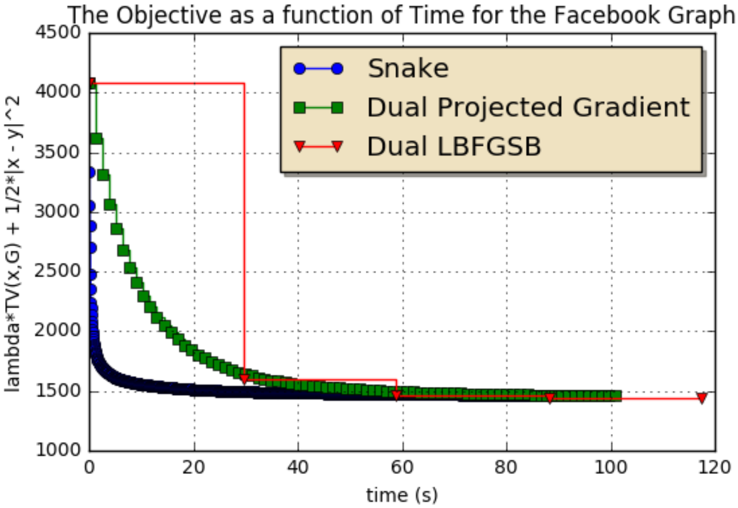


Figure 2: Snake: Trend Filtering over Facebook Graph [Leskovec *et al.*'16]

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

Y.F Atchadé, G. Fort, and E. Moulines.
On stochastic proximal gradient algorithms.
ArXiv e-prints, 1402.2365, 2014.

A. Salim, P. Bianchi, and W. Hachem.
Snake: a Stochastic Proximal Gradient Algorithm for Regularized Problems over Large Graphs.
2017.

P. Bianchi and W. Hachem.
Dynamical behavior of a stochastic forward-backward algorithm using random monotone operators.
J. Optim. Theory Appl., 171(1):90–120, 2016.

P. Bianchi, W. Hachem and A. Salim.
A constant step Forward-Backward algorithm involving random maximal monotone operators.
ArXiv e-prints, 1702.04144, 2017.