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:

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

with F, R convex over X, Euclidean space.

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

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

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

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

Proximal Stochastic Gradient algorithm

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

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

with

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

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

Constant step - Nonconvex analogous

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

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,

$$\limsup_{n \to +\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{P}(d(x_k; Z) > \varepsilon) \quad \xrightarrow[\gamma \to 0]{} 0.$$

Stochastic Proximal Gradient algorithm

What if both prox_{R} and ∇F are intractable?
Stochastic Proximal Gradient algorithm [BH'16] :

$$x_{n+1} = \mathrm{prox}_{\ _n r(\ ;\ _{n+1})}(x_n \quad _n \nabla_x f(x_n;\ _{n+1}))$$

with

- $(\ _n)$ iid
- $\mathrm{E}\ (f(x;\ )) = F(x)$
- $\mathrm{E}\ (r(x;\ )) = R(x).$

Theorem [BH'16] : If $\ _n \downarrow 0$, $x_n \quad \to_{n \to +\infty} x_\star$ where $x_\star \in \arg\min_X F + R$ a.s.

Constant step analogous

Theorem [BHS'17] : If ${}_{n}$ is constant, then

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

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]

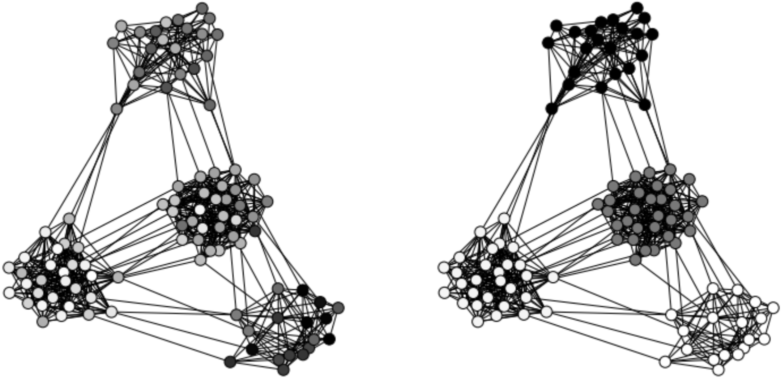


Figure 1: $\min_{x\in\mathbb{R}^V} \frac{1}{2}\|x - y\|^2 + \mathrm{TV}(x;G)$

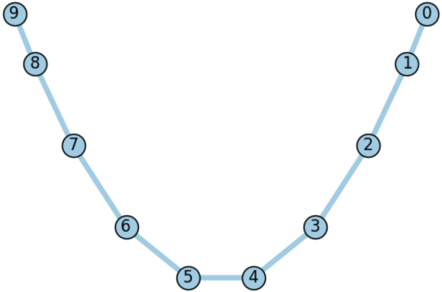
Problem Statement

Proximal Gradient algorithm

$$x_{n+1} = \mathrm{prox}_{\quad \mathrm{TV}(:;G)}(x_n \qquad r\, F(x_n))$$

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

$$\mathrm{E} \ \ (\mathrm{TV}(x; \ \ )) = \frac{jEj}{L}\mathrm{TV}(x; G):$$

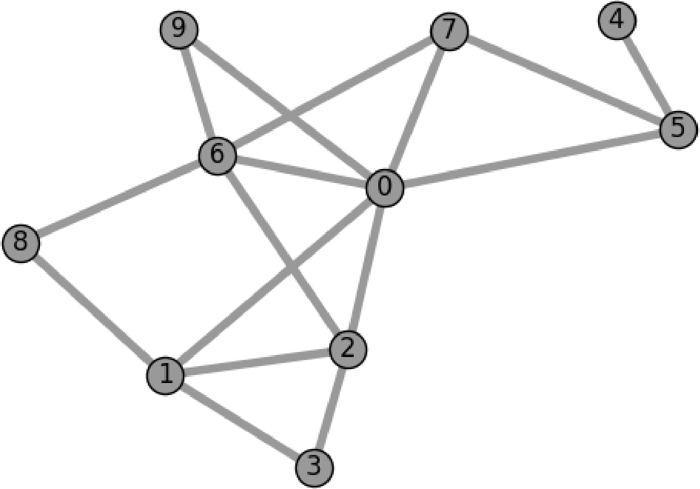
Our problem is equivalent to

$$\min_{x 2 \mathrm{R}^V} LF(x) + jEj\mathrm{E} \ \ (\mathrm{TV}(x; \ \ )):$$

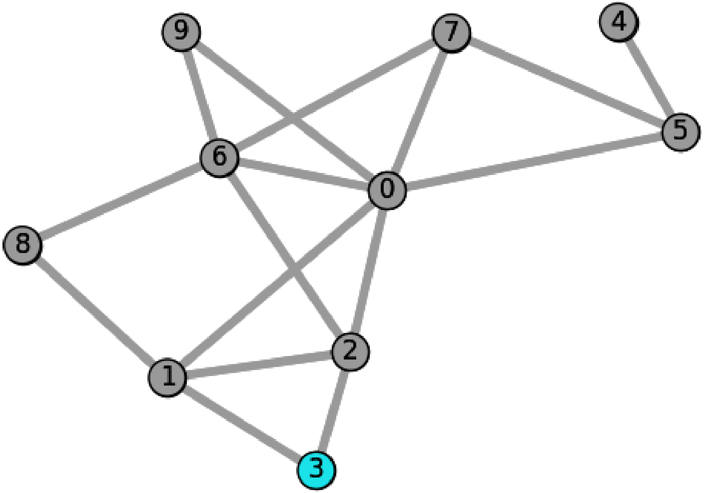
Stochastic Proximal Gradient algorithm:

Sample the Stationary Random Walk $_{n+1}$ with length $L + 1$
$x_{n+1} = \mathrm{prox}_{\ _n jEj \mathrm{TV}(\ ; \ _{n+1})}(x_n \qquad _n L \ r F(x_n))$

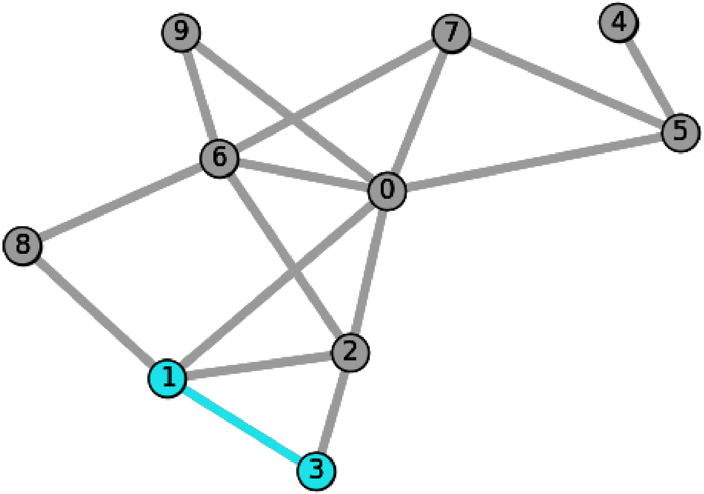
Example : The Graph G



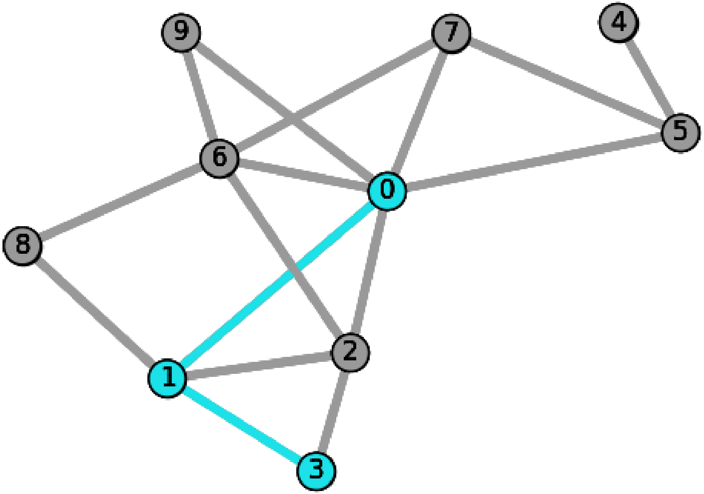
Example : Sampling the Random Walk $_{n+1}$



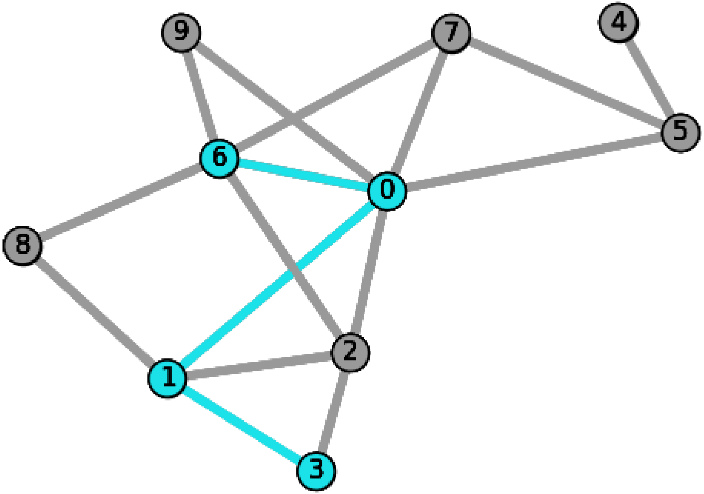
Example : Sampling the Random Walk $_{n+1}$



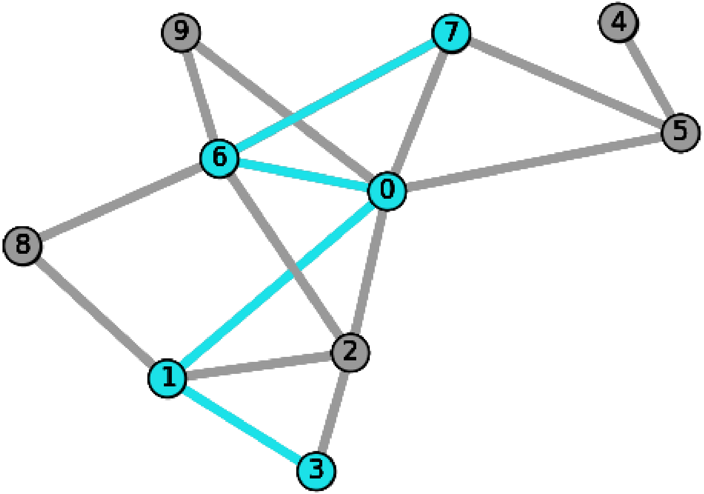
Example : Sampling the Random Walk $_{n+1}$



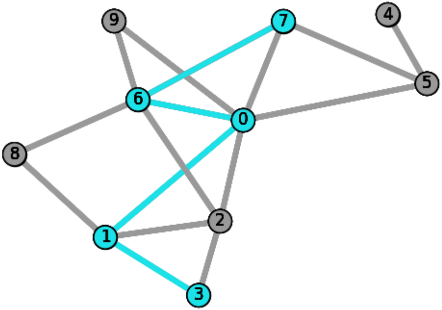
Example : Sampling the Random Walk $_{n+1}$



Example : Sampling the Random Walk $_{n+1}$



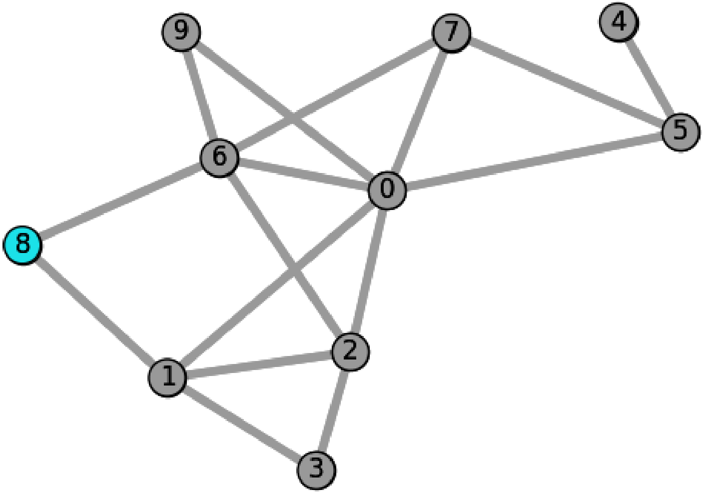
Example : Stochastic Proximal Gradient step



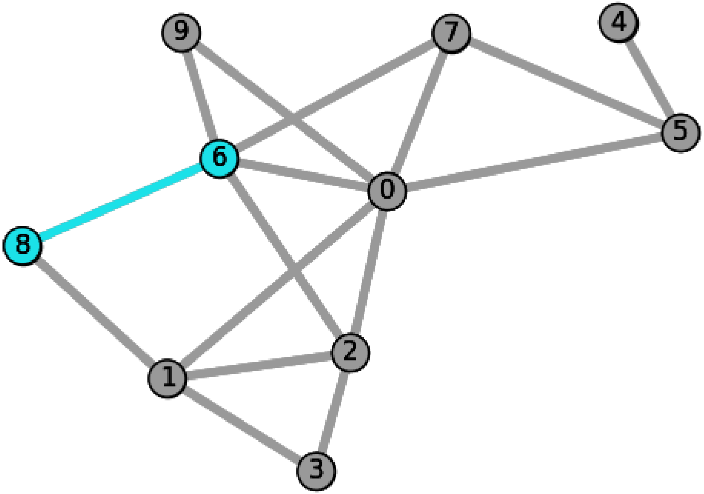
$$\mathrm{TV}(x; \ _{n+1}) = |x(3) - x(1)| + |x(1) - x(0)| + |x(0) - x(6)| + |x(6) - x(7)|$$

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

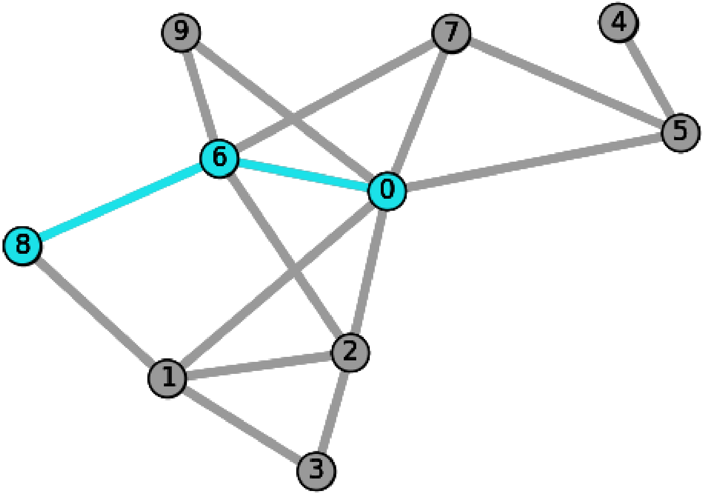
Example : Sampling the Random Walk $_{n+2}$



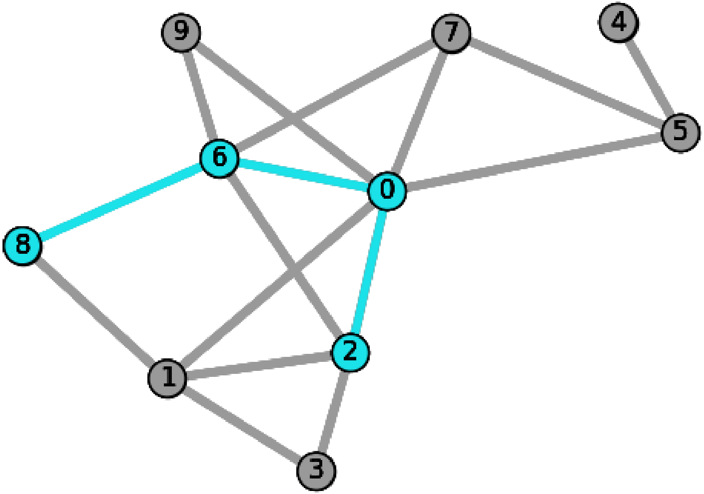
Example : Sampling the Random Walk $_{n+2}$



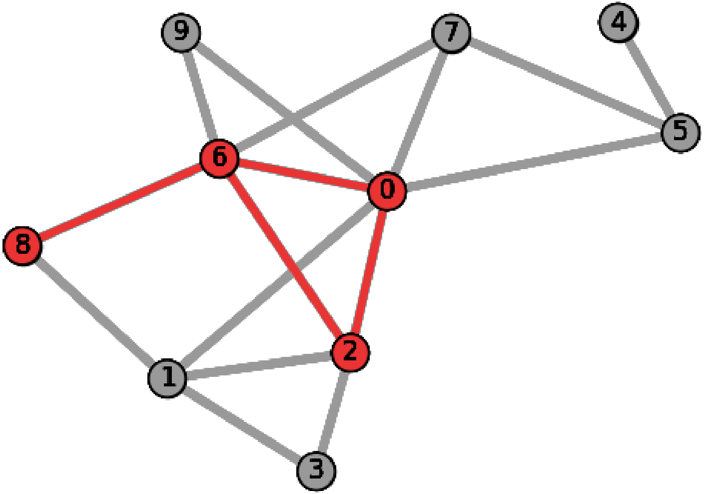
Example : Sampling the Random Walk $_{n+2}$



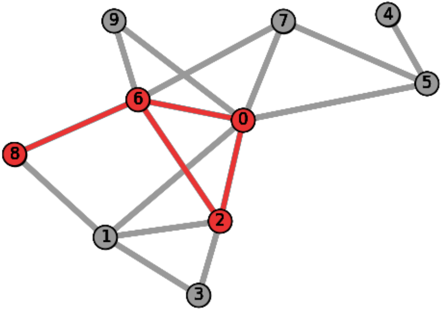
Example : Sampling the Random Walk $_{n+2}$



Example : Loop



Example : Stochastic Proximal Gradient step



$$\mathrm{TV}(x;\ _{n+2}) = jx(8) \quad x(6)j + jx(6) \quad x(0)j + jx(0) \quad x(2)j + jx(2) \quad x(6)j$$

$$x_{n+2} = \mathrm{prox}\ _{_{n+1}jEj\mathrm{TV}(\ ;\ _{n+2})}(x_{n+1} \qquad _{n+1}L\ r\ F(x_{n+1}))$$

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

Table of Contents

Snake algorithm

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

$$\mathrm{E}\,(\mathrm{TV}(x;\ )) = \frac{jEj}{L}\mathrm{TV}(x;G):$$

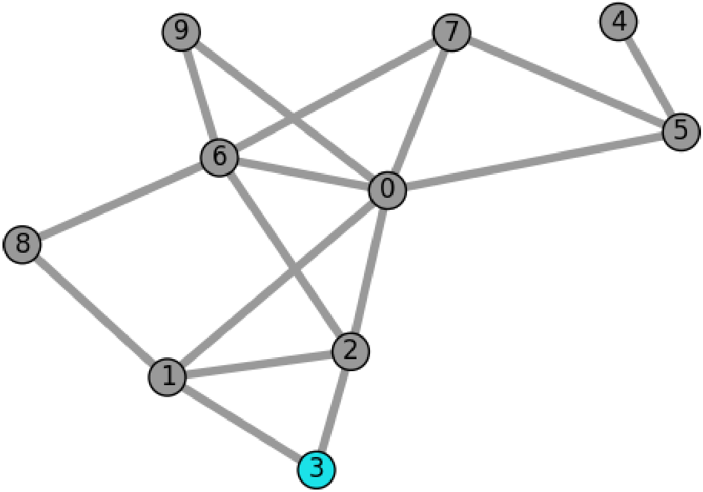
Our problem is equivalent to

$$\min_{x\,2\mathrm{R}^V} LF(x) + jEj\mathrm{E}\ \ (\mathrm{TV}(x;\ )):$$

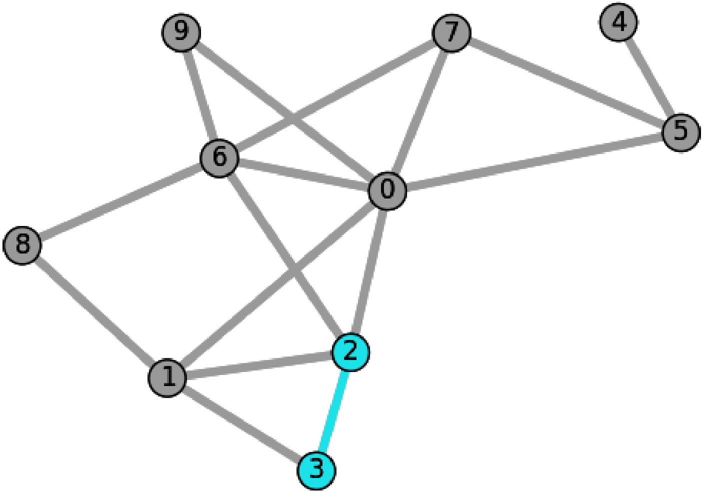
Snake algorithm:

Sample the Stationary Random Walk $_{n+1}$ **until Loop**
$x_{n+1} = \mathrm{prox}_{\ \ njEj\mathrm{TV}(\ ;\ n+1)}(x_n \qquad {}_nL(\ _{n+1})\, r\, F(x_n))$

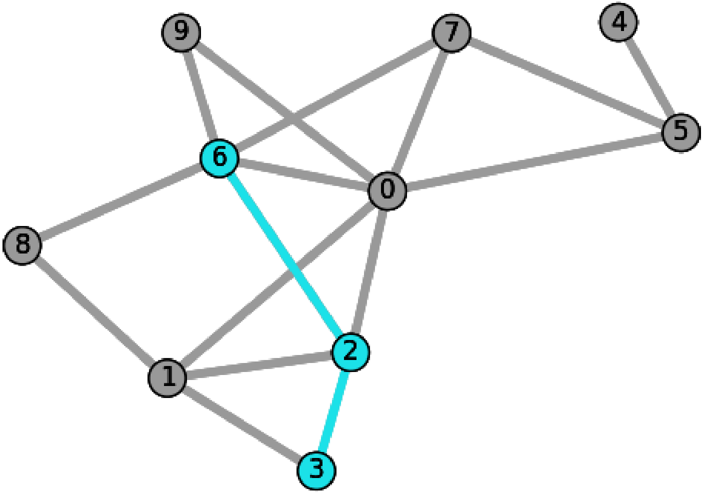
Example : Snake



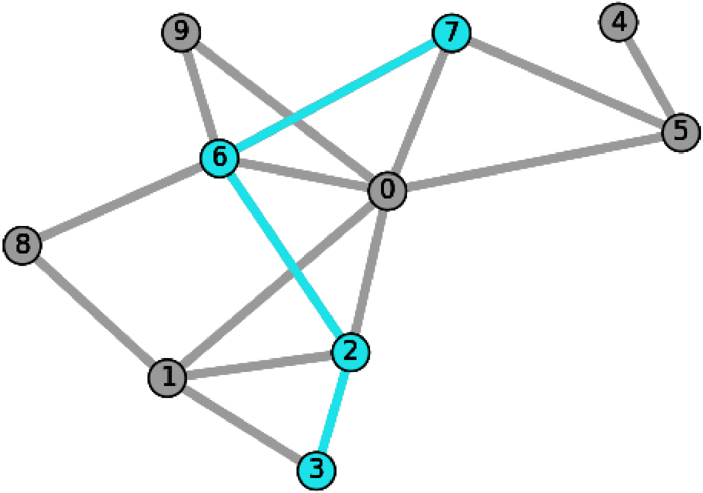
Example : Snake



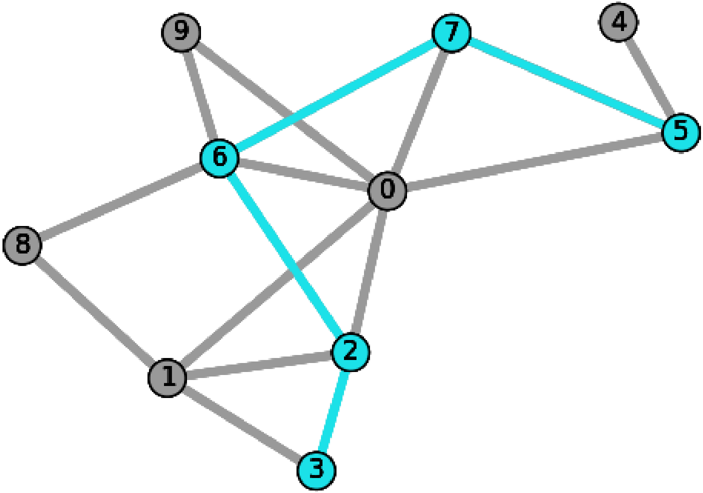
Example : Snake



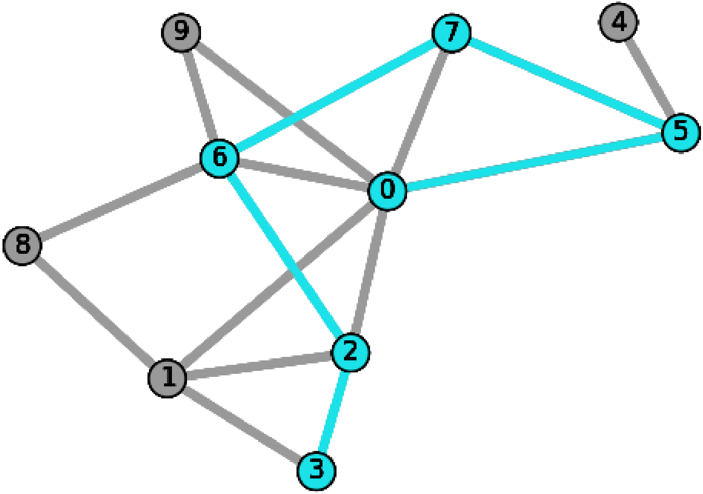
Example : Snake



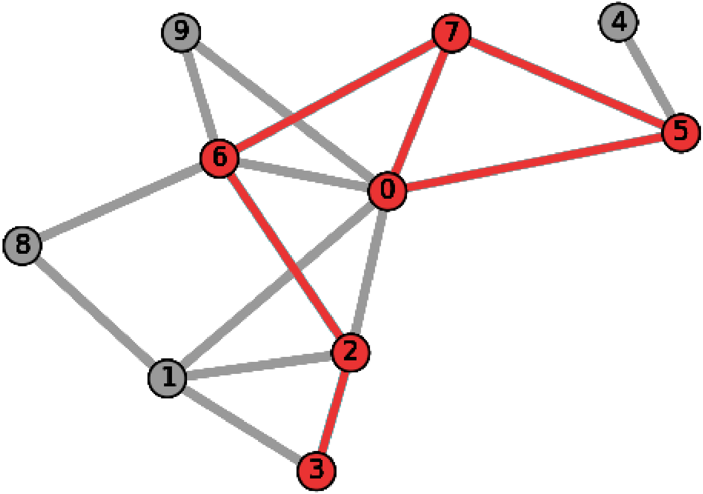
Example : Snake



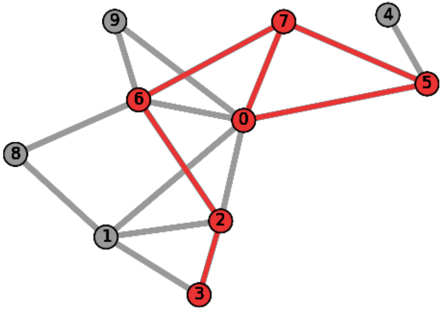
Example : Snake



Example : Snake

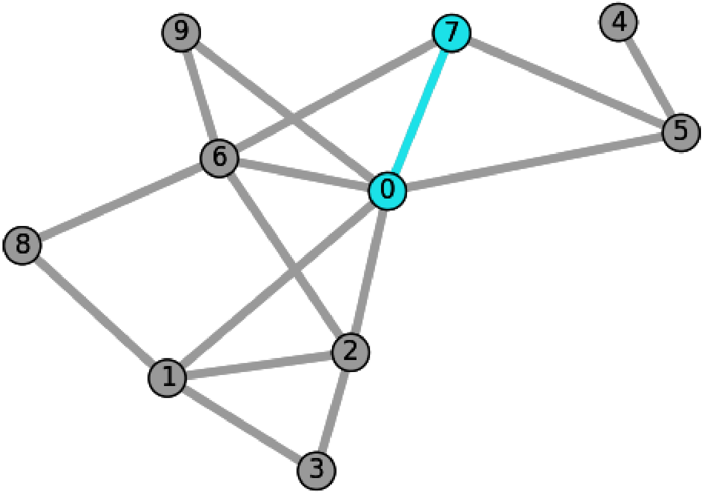


Example : Snake

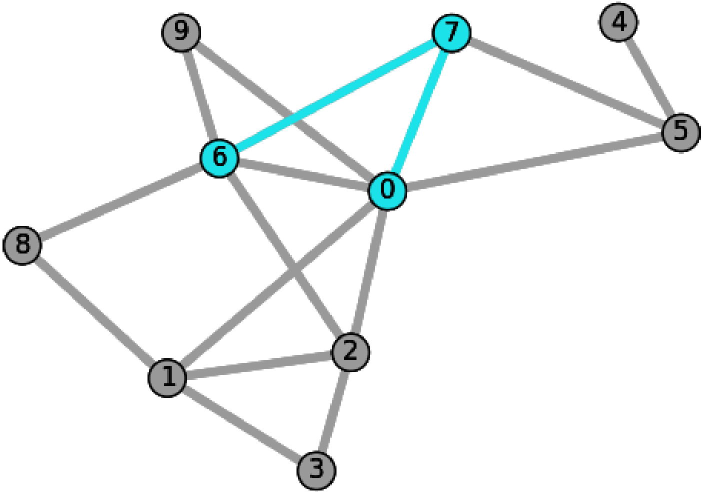


$$\mathrm{TV}(x; \ _{n+1}) = |x(3) - x(2)| + |x(2) - x(6)|$$
$$+ |x(6) - x(7)| + |x(7) - x(5)| + |x(5) - x(0)|$$
$$x_{n+1} = \mathrm{prox}_{\ _n |E| \mathrm{TV}(\ ; \ _{n+1})}(x_n - \ _n L(\ _{n+1}) \nabla F(x_n))$$

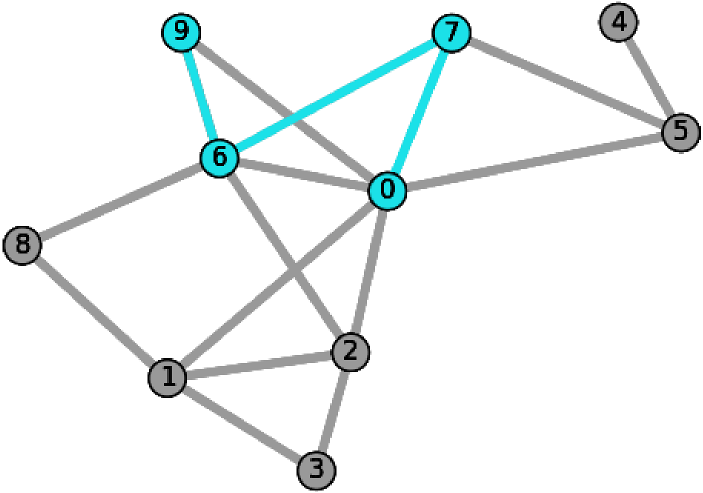
Example : Snake



Example : Snake



Example : Snake



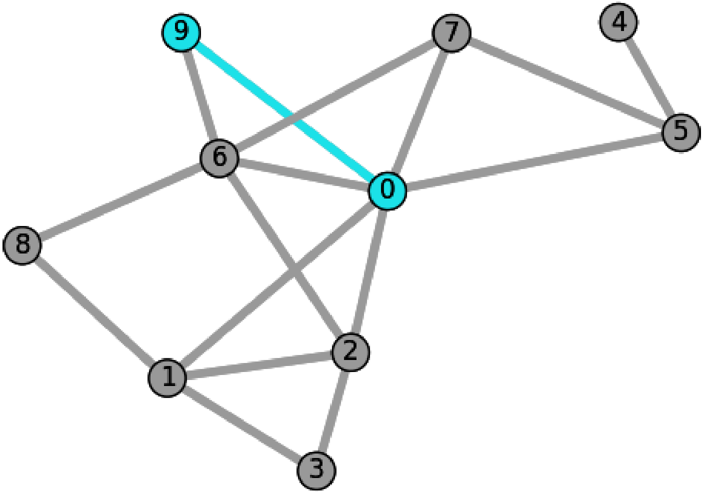
Example : Snake

Example : Snake

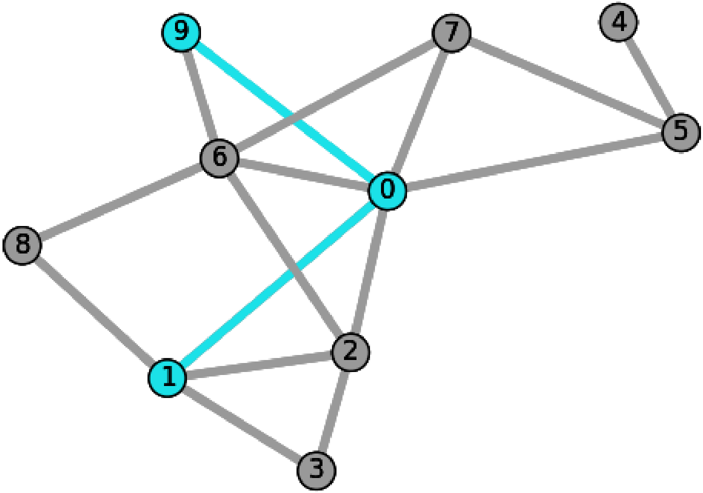
$$\mathrm{TV}(x;\ _{n+2}) = jx(0) \quad x(7)j + jx(7) \quad x(6)j + jx(6) \quad x(9)j$$

$$x_{n+2} = \mathrm{prox}_{\ _{n+1}jEj\mathrm{TV}(\ ;\ _{n+2})}(x_{n+1} \qquad _{n+1}L(\ _{n+2})\, r\, F(x_{n+1}))$$

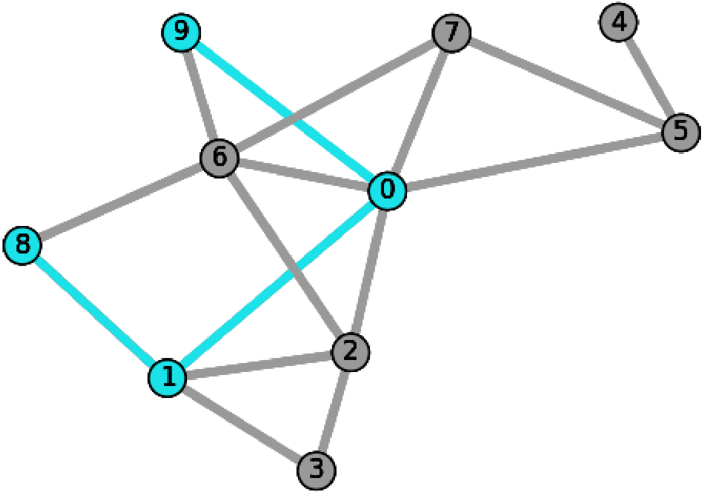
Example : Snake



Example : Snake



Example : Snake



Convergence of Snake algorithm

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

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

Proof:

- $\mathrm{E}\ (\mathrm{TV}(x;\ )) = \frac{|E|}{L} \mathrm{TV}(x;G)$
- **Convergence of a Generalized Stochastic Proximal Gradient Algorithm**

Illustration: Online Regularization

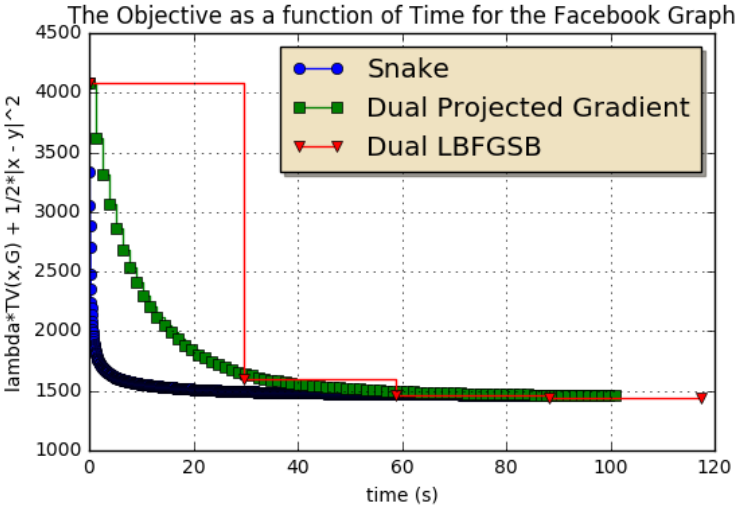


Figure 2: Snake: Trend Filtering over Facebook Graph [Leskovec *et al.*'16]

Structured Regularizations over Graphs

Other versions

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

where

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

with ${}_{i,j}$ symmetric convex.

Examples

- Weighted TV regularization, Laplacian regularization, Weighted/Normalized Laplacian regularization (**DCT**)
- $F(x) = \mathbb{E}\ (f(x;\ ))$ 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.