Stochastic proximal gradient algorithm

Adil Salim
joint work with Pascal Bianchi and Walid Hachem

Telecom ParisTech

09/21/2017

Presentation of the algorithm

Convergence results

Applications

Stochastic Gradient algorithm

General problem in Machine Learning :

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

where

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

where ξ is a random variable and $x \mapsto f(x,\xi)$ is a.s a **convex** function over X, Euclidean space.
Example :

$$\min_\theta \frac{1}{N}\sum_{i=1}^N \ell(\theta,(X_i,Y_i)), \quad \min_\theta \mathbf{E}_{(X,Y)}\ell(\theta,(X,Y)).$$

If $f(\cdot,\xi)$ smooth : **Stochastic gradient algorithm**

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

where $\gamma_n > 0$ and (ξ_n) i.i.d copies of ξ.

Stochastic Proximal Gradient algorithm

Regularized problem:

$$\min_{x \in \mathsf{X}} F(x) + G(x) \tag{1}$$

where

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

where ξ is a random variable, $f(\cdot,\xi)$ and $g(\cdot,\xi)$ are convex functions.

Stochastic Proximal Gradient algorithm :

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

where $\gamma_n > 0$ and (ξ_n) i.i.d copies of ξ and

$$\operatorname{prox}_g(x) = \arg\min_{y \in E} \frac{1}{2}\|x - y\|^2 + g(y)$$

for any convex function g.

Presentation of the algorithm

Convergence results

Applications

Decreasing step size

Theorem:
If $\gamma_n \to 0$, then, under mild assumptions ([BH'16]) : a.s,

$$x_n \longrightarrow_{n\to\infty} x_\star \in \arg\min F + G$$

Constant step size

If $\gamma_n \equiv \gamma > 0$, rewrite the algorithm

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

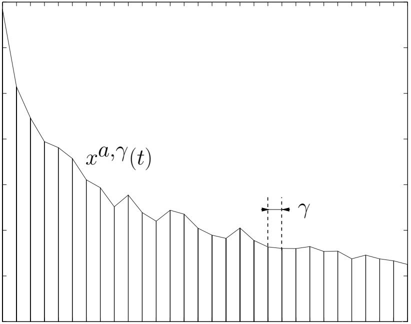

Figure 1: Continuous interpolated process : $x^{a,\gamma}(t)$ starting at $x^{a,\gamma}(0) = a$.

First step : Dynamical behavior

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

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

admits an unique solution x_a.

We look at $(x^{a,\gamma})_\gamma$ as a family of stochastic processes in $C(\mathbf{R}_+, \mathsf{X})$ 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 γ in order study its stability.

Stability assumptions :

- $F + G \longrightarrow_\infty +\infty$
- $\exists c > 0, x_\star \in \arg\min F + G$, for all $x \in \mathsf{X}$,

$$c\mathbf{E}\|\nabla f(x,\xi) - \nabla f(x_\star,\xi)\|^2 \leq \mathbf{E}\left(\langle\|\nabla f(x,\xi) - \nabla f(x_\star,\xi)|x - x_\star\rangle\right)$$

Then, using the dynamical behavior result,

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

Convergence result : Asymptotic behavior

Finally, **Theorem** ([BHS'17]) : Under the stability assumptions, and mild additional assumptions

$$\forall \varepsilon > 0, \quad \limsup_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbf{P} \left[d \left(x_k^\gamma, \arg\min F + G \right) \geq \varepsilon \right] \xrightarrow[\gamma\to 0]{} 0 \,.$$

Presentation of the algorithm

Convergence results

Applications

An application

Consider

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

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

Our problem:

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

with $F : \mathbf{R}^V \to \mathbf{R}$ convex, differentiable.

An application

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

$$\mathbf{E}\left(\frac{1}{L}\mathbf{TV}(x,\xi)\right) = \frac{1}{|E|}\mathbf{TV}(x,G).$$

Our problem is equivalent to

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

Stochastic Proximal Gradient algorithm ([SBH'16]):

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

Another application

Consider

- A family of closed convex sets $\mathcal{C}_1, \dots, \mathcal{C}_m$ of X
- Two convex functions F, G over X

Our problem:

$$\min_{x \in \mathcal{C}} F(x) + G(x), \quad \mathcal{C} := \bigcap_{i=1}^{m} \mathcal{C}_i \tag{3}$$

Let ι_C be the indicator function of a convex set C : $\iota_C(x) = 0$ if $x \in C$ and $\iota_C(x) = +\infty$ else.
Our problem is equivalent to

$$\min_{x \in \mathrm{X}} F(x) + G(x) + \sum_{i=1}^{m} \iota_{\mathcal{C}_i}(x).$$

Another application

Consider

- $\xi \sim \mathrm{Unif}(\{0, \dots, m\})$
- $h(x, 0) = (m+1)G(x)$
- $h(x, i) = \iota_{\mathcal{C}_i}(x)$ for all $i \in \{1, \cdots, m\}$

Then,

$$G(x) + \iota_{\mathcal{C}}(x) = \mathbf{E}(h(x, \xi)).$$

Our problem is equivalent to

$$\min_{x \in \mathsf{X}} F(x) + \mathbf{E}(h(x, \xi)).$$

Stochastic Proximal Gradient algorithm ([BH'16],[BHS'17]):

$$\begin{cases} \text{Sample} \quad \xi_{n+1} \sim \mathrm{Unif}(\{0, \dots, m\}) \\ \text{if } \xi_{n+1} = 0, \qquad x_{n+1} = \mathrm{prox}_{\gamma G}(x_n - \gamma \nabla F(x_n)) \\ \text{if } \xi_{n+1} = i > 0, \quad x_{n+1} = \mathrm{proj}_{\mathcal{C}_i}(x_n - \gamma \nabla F(x_n)) \end{cases}$$

Conclusion

- Constant step size stochastic approximation algorithm
- The ODE method
- Applications to structured penalizations

P. Bianchi and W. Hachem.
Dynamical behavior of a stochastic forward-backward algorithm using random monotone operators.
Journal of Optimization Theory and Applications, 2016.

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

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