Convergence of stochastic first order methods

Ion Necoara
University Politehnica Bucharest
Romania

Games, Dynamics and Optimization Workshop
Cluj, April 2019

Some results are based on some recent research with:

- Angelia Nedich

Motivation

- General stochastic optimization problem:

$$\min_{x\in\mathbb{R}^n} \mathbf{E}\left[f(x,\xi)\right] + g(x)$$

- Standard assumptions:
  - f (convex) function s.t. $f(x) = \mathbf{E}\left[f(x,\xi)\right]$ & g convex fct.
  - f is Lipschitz function or has Lipschitz gradient
- Many applications - computation of parameters for a system designed to make decisions based on yet unseen data (statistics, learning, estimation and control)
- "Almost all" learning problems can be formulated as above:
  - loss/fitting function $f(x) = \mathbf{E}\left[f(x,\xi)\right]$ with ξ random variable
  - Empirical risk minimization (finite sum): $f(x) = \frac{1}{m}\sum_{i=1}^{m} f(x,\xi_i)$
  - g regularizer (avoid overfitting, impose sparsity, or constraints)
- Solved "almost exclusively" by first order methods
- Stochastic (minibatch) first order methods have become *de facto* algorithmic choice for large-scale learning!

Algorithmic solution - stochastic case

General stochastic optimization problem:

$$\min_x \; f(x) + g(x) \qquad (:= \mathbf{E}\left[f(x,\xi)\right] + g(x))$$

Assume g admits a tractable *proximal operator*:

$$\mathrm{prox}_{\alpha g}(x) = \arg\min_{y \in \mathbb{R}^m} g(y) + \frac{1}{2\alpha}\|y - x\|^2.$$

Basic method - proximal gradient method:

$$(PG): \quad x_{k+1} = \mathrm{prox}_{\alpha_k g}(x_k - \alpha_k \nabla f(x_k))$$

- PG requires access to *full* gradient; e.g. in finite sum case we need to compute a large sum $\nabla f(x) = \frac{1}{m}\sum_{i=1}^{m} \nabla f(x, \xi_i)$
- difficult to implement when m large or data arrives in streams
- α_k is *global* stepsize (learning rate) - difficult to compute

Algorithmic solution - stochastic case cont.

Stochastic convex optimization problem:

$$\min_{x} \; f(x) + g(x) \qquad (:= \mathbf{E}\left[f(x,\xi)\right] + g(x))$$

Standard settings:

- function $f(x) = \mathbf{E}\left[f(x,\xi)\right]$ convex (random variable $\xi \in \Omega$)
- access to either unbiased stochastic estimate of gradient of f:

$$\nabla f(x;\xi) \quad \text{s.t.} \quad \nabla f(x) = \mathbb{E}[\nabla f(x;\xi)],$$

- or access to stochastic estimate of proximal operator of f:

$$\operatorname{prox}_{\alpha f(\cdot,\xi)}(x) = \arg\min_{y\in\mathbb{R}^m} f(y,\xi) + \frac{1}{2\alpha}\|y-x\|^2.$$

(assuming each $f(\cdot,\xi)$ admits a tractable proximal operator)

Existing work

Convex optimization problem:

$$F^* = \min_{x \in \mathbb{R}^m} F(x) \quad (= \mathbf{E}\left[f(x,\xi)\right] + g(x))$$

Since proximal gradient requires full information $\rightarrow$ use simple methods (mixing optimization and statistics):

- Stochastic gradient descent (g indicator function) has bee analyzed separately: for Lipschitz functions (Nedich & Bertsekas '00); for functions with Lipschitz gradient (Moulines & Bach '11) $\Rightarrow$ no common analysis!
- Stochastic proximal gradient (g general convex function) has been analyzed under more conservative assumptions: e.g. gradient Lipschitz with bounded variance (Rosasco et al '14)
- Stochastic proximal point has been analyzed for $g \equiv 0$ and gradient Lipschitz (Boyd '16, N'17) $\Rightarrow$ no general analysis!
- Convergence analysis for general g is partial/missing
- Most convergence results are for variable stepsize $\alpha_k = c/k$.

Algorithmic solution - stochastic case cont.

Stochastic convex optimization problem:

$$F^* = \min_x F(x) \qquad (:= \mathbf{E}\left[f(x,\xi)\right] + g(x))$$

- Denote X^* set of optima and for given x define $x^* = \Pi_{X^*}(x)$
- We provide unifying analysis under more general assumptions

Assumption: (restricted) Lipschitz type condition:

$$(RL): \quad M + L(F(x) - F^*) \geq \mathbf{E}_\xi[\|\nabla f(x,\xi) + \partial g(x)\|^2] \quad \forall x$$

Assumption: (restricted) strong convexity type condition (N'15):

$$(RSC): \quad F(x) - F^* \geq \frac{\mu}{2}\|x - x^*\|^2 \quad \forall x.$$

$$\Downarrow$$

Remark - (RL)/(RSC) covers several important functional classes:

- RL - class of Lipschitz functions or with Lipschitz gradients
- RSC - larger class than strong conv. $(f(x) = h(Ax) + c^T x)$

Stochastic first order methods

Stochastic convex optimization problem:

$$F^* = \min_x F(x) \qquad (:= \mathbf{E}\left[f(x,\xi)\right] + g(x))$$

Stochastic proximal gradient (SPG) method (for $g = 0$ we obtain Stochastic Gradient Descent (SGD) method) - sample ξ:

$$x_{k+1} = \text{prox}_{\alpha_k g}\left(x_k - \alpha_k \nabla f(x_k, \xi_k)\right)$$

Stochastic proximal point (SPP) method - sample ξ:

$$x_{k+1/2} = \text{prox}_{\alpha_k f(\cdot,\xi_k)}(x_k) \quad \text{and} \quad x_{k+1} = \text{prox}_{\alpha_k g}(x_{k+1/2})$$

- SPG/SPP have simple iteration: require evaluation of "partial" $\nabla f(x_k, \xi_k)/\text{prox}_{\alpha_k f(\cdot,\xi_k)}$, not entire gradient ∇f or entire prox operator $\text{prox}_{\alpha f}$ $\quad\rightarrow\quad$ m times cheaper!
- SPG/SPP adequate for applications - data arrive in streams
- α_k positive stepsize (learning rate) matters for SPG/SPP

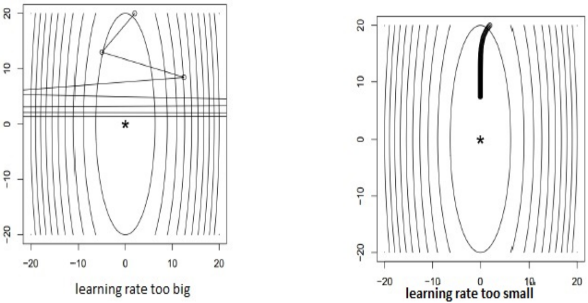
Stochastic first order methods

Stochastic proximal gradient (SPG) - sample ξ:

$$x_{k+1} = \text{prox}_{\gamma_k g}(x_k - \alpha_k \nabla f(x_k, \xi_k))$$

where α_k strictly positive stepsizes (learning rates).

- learning rate α_k matters for SPG



Question: When stochastic first order methods converge linearly?

Convergence rates of stochastic FOM - constant stepsize

Theorem (Descent Inequality)

Assume convexity and Lipschitz-like (sub)gradient condition RL hold. Then, the following recursive inequality holds for SPG/SPP:

$$\mathbf{E}\left[\|x_{k+1} - x_{k+1}^*\|^2\right] \\ \leq \mathbf{E}\left[\|x_k - x_k^*\|^2\right] - \alpha_k(2 - \alpha_k L)\mathbf{E}\left[F(x_k) - F(x_k^*)\right] + \alpha_k^2 M^2.$$

$$\text{Define}: R_0 = \|x_0 - x_0^*\|$$

Theorem (Constant stepsize)

SPG/SPP with $\alpha_k \equiv \alpha < 2/L$ under RL&RSC has "linear" conv.:

$$\mathbf{E}\left[\|x_k - x_k^*\|^2\right] \leq \left(1 - \mu\alpha + \frac{\mu L \alpha^2}{2}\right)^k R_0^2 + \frac{2M^2}{\mu(2 - L\alpha)}\alpha.$$

- linear convergence to noise dominated region whose radius $\sim \alpha$
- if $M = 0$ pure linear convergence!

Stochastic FOM - necessary & sufficient cond. linear conv.

Theorem (**Sufficient**)

SPG/SPP with $\alpha_k \equiv \alpha < 2/L$ *under RL & RSC has linear conv.:*

$$\mathbf{E}\left[\|x_k - x_k^*\|^2\right] \leq \left(1 - \mu\alpha + \frac{\mu L \alpha^2}{2}\right)^k R_0^2 + \frac{2M^2}{\mu(2 - L\mu)}\alpha.$$

- recall RL: $M + L(F(x) - F^*) \geq \mathbf{E}_\xi[\|\nabla f(x,\xi) + \partial g(x)\|^2]$
- linear convergence to noise dominated region whose radius$\sim \alpha$
- if $M = 0$ pure linear convergence!

Theorem (**Necessary**)

Assume $g \equiv 0$ *and* f *has unique minimizer satisfying RSC. Assume further that iterates of SPG/SPP with constant stepsize satisfy:*

$$\mathbf{E}_{\xi_k}[\|x_{k+1} - x_{k+1}^*\|^2] \leq c \cdot \|x_k - x_k^*\|^2, \quad \text{with} \quad c < 1.$$

Then, condition RL holds with $M \equiv 0$*! (i.e.* f *satisfies* $L(f(x) - f^*) \geq \mathbf{E}_\xi[\|\nabla f(x,\xi)\|^2]$*)*

Convergence rates - variable stepsize

Theorem (Sublinear convergence)

SPG/SPP with variable stepsize $\alpha_k = \min\left(\frac{1}{L}, \frac{c}{k+1}\right)$ *for some* $c > 0$ *under RL & RSC has sublinear convergence* $\mathcal{O}(1/k)$*:*

$$\mathbf{E}\left[\|x_k - x_k^*\|^2\right] \leq \frac{C(k_0, c, R_0)}{k} \qquad \text{if} \quad c\mu \geq 2$$

$$\mathbf{E}\left[\|x_k - x_k^*\|^2\right] \leq \frac{C(k_0, c, R_0)}{k^{0.5c\mu}} \qquad \text{if} \quad c\mu < 2$$

Remark 1: we can choose a larger stepsize α_k, with $\gamma \in (0,\ 1)$:

$$\alpha_k = \min\left(\frac{1}{L}, \frac{c}{(k+1)^\gamma}\right) \implies \mathcal{O}\left(\frac{1}{k^\gamma}\right) \text{ convergence rate}$$

Remark 2: Note that algorithm SPG is SPP scheme, but applied to the linearization of function $f(\cdot, \xi)$ at x:

$$l_f(y; x, \xi) = f(x; \xi) + \langle \nabla f(x; \xi), y - x \rangle \quad \leftrightarrow \quad f(y; \xi)$$

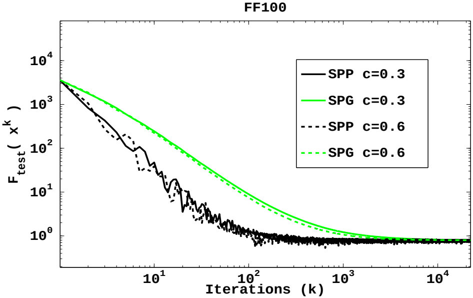
Thus, we expect faster convergence and more robustness for SPP!

Necoara, On the convergence rates of stochastic first order methods, Tech. Rep., 2018

Markowitz portfolio opt. - real data

$$\min_{x \in \mathbb{R}^m} \mathbf{E}\left[(a_\xi^T x - b)^2\right] + 1_X(x)$$

- $X = \{x : x \geq 0, e^T x = 1\}$ - easy to project $\mathcal{O}(m \log m)$ flops
- We compare SPG and SPP for learning rate α_k=min$(1/L, c/k)$, with $c = 0.3$ & 0.6. Dataset Fama and French (FF100, with 100 portfolios for 23.647 days)



- Plot value of objective function over datapoints in test partition F_{test} along iterations - one pass through data
- SPP is usually faster and more robust w.r.t. c than SPG

Application: convex feasibility

- SPG/SPP convergence linearly under restricted Lipschitz (RL with $M = 0$) & restricted strong convexity (RSC)
- Particular case of X represented as intersection of simple sets:

$$\text{find } x \in \cap_{\xi \in \Omega} X_\xi$$

reformulated as stochastic convex problems

$$(CFP): \quad \min_{x \in \mathbb{R}^m} \mathbf{E}\left[\|x - \Pi_{X_\xi}(x)\|^2\right] \quad \vee \quad \mathbf{E}\left[\mathbf{1}_{X_\xi}(x)\right]$$

- SPG (first formulation) $\vee$ SPP (second formulation) with $\alpha = 1$ becomes basic random projection algorithm:

$$(AP): \quad x_{k+1} = \Pi_{X_{\xi_k}}(x_k)$$

- For $X_\xi = \{x : a_\xi^T x = b_\xi\}$ AP becomes Kaczmarz algorithm
- If sets X_ξ satisfy linear regularity $\bar{\mu}\text{dist}(x, X) \leq \mathbf{E}[\text{dist}(x, X_\xi)]$, then *RSC* holds for CFP.
- Clearly, RL always holds for CFP, with $M = 0$ and $L = 1$.
- Hence, from previous theory recover linear convergence of AP.

N, Richtarik, Patrascu, *Randomized projection methods for convex feasibility problems*, Siopt, 2018

General convex feasibility with functional constraints

Consider convex feasibility problem (in functional constraints form):

$$\text{find } x \in X \equiv \{x \in X_0 : \ f_-(x,\xi) \leq 0 \quad \forall \xi \in \Omega\}$$

equivalently written as finite/infinite intersection of sets

$$\text{find } x \in X \equiv (\cap_{\xi \in \Omega} X_\xi) \cap X_0$$

where $X_\xi = \{x : \ f_-(x,\xi) \leq 0\}$. Note that if X_ξ not described by functional constraints we can just define $f_-(x,\xi) = \text{dist}^p(x, X_\xi)$, with $p = 1 \ \vee \ 2$ and $X_0 = \mathbb{R}^n$. Define a stochastic convex problem:

$$f(x,\xi) = \max{}^p(0, f_-(x,\xi)) \quad \Longrightarrow \quad \min_{x \in \mathbb{R}^m} f(x) \ \ (\equiv \mathbf{E}\,[f(x,\xi)])$$

Lemma 1: if X_0 compact and $f_-(\cdot,\xi)$ are Lipshitz or with gradient Lipschitz, then RL holds with $M = 0$, i.e. $L\,f(x) \geq \mathbf{E}[\|\nabla f(x,\xi)\|^2]$.

Lemma 2: if $f_-(\cdot,\xi)$ satisfy linear regularity $\bar{\mu}\text{dist}^2(x,X) \leq f(x)$ for all $\forall x \in X_0$, then RSC holds, i.e. $f(x) \geq \frac{\mu}{2}\|x - x^*\|^2$.

Remark: Linear regularity holds e.g. for polyhedral sets $f_-(x,\xi) = a_\xi^T x - b_\xi$ or more general under Slater type condition.

Convex feasibility with functional constraints cont.

Consider convex feasibility problem (in functional constraints form):

$$\text{find } x \in X \equiv \{x \in X_0 : \ f_{-}(x,\xi) \leq 0 \quad \forall \xi \in \Omega\}$$

reformulated as a stochastic convex problem:

$$f(x,\xi) = \max^p(0, f_{-}(x,\xi)) \quad \Longrightarrow \quad \min_{x \in \mathbb{R}^m} f(x) \ \ (\equiv \mathbf{E}\left[f(x,\xi)\right])$$

Consider Polyak's stochastic (sub)gradient algorithm:

$$x_{k+1} = \Pi_{X_0}\left[x_k - \alpha \frac{f(x_k,\xi_k)}{\|g_k\|^2}\, g_k\right]$$

where $g_k \in \partial f(x_k,\xi_k)$ if $f(x_k,\xi_k) > 0$ and $d_k \equiv d \neq 0$, otherwise.

Theorem

Assume X_0 compact, $f_{-}(\cdot,\xi)$ are Lipshitz or with gradient Lipschitz, and satisfying linear regularity, then linear converge:

$$\mathbf{E}\left[dist^2(x_k, X)\right] \leq (1-q)^k\, dist^2(x_0, X).$$

N, Nedich, *Random minibatch subgradient algorithms for convex feasibility problems*, CDC, 2019

Convex problems with functional constraints

After investigating feasibility problems, it is also natural to consider on top of intersection of sets some objective function:

$$\min_{x \in X_0} f(x) \quad \text{s.t.} \quad x \in \cap_{\xi \in \Omega} X_\xi$$

We assume X_ξ have functional representation, thus feasible set is given by finite intersection of convex sets of the form:

$$X = X_0 \cap \left(\cap_{\xi \in \Omega} X_\xi \right), \text{ with } X_\xi = \{x : f_-(x, \xi) \leq 0\}$$

- This model have appeared in Facchinei's talk today ("optimization problems with complex geometry").
- Many algorithms for solving this general problem:
  - Lagrangian methods: Hestenes'69,Sabach et al'18,Combettes et al'11,Eckstein'93,Rockafellar'76,...
  - Linearization methods: Nesterov'04, Teboulle et al'10, Drusvyatskiy et al'16, Bolte et al'18, Salzo&Villa'12,...
- Usually work with all $f_-(x, \xi) \implies$ subproblem is difficult!

Assumptions

We aim at solving problems with complex geometry (m large):

$$\min_{x \in X_0} f(x) \quad \text{s.t.} \quad f_{-}(x,\xi) \leq 0 \ \ \forall \xi \in \Omega$$

- Assume f and constraint functions $f_{-}(\cdot,\xi)$ convex and nonsmooth
- Objective function f is μ restricted strongly convex (RSC)
- Subgradients of f and $f_{-}(\cdot,\xi)$ uniformly bounded on X_0:

$$\|g_f(x)\| \leq M_f, \ \ \|g_\xi(x)\| \leq M \quad \forall x \in X_0$$

- If X_ξ simple for projection, then one may choose an alternative equivalent description of the constraint sets by letting $f_{-}(x,\xi) = \text{dist}(x, X_\xi)$, then $g_\xi(x) = \frac{x - \Pi_{X_\xi}(x)}{\|x - \Pi_{X_\xi}(x)\|} \in \partial f_{-}(x,\xi)$
- However, our approach allows to tackle "complicated" sets
- Assume linear regularity for sets ($f(x,\xi) = \max(0, f_{-}(x,\xi))$):

$$\mu \cdot \text{dist}^2(x, X) \leq \mathbf{E}\left[f(x,\xi)\right] \equiv f(x) \quad \forall x \in X_0$$

Subgradient with minibatch feasibility updates

Our method takes:

- one subgradient step for the objective function
- followed by $\tau = |J_k|$ feasibility updates (choose $J_k \subset [m], J_k \sim \mathbf{P}$)
- feasibility updates are taken in parallel or sequential!

$$v_k = \Pi_{X_0}(x_k - \alpha_k g_f(x_k))$$

$$z_k^i = v_k - \beta_k \frac{f(v_k, i)}{\|d_k^i\|^2} d_k^i \quad \forall i \in J_k$$

$$x_{k+1} = \Pi_{X_0}(\bar{z}_k), \quad \text{with } \bar{z}_k = \frac{1}{\tau} \sum_{i=1}^{\tau} z_k^i$$

- Here, $g_f(x_k) \in \partial f(x_k)$ and $d_k^i \in \partial f(v_k, i)$
- Do not require projections, just subgradient evaluation of g_i
- Variants of this algorithm for convex case and $|J_k| = 1$ considered in Polyak'69, Nedich'11, Nesterov'15 $\to O(1/\sqrt{k})$!
- **Question**: minibatch setting influences convergence rate?

Convergence rates

Story is long, but we get some recurrence relation in expectation that allows to obtain convergence rates:

- Consider stepsizes $\alpha_k = \frac{4}{\mu(k+1)}$ and extrapolated β_k
- Define average sequence $\hat{x}_k = 1/S \sum_{j=0}^{k-1} (j+1)^2 x_j$

Theorem (Sublinear convergence $\mathcal{O}(1/k)$)

Under above settings, average sequence $\hat{x}_k$ generated by parallel/sequential subgradient method with random minibatch feasibility updates converges as:

$$\mathbf{E}\left[dist_X(\hat{x}_k)\right] \leq \mathcal{O}\left(\frac{1}{c_\tau k}\right), \quad \mathbf{E}\left[|f(\hat{x}_k) - f^*|\right] \leq \mathcal{O}\left(\frac{1}{k} + \frac{1}{c_\tau k}\right).$$

- feasibility estimate depends explicitly on batchsize τ via c_τ
- suboptimality estimate contains a term not depending on τ

N, Nedich, *Random minibatch subgradient algorithms for convex problems with functional constraints*, 2019

Conclusions

This talk:

- Convergence analysis of stochastic first order methods (SPG & SPP) under general assumptions
- Cover important functional classes: functions with bounded/ Lipschitz (sub)gradients & restricted strong convexity
- Convergence rates for constant/variable stepsizes
- Derive conditions for linear convergence (necessary&sufficient)
- Extension to convex feasibility problems (linear convergence)
- Extension to convex problems with many functional constraints

Future work:

- More general stochastic models: $\min_{x \in \mathbb{R}^m} \mathbf{E}\left[f(x,\xi) + g(x,\xi)\right]$
- Using accelerated gradient schemes/second-order information
- Parallel and asynchronous implementations

References

- Necoara, Richtarik, Patrascu, *Randomized projection methods for convex feasibility problems: conditioning and convergence rates*, SIOPT, 2019.
- Necoara, *Convergence of stochastic first order methods for composite convex optimization*, Tech. Rep., 2018.
- Nedich, Necoara, *Random minibatch subgradient algorithms for convex problems with functional constraints*, Tech. Rep., 2019.
- Necoara, Nedich, *Random minibatch subgradient algorithms for convex feasibility problems*, CDC, 2019.