Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning

Francis Bach
INRIA - ENS

Eric Moulines
Telecom ParisTech

Stochastic approximation

- **Context**: Large-scale learning ("large p, large n, large k")
- **Goal**: Minimizing a function f defined on a Hilbert space $\mathcal{H}$
  - given only unbiased estimates $f'_n(\theta_n)$ of its gradients $f'(\theta_n)$ at certain points $\theta_n \in \mathcal{H}$
- **Stochastic approximation**
  - Observation of $f'_n(\theta_n) = f'(\theta_n) + \varepsilon_n$
  - ε_n = additive noise (typically i.i.d.)
- **Machine learning - statistics**
  - $f_n(\theta) = \ell(\theta, z_n)$ where z_n is an i.i.d. sequence
  - $f(\theta) = \mathbb{E} f_n(\theta)$ = generalization error of predictor θ
  - Typically $f_n(\theta) = \frac{1}{2}(\langle x_n, \theta\rangle - y_n)^2$ or $\log[1 + \exp(-y_n \langle x_n, \theta\rangle)]$

Convex stochastic approximation

- **Key properties of f and/or f_n**
  - Smoothness: f B-Lipschitz continuous, f' L-Lipschitz continuous
  - Strong convexity: f μ-strongly convex
- **Key algorithm:** Stochastic gradient descent (a.k.a. Robbins-Monro)

$$\boxed{\theta_n = \theta_{n-1} - \gamma_n f_n'(\theta_{n-1})}$$

  - Polyak-Ruppert averaging: $\bar{\theta}_n = \frac{1}{n}\sum_{k=0}^{n-1}\theta_k$
  - Which learning rate sequence γ_n? Classical setting: $\boxed{\gamma_n = Cn^{-\alpha}}$
- **Desirable practical behavior**
  - Applicable (at least) to least-squares and logistic regression
  - Robustness to (potentially unknown) constants (L,B,μ)
  - Adaptivity to difficulty of the problem (e.g., strong convexity)

Summary of new results

- Stochastic gradient descent with learning rate $\gamma_n = Cn^{-\alpha}$
- **Strongly convex smooth objective functions**
  - Old: $O(n^{-1})$ rate achieved without averaging for $\alpha = 1$
  - New: $O(n^{-1})$ rate achieved with averaging for $\alpha \in [1/2, 1]$
  - Non-asymptotic analysis with explicit constants
- **Non-strongly convex smooth objective functions**
  - Old: $O(n^{-1/2})$ rate achieved with averaging for $\alpha = 1/2$
  - New: $O(\max\{n^{1/2-3\alpha/2}, n^{-\alpha/2}, n^{\alpha-1}\})$ rate achieved without averaging for $\alpha \in [1/3, 1]$,
- **Take-home message**
  - Use $\alpha = 1/2$ with averaging to be adaptive to strong convexity