Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$

Francis Bach
INRIA - ENS

Eric Moulines
Telecom ParisTech

Large-scale supervised learning
Stochastic approximation

- **Context**: Learning from large datasets with a single pass

- **Goal**: Minimize generalization error $\boxed{\mathbb{E}_{p(x,y)}\ell(y, \theta^\top \Phi(x))}$
  - Linear predictions $\theta^\top \Phi(x)$, with $\Phi(x) \in \mathbb{R}^d$
  - Smooth loss ℓ (least-squares and logistic)
  - Learning from stream of i.i.d. data (x_n, y_n), $n \geqslant 1$

- **Main approach**: (averaged) stochastic gradient descent

$$\boxed{\theta_n = \theta_{n-1} - \gamma_n f_n'(\theta_{n-1})} \text{ with } \left\{ \begin{array}{l} f_n(\theta) = \ell(y_n, \theta^\top \Phi(x_n)) \\ f_n'(\theta) = \ell'(y_n, \theta^\top \Phi(x_n))\, \Phi(x_n) \end{array} \right.$$

  - Polyak-Ruppert averaging: $\bar{\theta}_n = \frac{1}{n+1} \sum_{k=0}^{n} \theta_k$

Convex stochastic approximation
Existing work

- **Known global minimax rates of convergence for non-smooth problems** (Nemirovski and Yudin, 1983)
  - Strongly convex: $O((\mu n)^{-1})$
    Attained by averaged stochastic gradient descent with $\gamma_n \propto (\mu n)^{-1}$
  - Non-strongly convex: $O(n^{-1/2})$
    Attained by averaged stochastic gradient descent with $\gamma_n \propto n^{-1/2}$

- **Breaking lower bounds**
  - A single algorithm for smooth problems with convergence rate $O(1/n)$ in all situations
  - Robustness to ill-conditioning and step-size selection

Provable convergence in $O(1/n)$ for smooth functions

- **Least-squares regression**
  - Constant step-size averaged stochastic gradient descent

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

- **Logistic regression**
  - Novel constant step-size online Newton algorithm
  - Same complexity of $O(d)$ per iteration

$$\boxed{\theta_n = \theta_{n-1} - \gamma \Big[f_n'(\bar{\theta}_{n-1}) + f_n''(\bar{\theta}_{n-1})(\theta_{n-1} - \bar{\theta}_{n-1}) \Big]}$$

- **Step-size** $\gamma = 1/4R^2$
  - State-of-the-art performance in theory and experiments