When does stochastic gradient descent work without variance reduction?
Chuan-Zheng Lee, Hüseyin İnan

Background

- Stochastic gradient descent (SGD) tries to minimize
  \[ J(x) = \frac{1}{m} \sum_{i=1}^{m} J_i(x) \]
  with initialization at \( x^{(0)} \) and iterations
  \[ x^{(t+1)} = x^{(t)} - \mu J_i(x^{(t)}) \]

  \((t \in \{1, \ldots, m\} \text{ randomly chosen for each } t)\)
- Constant \( \mu = \mu \Rightarrow \) no convergence in general due to variance of \( \nabla J_i(x^{(t)}) \)
- Variance reduction helps it converge, e.g. \( \mu_t \propto 1/\sqrt{t} \)

- [KO16] introduced an SGD-like algorithm for phase retrieval that converges with fixed step size \( \mu \)
- Possible explanations:
  - The form of the loss function
  - The assumption of sensing vectors being i.i.d. \( \mathcal{N}(0, I) \)

Least squares

- Consider the problem \( \{ y \in \mathbb{R}^m, A \in \mathbb{R}^{m \times n}, m \geq n \} \)
  \[ \min_{x \in \mathbb{R}^n} \frac{1}{2m} \| y - Ax \|^2 = \frac{1}{m} \sum_{i=1}^{m} (y_i - a_i^T x)^2 / J_i(x) \]
- Denote \( x^* = \arg \min_{x} J(y, Ax) \)
- Assume rows of \( A \) i.i.d. \( \sim \mathcal{N}(0, I), i = 1, \ldots, m \)

Our project objective

For what other classes of problems can we find similar SGD convergence conditions with constant \( \mu ? \)

Reference


Theorem 1

If \( y \in \mathcal{R}(A) \), then there exist universal constants \( C, c_0, c_1, c_2 > 0, \rho \in (0, 1], \) such that with \( \mu = c_2/n \) and \( m \geq c_0 n \), the SGD iterates satisfy w.p. \( \geq 1 - C \rho^{-cn} \)

\[ E_{[\{ Y_{i=1}^{m} \}} \| x^{(t)} - x^* \|^2 \leq \left(1 - \frac{\mu}{n} \right)^t \| x^{(0)} - x^* \|^2 \]

where the last step uses the following concentration bounds:

Lemma 1. \( \forall \delta > 0, \) there exist universal constants \( C, c_0, c_1 > 0 \) such that if \( m \geq c_0 \delta^{-2} \), then w.h.p. for all \( h \in \mathbb{R}^n \)

\[ (1 - \delta) \| h \|^2 \leq E_{[\{ Y_{i=1}^{m} \}} \| h_{\mu} \|^2 \leq (1 + \delta) \| h \|^2 \]

Lemma 2. For all \( i = 1, \ldots, m, \| a_i \|^2 \leq 6 n \) w.h.p.

The following plot confirms the \( (1 - \frac{\mu}{n}) \) relationship:

Support vector machines

- Consider the problem
  \[ \min_{x \in \mathbb{R}^n} J(x) = \frac{1}{2m} \sum_{i=1}^{m} \max \{ 1 - y_i a_i^T x, 0 \} / J_i(x) = g_{\max}(y, x) \]
- Assume \( a_i \sim \mathcal{N}(0, I) \) i.i.d. and \( y_i a_i^T x^* \geq 1 + \delta, i = 1, \ldots, m \) (i.e. it is the minimum excess margin)

Future work

- Apply methods to other common loss functions
- Introduce regularization term to Theorem 3
- Having a perfect fit seems to help—try generalizing
- Simulations indicate \( a_i \) need not be Gaussian—explore i.i.d. with arbitrary distribution