EE270
Large scale matrix computation, optimization and learning

Instructor : Mert Pilanci

Stanford University

Thursday, Feb 27 2020
Randomized Linear Algebra and Optimization
Lecture 16: Stochastic Gradient Methods and Randomized Kaczmarz Algorithm
Empirical Risk Minimization

- Let \( \{a_i, y_i\}, \ i = 1, \ldots, n \) be training data
- Empirical risk minimization

\[
\min_x \frac{1}{n} \sum_{i=1}^{n} f(x, a_i, y_i)
\]

- Examples:
  - Least-Squares problems: \( f(x, a_i, y_i) = (a_i^T x - y_i)^2 \)
  - Logistic regression: \( f(x, a_i, y_i) = \log(1 + e^{a_i^T x_i y_i}) \)
Empirical Risk Minimization

- Let \( \{a_i, y_i\}, \ i = 1, \ldots, n \) be training data
- Empirical risk minimization

\[
\min_x \frac{1}{n} \sum_{i=1}^{n} f(x, a_i, y_i)
\]

- Examples:
  - Least-Squares problems: \( f(x, a_i, y_i) = (a_i^T x - y_i)^2 \)
  - Logistic regression: \( f(x, a_i, y_i) = \log(1 + e^{a_i^T x; y_i}) \)
- empirical risk approximates the population (expected) risk:

\[
\mathbb{E} f(x, a_i, y_i)
\]

where the expectation is taken over the data
Stochastic Programming

\[
\min_x \mathbb{E} f(x, a_i, y_i) \\
F(x)
\]

▶ A simple approach:

\[
x_{t+1} = x_t - \mu \nabla F(x_t) \\
= x_t - \mu \mathbb{E} f(x, a_i, y_i) \\
\approx x_t - \mu f(x, a_{i_t}, y_{i_t})
\]

where \(i_t\) is a random index
Stochastic Gradient Descent (SGD)

\[
\min_x \mathbb{E} f(x, a_i, y_i) \\
F(x)
\]

Consider the iterative algorithm

\[ x_{t+1} = x_t - \mu_t g_t \]

where \( g_t \) is an unbiased estimate of \( \nabla F(x_t) \)

\[ \mathbb{E} g_t = \nabla F(x_t) \]
SGD for Empirical Risk Minimization

- Let \( \{a_i, y_i\}, \ i = 1, ..., n \) be training data
- Empirical risk minimization

\[
\min_x \frac{1}{n} \sum_{i=1}^{n} f(x, a_i, y_i)
\]

- Choose an index \( i_t \) uniformly at random and let

\[
x_{t+1} = x_t - \mu_t \nabla_t f(x, a_{i_t}, y_{i_t})
\]
Convergence of SGD for strongly convex problems

\[
\min_x \mathbb{E} f(x, a_i, y_i) \quad \text{subject to} \quad F(x)
\]

- SGD with constant step size \( \mu \)

\[
x_{t+1} = x_t - \mu \nabla_t f(x, a_t, y_t)
\]

Assumptions

- \( F \) is strongly convex with parameters \( \beta_- \) and \( \beta_+ \)
- \( g_t \) is an unbiased estimate of \( \nabla F(x_t) \) and it holds that
  \[
  \mathbb{E} \| g_t \|_2^2 \leq \sigma_g^2 + c_g \| \nabla F(x) \|_2^2
  \]
- step size \( \mu \leq \frac{1}{\beta_+ c_g} \)
Convergence of SGD for strongly convex problems

\[
\min_x \mathbb{E} f(x, a_i, y_i)
\]

\[F(x)\]

- SGD with constant step size \( \mu \)

\[
x_{t+1} = x_t - \mu \nabla_t f(x, a_i, y_i)
\]

Assumptions

- \( F \) is strongly convex with parameters \( \beta_- \) and \( \beta_+ \)
- \( g_t \) is an unbiased estimate of \( \nabla F(x_t) \) and its holds that
- \( \mathbb{E}\|g_t\|^2 \leq \sigma_g^2 + c_g \|\nabla F(x)\|^2 \)
- step size \( \mu \leq \frac{1}{\beta_+ c_g} \)
- Theorem:

\[
\mathbb{E} [F(x_t) - F(x^*)] \leq \mu \frac{\beta_+ \sigma_g^2}{2 \beta_-} + (1 - \mu \beta_-)^t (F(x_0) - F(x^*))
\]
Convergence of SGD for strongly convex problems

Assumptions

► $F$ is strongly convex with parameters $\beta_-$ and $\beta_+$
► $g_t$ is an unbiased estimate of $\nabla F(x_t)$ and its holds that
► $\mathbb{E}\|g_t\|_2^2 \leq \sigma_g^2 + c_g\|\nabla F(x)\|_2^2$
► **Theorem:**

$$\mathbb{E}[F(x_t) - F(x^*)] \leq \mu \frac{\beta + \sigma_g^2}{2\beta_-} + (1 - \mu \beta_-)^t(F(x_0) - F(x^*))$$

► converges to a neighborhood of the optimum $x^*$
► converges to $x^*$ when the $\sigma_g = 0$, i.e., gradient is noise-free
► in practice we can reduce the stepsize whenever the progress stalls
Convergence of SGD with diminishing step-sizes

Assumptions

- $F$ is strongly convex with parameters $\beta_-$ and $\beta_+$
- $g_t$ is an unbiased estimate of $\nabla F(x_t)$ and its holds that
- $\mathbb{E} \| g_t \|_2^2 \leq \sigma_g^2$
- $\mu_t = \frac{\mu}{t+1}$ for some $\mu > \frac{1}{2\beta_-}$
- **Theorem:**

$$\mathbb{E} [F(x_t) - F(x^*)] \leq \frac{C_\mu}{t + 1}$$

where $C_\mu = \max(\frac{2\mu^2\sigma_g^2}{2\beta_-\mu - 1}, \|x_0 - x^*\|_2)$
Comparison with Gradient Descent

- **Stochastic Gradient Descent**
  - per iteration cost $O(d)$
  - number of iterations $O\left(\frac{1}{\epsilon}\right)$
  - total cost $O\left(\frac{d}{\epsilon}\right)$

- **Gradient Descent**
  - per iteration cost $O(nd)$
  - number of iterations $O\left(\log\left(\frac{1}{\epsilon}\right)\right)$
  - total cost $O\left(nd \log\left(\frac{1}{\epsilon}\right)\right)$

SGD can be faster for large $n$ and low accuracy $\epsilon$
SGD for Least Squares Problems

\[
\min \|Ax - b\|_2^2 = \sum_{i=1}^{n} (a_i^T x - b_i)^2
\]

- Gradient: \( \nabla f(x) = A^T (Ax - b) = \sum_{i=1}^{n} a_i (a_i^T x - b_i) \)
- A stochastic gradient: \( g_t = a_{i_t} (a_{i_t}^T x - b_{i_t}) \) where \( i_t \) is a random index
- SGD iterations

\[
x_{t+1} = x_t - \mu_t (a_{i_t}^T x_t - b_{i_t}) a_i
\]

- Sketched Gradient Descent

\[
x_{t+1} = x_t - \mu_t A^T S_t^T S_t (Ax_t - b)
\]

where \( \mathbb{E} S_t^T S_t = I \)
SGD for Least Squares Problems

\[
\min \|Ax - b\|^2_2 = \sum_{i=1}^{n} (a_i^T x - b_i)^2
\]

▶ SGD iterations

\[
x_{t+1} = x_t - \mu_t (a_i^T x_t - b_i) a_i
\]

▶ step-size \( \mu_t = \frac{1}{\|a_{it}\|^2_2} \)

\[
x_{t+1} = x_t - \frac{a_{it}^T x_t - b_{it}}{\|a_{it}\|^2_2} a_i
\]
Assume that $b = Ax^*$ and define $\Delta_t = A(x_t - x^*)$

$\Delta_{t+1} = \Delta_t - \frac{a_t a_t^T}{\|a_t\|_2^2} \Delta_t = P_t \Delta_t$

where $P_t := I - \frac{a_t a_t^T}{\|a_t\|_2^2}$ is a projection matrix

after $T$ iterations

$\Delta_T = P_{T-1} \ldots P_1 \Delta_1$
Consider a sampling distribution $p_1, \ldots, p_n$, i.e., we sample the $i$-th data row $a_i, y_i$ with probability $p_i$

SGD iterations with sampling distribution $\{p_i\}_{i=1}^n$

$$x_{t+1} = x_t - \mu_t g_t$$

where $g_t = \frac{1}{p_{it}} (a_{it}^T x_t - b_{it}) a_i$

unbiased gradient estimate

$$\mathbb{E} g_t = A^T (Ax_t - b)$$
Convergence Analysis: General Sampling Distributions

▶ Assume that \( b = Ax^* \) and define \( \Delta_t = A(x_t - x^*) \)

▶ set step-size \( \mu_t = 1 \)

\[
x_{t+1} = x_t - \frac{1}{p_{it}} (a_{it}^T x_t - b_{it}) a_i
\]

▶ \( \Delta_{t+1} = \Delta_t - \frac{a_i a_{it}^T}{p_{it}} \Delta_t \)

\[
\mathbb{E}\|\Delta_{t+1}\|^2_2 = \mathbb{E}\|\Delta_t - \frac{a_i a_{it}^T}{p_{it}} \Delta_t\|^2_2
\]

\[
= \mathbb{E}\|\Delta_t\|^2_2 - 2\Delta_t^T \frac{a_i a_{it}^T}{p_{it}} \Delta_t + \left\| \frac{a_i a_{it}^T}{p_{it}} \Delta_t \right\|^2_2
\]

\[
= \mathbb{E} \Delta_t^T \left( I - 2 \frac{a_i a_{it}^T}{p_{it}} + \frac{a_i a_{it}^T \|a_it\|^2_2}{p_{it}^2} \right) \Delta_t
\]
Convergence Analysis: General Sampling Distributions

Taking expectations

\[ \mathbb{E}\|\Delta_{t+1}\|_2^2 = \Delta_t^T \left( I - \sum_{i=1}^n 2a_i a_i^T + \sum_{i=1}^n \frac{a_i a_i^T \|a_i\|_2^2}{p_i} \right) \Delta_t \]

- note that right-hand-side, hence the optimal distribution depends on the previous error \( \Delta_t \)
- we can minimize the upper-bound with respect to the sampling distribution

\[ \Delta_t^T \left( \sum_{i=1}^n \frac{a_i a_i^T \|a_i\|_2^2}{p_i} \right) \Delta_t \leq \lambda_{\text{max}} \left( \sum_{i=1}^n \frac{a_i a_i^T \|a_i\|_2^2}{p_i} \right) \|\Delta_t\|_2^2 \]
Convergence Analysis: General Sampling Distributions

- Taking expectations

\[
\mathbb{E} \| \Delta_{t+1} \|^2_2 = \Delta_t^T \left( I - \sum_{i=1}^{n} 2a_i a_i^T + \sum_{i=1}^{n} \frac{a_{it} a_{it}^T \|a_i\|^2}{p_i} \right) \Delta_t
\]

- note that right-hand-side, hence the optimal distribution depends on the previous error \( \Delta_t \)

- we can minimize the upper-bound with respect to the sampling distribution

\[
\Delta_t^T \left( \sum_{i=1}^{n} \frac{a_{it} a_{it}^T \|a_i\|^2}{p_i} \right) \Delta_t \leq \lambda_{\text{max}} \left( \sum_{i=1}^{n} \frac{a_{it} a_{it}^T \|a_i\|^2}{p_i} \right) \| \Delta_t \|^2_2
\]

\[
\leq \text{Tr} \left( \sum_{i=1}^{n} \frac{a_{it} a_{it}^T \|a_i\|^2}{p_i} \right) \| \Delta_t \|^2_2
\]
Convergence Analysis: General Sampling Distributions

- minimizing the upper-bound

$$\min_{\mathbf{p}} \sum_{i=1}^{n} p_i \left[ \sum_{i=1}^{n} a_i a_i^T \frac{\|a_i\|^2}{p_i} \right]$$

- equivalent to

$$\min_{\mathbf{p}} \sum_{i=1}^{n} p_i \left[ \sum_{i=1}^{n} \frac{\|a_i\|^4}{p_i} \right]$$
Convergence Analysis: General Sampling Distributions

- minimizing the upper-bound

\[ \min_{p \sum_{i=1}^n p_i=1, p_i \geq 0} \text{Tr} \left( \sum_{i=1}^n \frac{a_i a_i^T \|a_i\|_2^2}{p_i} \right) \]

- equivalent to

\[ \min_{p \sum_{i=1}^n p_i=1, p_i \geq 0} \sum_{i=1}^n \frac{\|a_i\|_2^4}{p_i} \]

- optimal sampling distribution

\[ p_i^* = \frac{\|a_i\|_2^2}{\sum_{j=1}^n \|a_j\|_2^2} = \frac{\|a_i\|_2^2}{\|A\|_F^2} \]

- same distribution as in approximate matrix multiplication

\[ A^T A \sim A^T S^T SA \]
Randomized Kaczmarz Algorithm

- optimal sampling distribution

\[ p_i^* = \frac{\|a_i\|^2}{\sum_{j=1}^n \|a_j\|^2} = \frac{\|a_i\|^2}{\|A\|^2_F} \]

- consider step-size \( \mu_t \)

\[ x_{t+1} = x_t - \mu_t \frac{1}{p_{it}} a_{it} (a_{it}^T x - b_{it}) = x_t - \mu_t \frac{\|A\|^2_F}{\|a_{it}\|^2} a_{it} (a_{it}^T x - b_{it}) \]

- set the step-size \( \mu_t = \frac{1}{\|A\|^2_F} \)

- this is called Randomized Kaczmarz Algorithm

\[ x_{t+1} = x_t - \frac{1}{\|a_{it}\|^2} a_{it} (a_{it}^T x - b_{it}) \]

- convergence analysis yields

\[ \Delta_{t+1} = \left( I - \frac{a_i a_i^T}{\|a_{it}\|^2} \right) \Delta_t \]

\[ = P_t \Delta_t \]

- where \( P_t = I - \frac{a_i a_i^T}{\|a_{it}\|^2} \)
Convergence rate

\[ \mathbb{E}\|\Delta_{t+1}\|^2 = \Delta_t^T (I - \frac{1}{\|A\|_F^2} A^T A) \Delta_t \]

\[ \geq (1 - \frac{\lambda_{\text{min}}}{\|A\|_F^2}) \|\Delta_t\|^2 \]

- recursively applying the above bound and taking conditional expectations

after \( T \) iterations we obtain

\[ \mathbb{E}\|\Delta_T\|^2 \leq (1 - \frac{\lambda_{\text{min}}}{\|A\|_F^2})^T \]