EE270
Large scale matrix computation, optimization and learning

Instructor : Mert Pilanci

Stanford University

Randomized Linear Algebra and Optimization Lecture 16: Stochastic Gradient Methods and Randomized Kaczmarz Algorithm

Empirical Risk Minimization

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

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

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

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

- Examples:

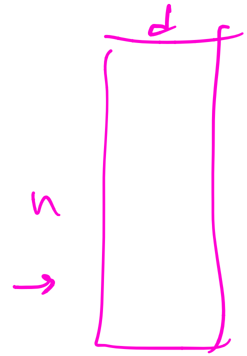
  Least-Squares problems: $f(x, a_i, y_i) = (a_i^T x - y_i)^2$

  Logistic regression: $f(x, a_i, y_i) = \log(1 + e^{a_i^T x_i y_i})$
- empirical risk approximates the population (expected) risk:

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

where the expectation is taken over the data

Stochastic Programming



$$\min_{x} \underbrace{\mathbb{E} f(x, a_i, y_i)}_{F(x)}$$

- A simple approach:

$$\begin{aligned} x_{t+1} &= x_t - \mu \nabla F(x_t) \\ &= x_t - \mu \mathbb{E} f(x, a_i, y_i) \\ &\approx x_t - \mu f(x, a_{i_t}, y_{i_t}) \end{aligned}$$

where i_t is a random index

Stochastic Gradient Descent (SGD)

$$\min_x \underbrace{\mathbb{E} f(x, a_i, y_i)}_{F(x)}$$

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

Stoch. Grad Descent

- 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

$$\min_x \underbrace{\mathbb{E} f(x, a_i, y_i)}_{F(x)}$$

- SGD with constant step size μ

$$x_{t+1} = x_t - \mu \nabla_t f(x, a_{i_t}, y_{i_t})$$

Assumptions

- F is strongly convex with parameters β_- and β_+
- g_t is an unbiased estimate of $\nabla F(x_t)$ and its holds that
- $\mathbb{E}\|g_t\|_2^2 \leq \sigma_g^2 + c_g \|\nabla F(x)\|_2^2$
- step size $\mu \leq \frac{1}{\beta_+ c_g}$

Convergence of SGD for strongly convex problems

$$\min_x \underbrace{\mathbb{E} f(x, a_i, y_i)}_{F(x)}$$

- SGD with constant step size μ

$$x_{t+1} = x_t - \mu \nabla_t f(x, a_{i_t}, y_{i_t})$$

Assumptions

- F is strongly convex with parameters β_- and β_+
- g_t is an unbiased estimate of $\nabla F(x_t)$ and its holds that
- $\mathbb{E}\|g_t\|_2^2 \leq \sigma_g^2 + c_g \|\nabla F(x)\|_2^2$
- step size $\mu \leq \frac{1}{\beta_+ c_g}$
- **Theorem:**

$$\mathbb{E}\left[F(x_t) - F(x^*)\right] \leq \overbrace{\mu \frac{\beta_+ \sigma_g^2}{2\beta_-}}^{\text{error floor}} + \overbrace{(1 - \mu\beta_-)^t (F(x_0) - F(x^*))}^{\to 0}$$

Convergence of SGD for strongly convex problems

Assumptions

- F is strongly convex with parameters β_- and β_+
- 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:**

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

- converges to a neighborgood 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

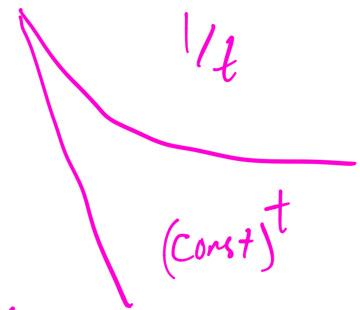
Convergence of SGD with diminishing step-sizes

Assumptions

- F is strongly convex with parameters β_- and β_+
- 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:**

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

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



Comparison with Gradient Descent

n $\underbrace{\boxed{}}^{d}$

$$\nabla F(x) = \sum_i a_i^T (a_i^T x - b)$$

$$\approx a_j^T (a_j^T x - b_j)$$

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

SGD can be faster for large n and low accuracy ϵ

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

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

sketch size
m=1
uniform row sampling

- Sketched Gradient Descent

$$x_{t+1} = x_t - \mu_t A^T S_t^T S\overset{t}{\cancel{A}}(Ax_t - b)$$

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

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

- step-size $\mu_t = \frac{1}{\|a_{i_t}\|_2^2}$

$$x_{t+1} = x_t - \frac{a_{i_t}^T x_t - b_{i_t}}{\|a_{i_t}\|_2^2} a_i$$

$$= \left(I - \frac{a_{i_t} a_{i_t}^T}{a_{i_t}^T a_{i_t}}\right) x_t + \frac{b_{i_t}}{\|a_{i_t}\|_2^2} \cdot a_i$$

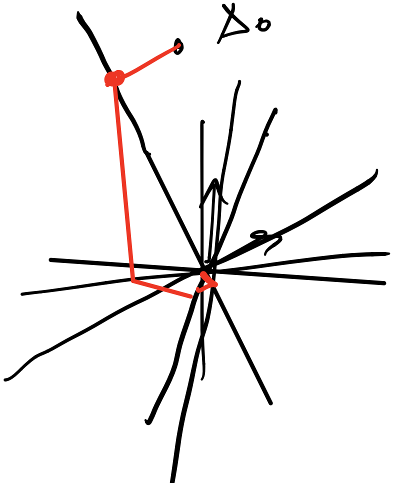
Convergence Analysis

$A(x_t - x^*) = \Delta_t$

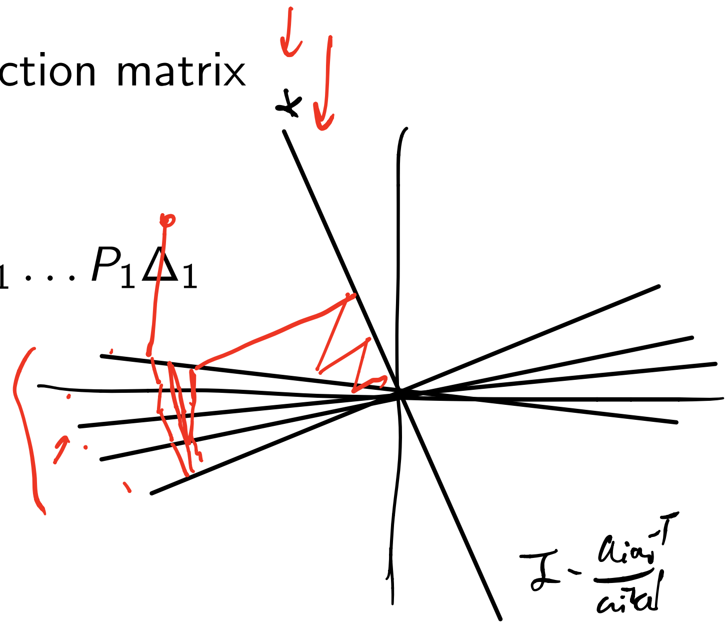
- Assume that $b = Ax^*$ and define $\Delta_t = A(x_t - x^*)$
- $\Delta_{t+1} = \Delta_t - \frac{a_{i_t} a_{i_t}^T}{\|a_{i_t}\|_2^2} \Delta_t = P_t \Delta_t$

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

Δ_0



$$\Delta_T = P_{T-1} \ldots P_1 \Delta_1$$



$I - \frac{a_i a_i^T}{a_i^T a_i}$

Convergence Analysis: General Sampling Distributions

- 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 $g_t = \frac{1}{p_{i_t}}(a_{i_t}^T x_t - b_{i_t})a_i$ $\Rightarrow \mathbb{E}\, g_t = \sum_j \left[\frac{1}{\cancel{p_j}}(a_j^T x_t - b_j)a_j\right]\cancel{p_j}$
  $= \nabla f(x)$
- 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 $\mu_t = 1$

$$\underline{x_{t+1}}^{\,-x^*} = x_t^{\,-x^*} - \frac{1}{p_{i_t}} \overbrace{(a_{i_t}^T x_t - b_{i_t})}^{a_{i_t}^T(x_t - x^*)} a_i$$

- $\Delta_{t+1} = \Delta_t - \frac{a_{i_t} a_{i_t}^T}{p_{i_t}} \Delta_t = \left(I - \frac{a_{i_t} a_{i_t}^T}{p_{i_t}}\right) \cdot \Delta_t$

$$\mathbb{E}\|\Delta_{t+1}\|_2^2 = \mathbb{E}\|\Delta_t - \frac{a_{i_t} a_{i_t}^T}{p_{i_t}} \Delta_t\|_2^2$$

$$= \mathbb{E}\|\Delta_t\|_2^2 - 2\Delta_t^T \frac{a_{i_t} a_{i_t}^T}{p_{i_t}} \Delta_t + \|\frac{a_{i_t} a_{i_t}^T}{p_{i_t}} \Delta_t\|_2^2$$

$$= \mathbb{E}\Delta_t^T \left(I - 2\boxed{\frac{a_{i_t} a_{i_t}^T}{p_{i_t}}} + \underline{\frac{a_{i_t} a_{i_t}^T \|a_{i_t}\|_2^2}{p_{i_t}^2}}\right) \Delta_t$$

$$\sum \frac{a_j a_j^T \|a_j\|^2}{p_j^{\cancel{2}}} \cdot \cancel{p_j}$$

$$\sum \frac{a_j a_j^T}{\cancel{p_j}} \cdot \cancel{p_j} = \sum a_j a_j^T$$

Convergence Analysis: General Sampling Distributions

$\Delta^T Q \Delta \le \lambda_{max}(Q) \cdot \|\Delta\|_2^2$
holds for $Q = Q^T$

- Taking expectations

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

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

$$\Delta_t^T \left(\sum_{i=1}^{n} \frac{a_{i_t} a_{i_t}^T \|a_i\|_2^2}{p_i} \right) \Delta_t \le \lambda_{\max} \left(\sum_{i=1}^{n} \frac{a_{i_t} a_{i_t}^T \|a_i\|_2^2}{p_i} \right) \|\Delta_t\|_2^2$$

tight
there is an x^* that achieves this

Convergence Analysis: General Sampling Distributions

- Taking expectations

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

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

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

PSD

$$\leq \mathbf{Tr} \left(\sum_{i=1}^n \frac{a_{i_t} a_{i_t}^T \|a_i\|_2^2}{p_i} \right) \|\Delta_t\|_2^2$$

$\lambda_{max}(X)$ is convex
but not differentiable

trace

Convergence Analysis: General Sampling Distributions

$\sum_t \mathrm{tr}\ \frac{a_{i_t} a_{i_t}^T \cdot \|a_i\|_2^2}{p_i}$

- minimizing the upper-bound

$$\min_{p \sum_{i=1} p_i = 1, p_i \geq 0} \mathbf{Tr}\left(\sum_{i=1}^{n} \frac{a_{i_t} a_{i_t}^T \|a_i\|_2^2}{p_i}\right)$$

- equivalent to

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

↑ Same form in AMM;

Form Lagrange mult.

Convergence Analysis: General Sampling Distributions

- minimizing the upper-bound

$$\min_{p\ \sum_{i=1} p_i = 1, p_i \geq 0} \mathbf{Tr}\left(\sum_{i=1}^{n} \frac{a_{i_t} a_{i_t}^T \|a_i\|_2^2}{p_i} \right)$$

- equivalent to

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

- 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 $A^T A \sim A^T S^T S A$

Randomized Kaczmarz Algorithm

- optimal sampling distribution

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

- consider step-size μ_t
- $x_{t+1} = x_t - \mu_t \frac{1}{p_{i_t}} a_{i_t}(a_{i_t}^T x - b_{i_t}) = x_t - \mu_t \frac{\|A\|_F^2}{\|a_{i_t}\|_2^2} a_{i_t}(a_{i_t}^T x - b_{i_t})$
- set the step-size $\mu_t = \frac{1}{\|A\|_F^2}$
- this is called **Randomized Kaczmarz Algorithm**
- $x_{t+1} = x_t - \frac{1}{\|a_{i_t}\|_2^2} a_{i_t}(a_{i_t}^T x - b_{i_t})$
- convergence analysis yields

$$\begin{aligned}\Delta_{t+1} &= \left(I - \frac{a_i a_i^T}{\|a_{i_t}\|_2^2}\right)\Delta_t \\ &= P_t \Delta_t\end{aligned}$$

- where $P_t = I - \frac{a_i a_i^T}{\|a_{i_t}\|_2^2}$

Convergence rate

$$
\begin{aligned}
\mathbb{E}\|\Delta_{t+1}\|_2^2 &= \Delta_t^T (I - \frac{1}{\|A\|_F^2} A^T A)\Delta_t \\
&\leq (1 - \frac{\lambda_{\min}}{\|A\|_F^2})\|\Delta_t\|_2^2
\end{aligned}
$$

- recursively applying the above bound and taking conditional expectations

  after T iterations we obtain

$$
\mathbb{E}\|\Delta_T\|_2^2 \leq (1 - \frac{\lambda_{\min}}{\|A\|_F^2})^T \cdot \mathbb{E}\|\Delta_0\|_2^2
$$

$$
\operatorname{tr} A^T A = \sum \lambda_i (A^T A)
$$