THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Statistics and Data Sciences
College of Natural Sciences

SDS 385: Stat Models for Big Data

Lecture 3: GD and SGD cont.

Purnamrita Sarkar
Department of Statistics and Data Science
The University of Texas at Austin

https://psarkar.github.io/teaching

Scalability concerns

- You have to calculate the gradient every iteration.
- Take ridge regression.
- You want to minimize $1/n\left((\boldsymbol{y}-\boldsymbol{X}\beta)^T(\boldsymbol{y}-\boldsymbol{X}\beta)-\lambda\beta^T\beta\right)$
- Take a derivative: $(-2\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{X}\beta)-2\lambda\boldsymbol{\beta})/n$
- Grad descent update takes $\boldsymbol{\beta}_{t+1} \leftarrow \boldsymbol{\beta}_t + \alpha(\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_t)+\lambda\boldsymbol{\beta}_t)$
- What is the complexity?
    - Trick: first compute $\boldsymbol{y}-\boldsymbol{X}\beta$.
    - *np* for matrix vector multiplication, $nnz(\boldsymbol{X})$ for sparse matrix vector multiplication.
    - Remember the examples with humongous n and p?

What will you need for this class

- Stuff you should know from the last lecture.
- The knowledge of conditional expectation.
- Law of total expectation, which is also known as the tower property.

So what to do?

- For $t = 1 : T$
  - Draw σ_t with replacement from n
  - $\beta_{t+1} = \beta_t - \alpha \nabla f(x_{\sigma_t}; \beta_t)$
- In expectation (over the randomness of the index you chose), for a fixed β,

$$E[\nabla f(x_{\sigma_t}; \beta)] = \frac{\sum_i \nabla f(x_i; \beta)}{n}$$

- Does this also converge?

Convergence

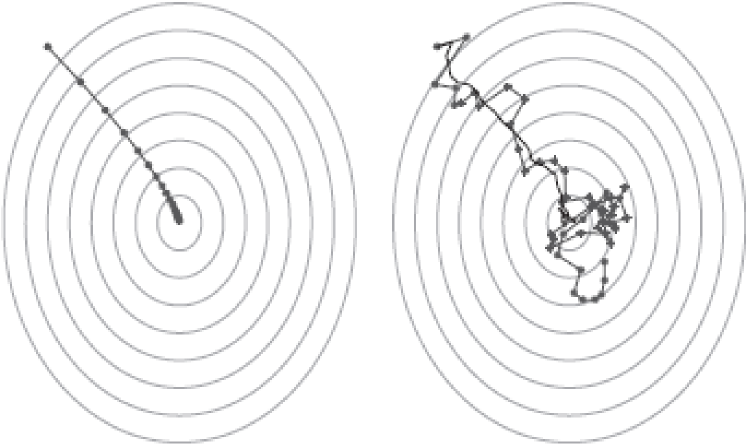

Figure 1: Gradient descent vs Stochastic gradient descent

Convergence

- Let $\nabla f(X; \beta)$ be the full derivative.

$$\beta_{t+1} - \beta^* = \beta_t - \beta^* - \alpha \nabla f(x_{\sigma_t}; \beta_t)$$
$$\|\beta_{t+1} - \beta^*\|^2$$
$$= \|\beta_t - \beta^*\|^2 + \alpha^2 \|\nabla f(x_{\sigma_t}; \beta_t)\|^2 - 2\alpha \langle \nabla f(x_{\sigma_t}; \beta_t), \beta_t - \beta^* \rangle$$

- Take the expectation

$$E[\|\beta_{t+1} - \beta^*\|^2] = E[\|\beta_t - \beta^*\|^2] + \alpha^2 E\|\nabla f(x_{\sigma_t}; \beta_t)\|^2 - 2\alpha E\langle \nabla f(x_{\sigma_t}; \beta_t), \beta_t - \beta^* \rangle$$

Convergence

- Let $\nabla f(X;\beta)$ be the full derivative.
- How do we do expectation of the cross product

$$\begin{aligned} E\langle \nabla f(x_{\sigma_t};\beta_t), \beta_t - \beta^*\rangle &= EE[\langle \nabla f(x_{\sigma_t};\beta_t), \beta_t - \beta^*\rangle | \sigma_1, \ldots, \sigma_{t-1}] \\ &= E\langle \nabla f(X;\beta_t), \beta_t - \beta^*\rangle \end{aligned}$$

Convergence

- Let $\nabla f(X;\beta)$ be the full derivative.
- How do we do expectation of the cross product

$$\begin{aligned} E\langle \nabla f(x_{\sigma_t};\beta_t), \beta_t - \beta^*\rangle &= EE[\langle \nabla f(x_{\sigma_t};\beta_t), \beta_t - \beta^*\rangle | \sigma_1, \ldots, \sigma_{t-1}] \\ &= E\langle \nabla f(X;\beta_t), \beta_t - \beta^*\rangle \end{aligned}$$

- Now we will use strong convexity. Recall:

$$\langle \beta - \beta', \nabla f(X;\beta) - \nabla f(X;\beta')\rangle \geq \mu\|\beta - \beta'\|^2$$

Convergence

- Let $\nabla f(X;\beta)$ be the full derivative.
- How do we do expectation of the cross product

$$\begin{aligned} E\langle \nabla f(x_{\sigma_t};\beta_t), \beta_t - \beta^*\rangle &= EE[\langle \nabla f(x_{\sigma_t};\beta_t), \beta_t - \beta^*\rangle | \sigma_1, \ldots, \sigma_{t-1}] \\ &= E\langle \nabla f(X;\beta_t), \beta_t - \beta^*\rangle \end{aligned}$$

- Now we will use strong convexity. Recall:

$$\langle \beta - \beta', \nabla f(X;\beta) - \nabla f(X;\beta')\rangle \geq \mu \|\beta - \beta'\|^2$$

- Take $\beta = \beta_t$ and $\beta' = \beta^*$:

$$\langle \beta_t - \beta^*, \nabla f(X;\beta_t) - \underbrace{\nabla f(X;\beta^*)}_{0}\rangle \geq \mu \|\beta_t - \beta^*\|^2$$

Convergence

- Let $\nabla f(X;\beta)$ be the full derivative.
- How do we do expectation of the cross product

$$\begin{aligned} E\langle \nabla f(x_{\sigma_t};\beta_t), \beta_t - \beta^*\rangle &= EE[\langle \nabla f(x_{\sigma_t};\beta_t), \beta_t - \beta^*\rangle | \sigma_1, \ldots, \sigma_{t-1}] \\ &= E\langle \nabla f(X;\beta_t), \beta_t - \beta^*\rangle \\ &\geq \mu \|\beta_t - \beta^*\|^2 \end{aligned}$$

Convergence

- $$\begin{aligned} E\|\nabla f(x_{\sigma_t};\beta_t)\|^2 &= EE\left[\|\nabla f(x_{\sigma_t};\beta_t)\|^2 \,\middle|\, \sigma_1,\ldots,\sigma_{t-1}\right] \\ &= \frac{1}{n}\sum_i E\left[\|\nabla f(x_i;\beta_t)\|^2\right] \\ &\leq M \qquad \textbf{We assume this} \end{aligned}$$

SGD cont.

- So by total expectation rule,

$$E[\|\beta_{t+1} - \beta^*\|^2] \leq (1 - 2\alpha\mu)E[\|\beta_t - \beta^*\|^2] + \alpha^2 M$$

- So SGD is converging to a noise ball.
- How to remedy this?

SGD stepsize

- Assume you are far away from the noise ball.
- $\|\beta_t - \beta^*\|^2 \geq \alpha M/\mu$.
- Then,

$$\begin{aligned} E[\|\beta_{t+1} - \beta^*\|^2 | \beta_t] &\leq (1 - 2\alpha\mu)\|\beta_t - \beta^*\|^2 + \alpha\mu\|\beta_t - \beta^*\|^2 \\ &\leq (1 - \alpha\mu)\|\beta_t - \beta^*\|^2 \qquad \text{If } \alpha\mu < 1 \\ E[\|\beta_T - \beta^*\|^2] &\leq e^{-\alpha\mu T} C, \end{aligned}$$

- C is the initial loss
- It takes $1/\alpha\mu \log M$ steps to achieve M factor contraction.

Tradeoff

- Recall that the size of the noise ball is

$$\lim_{t\to\infty} E[\|\beta_{t+1} - \beta^*\|^2] \leq \frac{\alpha M}{2\mu}$$

- So the size is $O(\alpha)$, i.e. for larger α we converge to a larger noise ball.
- But convergence time inversely proportional to step size α.
- So there is a tradeoff.

What if we allow the step size to vary

- We will set the stepsize as $1/t$, and check the following by induction.

Theorem
If we use $\alpha_t = a/(t+1)$, for $a > 1/2\mu$ we have:

$$E[\|\beta_t - \beta_0\|^2] \leq \frac{\max(\|\beta_1 - \beta^*\|^2, Y)}{t+1}$$

where $Y = \dfrac{Ma^2}{2a\mu - 1}$.

Proof.
We will do this by induction. First note Step 1 is obviously true. Now assume that the above holds for t. We will show that it holds for $t+1$. $\square$

What if we allow the step size to vary

- Let $C = \max(\|\beta_1 - \beta^*\|^2, Y)$
- Recall that we have:

$$\begin{aligned} E[\|\beta_{t+1} - \beta^*\|^2] &\leq (1 - 2\alpha_t\mu)E\|\beta_t - \beta^*\|^2 + \alpha_t^2 M \\ &\leq (1 - 2a\mu/(t+1)))\frac{Y}{t+1} + \frac{Ma^2}{(t+1)^2} \\ &= \frac{Y}{t+1} - \frac{a}{(t+1)^2}(2\mu Y - Ma) \end{aligned}$$

- Set $a(2Y\mu - Ma) = Y$, i.e. $Y = \dfrac{Ma^2}{2a\mu - 1}$
- So

$$E[\|\beta_{t+1} - \beta^*\|^2] \leq Y\left(\frac{1}{t+1} - \frac{1}{(t+1)(t+2)}\right) = \frac{Y}{t+2}$$

Mini batch Stochastic Gradient Descent

- SGD uses one data-point at a time.
  - Number of iterations to reach ϵ error is $1/\epsilon$
  - Work per iteration $O(p)$
  - Total work p/ϵ
- GD uses all data-points at a time.
  - Number of iterations to reach ϵ error is $\log(1/\epsilon)$
  - Work per iteration $O(np)$
  - Total work $np \log(1/\epsilon)$

A compromise

,

- Pick B_t without replacement from $\{1, \ldots, n\}$ with $|B_t| = b$
- $\beta_{t+1} = \frac{1}{b} \sum_{i \in B_t} \nabla f(x_i; \beta_t)$
- $b \ll N$

Hope

- Takes b times more time than Stochastic Gradient Descent
- Hopefully converges **sooner**?

Convergence

$$
\begin{aligned}
&\beta_{t+1} - \beta^* \\
&= \beta_t - \beta^* - \alpha \frac{1}{b} \sum_{i \in B_t} \nabla f(x_i; \beta_t) \\
&= \beta_t - \beta^* - \alpha(\nabla f(X; \beta_t) - \nabla f(X; \beta^*)) + \alpha(\nabla f(X; \beta_t) - \nabla f(x_{\sigma_t}; \beta_t)) \\
&= \beta_t - \beta^* - \alpha(\nabla f(X; \beta_t) - \nabla f(X; \beta^*)) - \alpha \left(\frac{1}{b} \sum_{i \in B_t} \nabla f(x_i; \beta_t) - \nabla f(X; \beta_t) \right)
\end{aligned}
$$

Lets look at the variance of

$$
\operatorname{var} \left(\frac{1}{b} \sum_{i \in B_t} \nabla f(x_i; \beta_t) - \nabla f(X; \beta_t) \right)
$$

Variance reduction

- Let $\Delta_i := f(x_i; \beta_t) - \nabla f(X; \beta_t)$
- Let $Y_i \in \{0, 1\}$ be a random variable that denotes whether $i \in B_t$ or not.
- Expectation:

$$E\left[\frac{1}{b}\sum_{i \in B_t} \nabla f(x_i; \beta_t) - \nabla f(X; \beta_t)\right] = E\left[\frac{1}{b}\sum_{i} Y_i \nabla f(x_i; \beta_t) - \nabla f(X; \beta_t)\right] = 0$$

- Let $\Delta_i = \nabla f(x_i; \beta_t) - \nabla f(X; \beta_t)$
- Variance:

$$\begin{aligned} E\left[\frac{1}{b}\sum_{i \in B_t} \nabla f(x_i; \beta_t) - \nabla f(X; \beta_t)\right]^2 &= E\left[\frac{1}{b}\sum_{i} Y_i \Delta_i\right]^2 \\ &= \sum_{ij} \Delta_i \Delta_j E(Y_i Y_j)/b^2 \end{aligned}$$

Variance

- $$\begin{aligned}
\sum_{ij} \Delta_i \Delta_j E(Y_i Y_j) &= \sum_{i \neq j} \frac{b(b-1)}{n(n-1)} \Delta_i \Delta_j + \sum_i \frac{b}{n} \Delta_i^2 \\
&= \frac{b}{n} \left(\frac{b-1}{n-1} \sum_{i \neq j} \Delta_i \Delta_j + \sum_i \Delta_i^2 \right) \\
&= \frac{b}{n} \left(\frac{b-1}{n-1} (\sum_i \Delta_i)^2 + \sum_i \Delta_i^2 (1 - \frac{b-1}{n-1}) \right) \\
&= \frac{b}{n} \sum_i \Delta_i^2 (1 - \frac{b-1}{n-1})
\end{aligned}$$

- So

$$E_{X,B_t} \left[\frac{1}{b} \sum_{i \in B_t} \nabla f(x_i; \beta_t) - \nabla f(X; \beta_t) | \beta_t \right]^2 \leq \sum_i E_X[\Delta_i^2]/bn \leq M/b$$

Acknowledgment

Cho-Jui Hsieh and Christopher De Sa's large scale ML classes.