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?

So what to do?

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

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

- Does this also converge?

Convergence

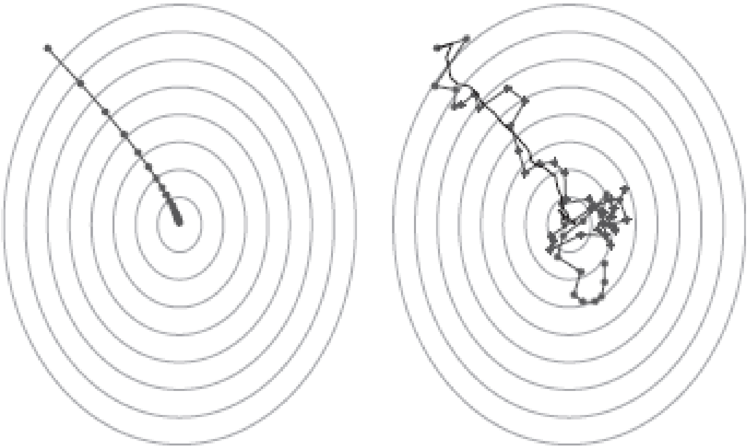

Figure 1: Gradient descent vs Stochastic gradient descent

Convergence

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

$$\begin{aligned}
&\beta_{t+1} - \beta^* \\
&= \beta_t - \beta^* - \alpha \nabla f(x_{\sigma_i};\beta_t) \\
&= \beta_t - \beta^* - \alpha(\nabla f(X;\beta_t) - \nabla f(X;\beta^*)) + \alpha(\nabla f(X;\beta_t) - \nabla f(x_{\sigma_i};\beta_t)) \\
&= \underbrace{(I - \alpha H(z_t))(\beta_t - \beta^*)}_{g(\beta_t)} + \alpha \underbrace{(\nabla f(X;\beta_t) - \nabla f(x_{\sigma_i};\beta_t))}_{h(\sigma_i,\beta_t)}
\end{aligned}$$

- Take the expected squared length:

$$E[\|\beta_{t+1} - \beta^*\|^2 | \beta_t] = \underbrace{\|g(\beta_t)\|^2}_{\text{Same as before}} + \alpha^2 \underbrace{E[\|h(\sigma_i,\beta_t)\|^2 | \beta_t]}_{\substack{\text{variance of} \\ \text{gradient update} \\ \text{at a random point}}}$$

Convergence

-

$$
\begin{aligned}
E[\|h(\sigma_i, \beta_t)\|^2 | \beta_t] &= E_X E_\sigma[\|h(\sigma_i, \beta_t)\|^2 | \beta_t] \\
&= E_X E_\sigma[\|\nabla f(x_{\sigma_i}; \beta_t) - \nabla f(X; \beta_t)\|^2 | \beta_t] \\
&= E_X \frac{1}{n} \sum_i [\|\nabla f(x_i; \beta_t) - \nabla f(X; \beta_t)\|^2 | \beta_t] \\
&= E_X[\|\nabla f(x_i; \beta_t) - \nabla f(X; \beta_t)\|^2 | \beta_t] =: M
\end{aligned}
$$

SGD cont.

- So by total expectation rule,

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

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

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

SGD stepsize

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

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

- C is the initial loss
- It takes $2/\alpha\mu \log M$ steps to achieve M factor contraction.

Tradeoff

- Recall that the size of the noise ball is

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

- So the size is $O(\alpha)$, i.e. for larger α we converge to a larger noise ball.
- But convergence time is $2/\alpha\mu \log M$, i.e. inversely proportional to step size α.
- So there is a tradeoff.

What if we allow the step size to vary

- $\beta_{t+1} = \beta_t - \alpha_t \nabla f(x_i; \beta_t)$
- How do we choose this optimally?
- Recall our bound, and assume $\alpha_t \mu < 1$

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

- Define $d_t := E[\|\beta_{t+1} - \beta^*\|^2]$
- Differentiate and set to zero. This gives,

$$
-\mu d_t + 2\alpha_t M = 0 \rightarrow \alpha_t = \frac{\mu d_t}{2M}
$$

Varying step size

$$
\begin{aligned}
d_{t+1} &\leq (1 - \mu^2 d_t/2M) d_t + \mu^2 d_t^2/4M \\
&= d_t - \mu^2 d_t^2/4M \\
\frac{1}{d_{t+1}} &\geq \frac{1}{d_t} \frac{1}{1 - \mu^2 d_t/4M} \\
&\geq \frac{1}{d_t} \left(1 + \frac{\mu^2 d_t}{4M} \right) \\
&= \frac{1}{d_t} + \frac{\mu^2}{4M}
\end{aligned}
$$

- If you think of $1/d_t$ to be analogous to the accuracy of the score, then this is saying at each iteration the accuracy is increasing by some increment.

Varying step size

- So $\frac{1}{d_T} \geq \frac{1}{d_0} + \frac{\mu^2 T}{4M}$
- Take $\alpha_t = \frac{\mu d_t}{2M} = \frac{\mu \left(\frac{1}{d_0} + \frac{\mu^2 T}{4M} \right)^{-1}}{2M} \approx 1/t$

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

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

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.