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: $(-2\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\beta) - 2\lambda\beta)/n$

- Grad descent update takes $\beta_{t+1} \leftarrow \beta_t + \alpha(\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\beta_t) + \lambda\beta_t)$

- 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?

1

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.

- 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?

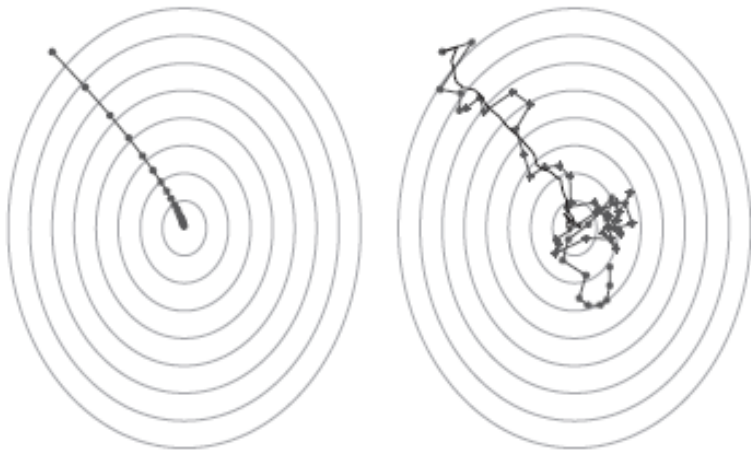**Figure 1:** Gradient descent vs Stochastic gradient descent

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

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

5

Convergence

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

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

Convergence

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

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

- 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

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

- 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

-

$$E\|\nabla f(x_{\sigma_t}; \beta_t)\|^2 = EE\left[\|\nabla f(x_{\sigma_t}; \beta_t)\|^2 \,\middle|\, \sigma_1, \ldots, \sigma_{t-1}\right]$$

$$= \frac{1}{n} \sum_i E\left[\|\nabla f(x_i; \beta_t)\|^2\right]$$

$$\leq M \qquad \textbf{We assume this}$$

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,

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

- 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

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

- 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
$t+1$. □

What if we allow the step size to vary

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

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

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

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

An example

- $\min\limits_{\beta} \dfrac{1}{n} \sum\limits_{i}(x_i - \beta)^2$

- Assume that $\bar{x} = 0$

- SGD update with fixed α is as follows:
  $\beta_1 = \beta_0 + \alpha\alpha(x_{\sigma_t} - \beta_0) = (1-\alpha)\beta_0 + \alpha x_{\sigma_0}$

$$\beta_t = (1-\alpha)^t \beta_0 + \alpha \underbrace{\sum_{i<t}(1-\alpha)^i x_{\sigma_i}}_{\text{Behaves like a } N(0, C_t)}$$

  where $C_t = \dfrac{1 - (1-\alpha)^{2t}}{1 - (1-\alpha)^2}\sigma^2$, where $\sigma^2 = \sum\limits_{i} x_i^2/n$.

- Doesn't go away, unless α goes to zero.

An example - minibatch SGD

- Use average of a batch of size b.

- $\beta_t = (1-\alpha)^t \beta_0 + \alpha \sum_{i<t} (1-\alpha)^i \underbrace{g_i}_{\text{average of B datapoints.}}$

- Let $Y_i = 1$ if $i \in$ minibatch and 0 else.

- If you do with replacement sampling, $P(Y_i = 1) = E[Y_i] = B/n$.

-

$$g_i = \frac{1}{B} \sum_{i=1}^{n} Y_i x_i$$

$$E[g_i^2] = \frac{1}{B^2} \sum_{i=1}^{n} E[Y_i^2] x_i^2 = \frac{1}{B} \frac{1}{n} \sum_i x_i^2 = \frac{\sigma^2}{B}$$

- So, from the previous page, the variance term gets shrunk by B.

An example - averaged SGD

- Use average of SGD updates.

-

$$\beta_t = (1-\alpha)^t \beta_0 + \alpha \sum_{i<t} (1-\alpha)^i x_{\sigma_i}$$

$$\frac{1}{T} \sum_t \beta_t = \frac{1}{T} \sum_{t=1}^{T} (1-\alpha)^t \beta_0 + \alpha \frac{1}{T} \sum_{t=1}^{T} \sum_{i<t} (1-\alpha)^i x_{\sigma_i}$$

$$= \underbrace{\frac{1-\alpha}{T} \frac{1-(1-\alpha)^T}{\alpha} \beta_0}_{\text{Goes to zero}} + \alpha \underbrace{\frac{1}{T} \sum_{j=0}^{T-t-1} (1-\alpha)^j x_{\sigma_t}}_{\text{Behaves like } N(0, D/t)}$$

- So, by averaging, and not using diminishing step size, you basically are reducing the noise ball size and converging to the truth.

Acknowledgment

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