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

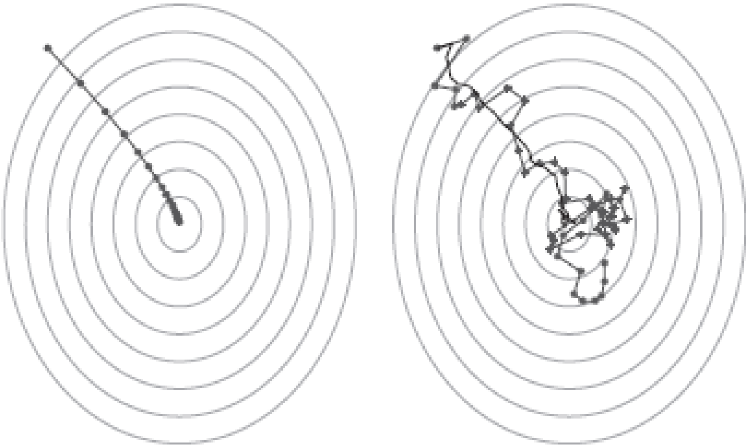

Figure 1: Gradient descent vs Stochastic gradient descent

Convergence

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

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

- Take the expectation

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

Convergence

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

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

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,

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

- 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

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

An example

- $\min_{\beta} \frac{1}{n} \sum_{i} (x_i - \beta)^2$
- Assume that $\bar{x} = 0$
- SGD update with fixed α is as follows:

  $\beta_1 = \beta_0 + \alpha(x_{\sigma_t} - \beta_0) = (1-\alpha)\beta_0 + \alpha x_{\sigma_0}$

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

  where $C_t = \frac{1-(1-\alpha)^{2t}}{1-(1-\alpha)^2}\sigma^2$, where $\sigma^2 = \sum_i x_i^2 / n$.

- 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 without replacement sampling, $P(Y_i = 1) = E[Y_i] = B/n$.
-

$$g_i = \frac{1}{B} \sum_{j=1}^{n} Y_j x_j$$

$$E[g_i^2] \approx \frac{1}{B^2} \sum_{j=1}^{n} E[Y_j^2] x_i^2 = \frac{1}{B} \frac{1}{n} \sum_j x_j^2 = \frac{\sigma^2}{B}$$

- Can you make the above step rigorous?
- So, from the previous page, the variance term gets shrunk by B.

An example - averaged SGD

- Use average of SGD updates.
-

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

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

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

- So, by averaging, you basically are reducing the noise ball size and converging to the truth.

Averaging

- Averaging for more general quadratic functions have the same behavior.
- For general strongly convex functions, we can't have "constant" step sizes, but we **can** have much larger stepsizes – $t^{-\alpha}$ for $\alpha \in (1/2, 1)$.
  - Compare this to $1/t$ for SGD.
- One can do statistical inference with averaging, since we know that the averaged vector converges to a normal ball of a certain variance. If you can estimate this variance, then, you can give confidence intervals for your parameter of interest, not just point estimates.

Final thoughts

- As it turns out, the stepsize for SGD is "optimal"
  - For strongly convex function minimization, no algorithm making T noisy gradient computations will have accuracy better than c/T.

	Error in T iterations	computation per iter	stepsize
GD	$\exp(-cT)$	O(n)	Fixed
SGD	$1/T$	O(1)	$\alpha_t = 1/t$
batch SGD	$1/TB$	O(B)	$\alpha_t = 1/t$
average SGD	$1/T$	O(1)	$\alpha_t = 1/t^{\alpha}, \alpha \in (0.5, 1)$

- Typically you use 32 as batchsize as default.