CSE 547: Machine Learning for Big Data
Homework 4

Academic Integrity We take academic integrity extremely seriously. We strongly encourage students to form study groups. Students may discuss and work on homework problems in groups. However, each student must write down the solutions and the code independently. In addition, each student should write down the set of people whom they interacted with.

Discussion Group (People with whom you discussed ideas used in your answers):

- *Hwai-Jin Peng*

On-line or hardcopy documents used as part of your answers:

- CSE 547 Lecture Slide (Tim Althoff, UWashington)

I acknowledge and accept the Academic Integrity clause.

(Signed) ***Cheng-Yen Yang***

Answer to Question 1

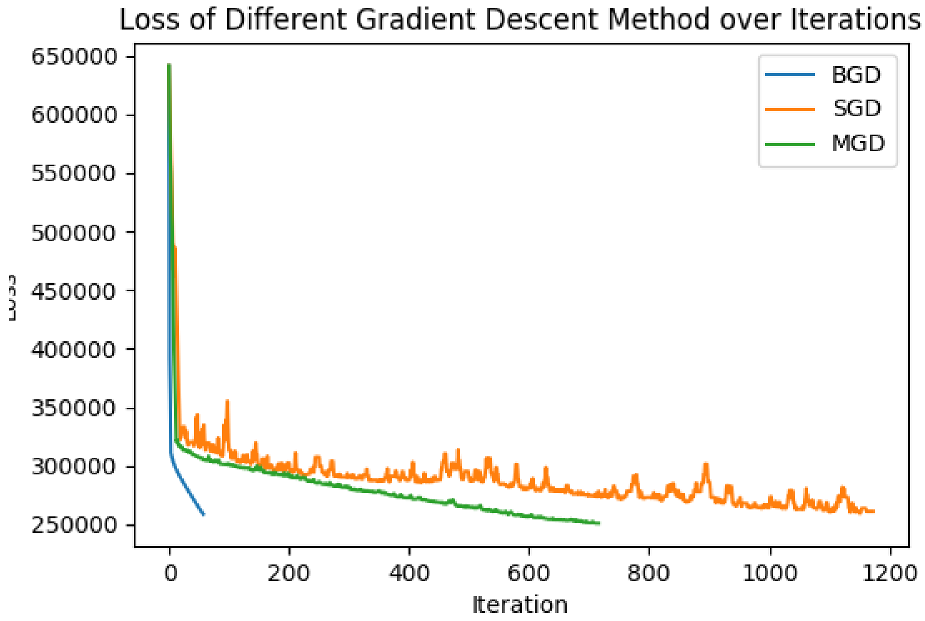

Figure 1: Loss vs Iterations for different gradient descent method

- Batch Gradient Descent takes 57 iterations to converge in 223.48 seconds (3.92sec/iter).
- Stochastic Gradient Descent takes 1173 iterations to converge in 1.29 seconds (0.001sec/iter)
- Minibatch Gradient Descent takes 716 iterations to converge in 9.69 seconds (0.014sec/iter)

Batch Gradient Descent takes fewer updates and is more stable in terms of loss, but is slow as it calculates the loss of the entire dataset per update. Stochastic Gradient Descent updates for every individual sample so it is more unstable, but is very fast in tern of execution time. Minibatch Gradient Descent combine the ideas from Batch Gradient Descent and Stochastic Gradient Descent therefore have a more robust performance in comparison to Stochastic Gradient Descen and a faster update time per iteration in comparison to Batch Gradient Descent.

Answer to Question 2(a)

The algorithm has two parameters δ and ϵ, and $\lceil \ln \frac{1}{\delta} \rceil$ independent hash functions:

$$h_j : \{1, 2, \dots, n\} \rightarrow \{1, 2, \dots, \lceil \frac{e}{\epsilon} \rceil\}, \tag{1}$$

and for each bucket b of each hash function j, the algorithm has a counter, summing up to a total of $\lceil \frac{e}{\epsilon} \rceil \times \lceil \ln \frac{1}{\delta} \rceil$ buckets/counters. Therefore the memory usage of this algorithm in Big-$\mathcal{O}$ notation is $\Theta(\frac{1}{\epsilon} \ln \frac{1}{\delta})$.

Answer to Question 2(b)

Define the variable $X_{i,j}$ to be:

$$X_{i,j} = \sum_{k=1}^{n} I_{i,k}^{(i)} F[k] \tag{2}$$

with $I_{i,k}^{j}$ being the indicator variable which is set to 1 when hash collisions happened between i and k on h_j and 0 otherwise. As we obtain:

$$\mathbb{E}[I_{i,k}^{(i)}] = \mathbb{P}[h_j(i) = h_j(k)] \leq \frac{1}{|h_j|} = \frac{\epsilon}{e} > 0 \tag{3}$$

Therefore $X_{p,q}$ is a non-negative variable. Then we have:

$$\tilde{F}[i] = min_j\{c_{j,h_j(i)}\} = F[i] + X_{i,j} \geq F[i] \tag{4}$$

Answer to Question 2(c)

For any $1 \leq i \leq n$ and $1 \leq j \leq \lceil \ln \frac{1}{\delta} \rceil$, by defining the variable $X_{i,j}$ to be:

$$X_{i,j} = \sum_{k=1}^{n} I_{i,k}^{(j)} F[k] \tag{5}$$

with $I_{i,k}^{(j)}$ being the indicator variable which is set to 1 when hash collisions happened between i and k on h_j and 0 otherwise. We have:

$$\begin{align}
\mathbb{E}[c_{j,h_j(i)}] &= \mathbb{E}\big[F[i] + X_{i,j}\big] \tag{6}\\
&= F[i] + \mathbb{E}[X_{i,j}] \tag{7}\\
&= F[i] + \mathbb{E}\left[\sum_{k=1}^{n} I_{i,k}^{(j)} F[k]\right] \tag{8}\\
&\leq F[i] + \sum_{k=1}^{n} F[k]\mathbb{E}\big[I_{i,k}^{(j)}\big] \tag{9}\\
&\leq F[i] + \sum_{k=1}^{n} F[k]\mathbb{P}\big[h_j(i) = h_j(k)\big] \tag{10}\\
&= F[i] + \frac{\epsilon}{e}\sum_{k=1}^{n} F[k] \tag{11}
\end{align}$$

and as mentioned, t is the length of the entire stream where $\sum_{k=1}^{n} F[k] = t$, therefore:

$$\mathbb{E}[c_{j,h_j(i)}] \leq F[i] + \frac{\epsilon}{e}\sum_{k=1}^{n} F[k] = F[i] + \frac{\epsilon}{e}t \tag{12}$$

Answer to Question 2(d)

From the problem statement:

$$\begin{aligned}
\mathbb{P}\Big[\tilde{F}[i] \leq F[i] + \epsilon t\Big] &= 1 - \mathbb{P}\Big[\tilde{F}[i] \geq F[i] + \epsilon t\Big] && (13)\\
&= 1 - \mathbb{P}\Big[min_j\{c_{j,h_j(i)}\} \geq F[i] + \epsilon t\Big] && (14)\\
&= 1 - \mathbb{P}\Big[c_{j,h_j(i)} \geq F[i] + \epsilon t,\, \forall 1 \leq j \leq \lceil \ln \frac{1}{\delta} \rceil\Big] && (15)\\
&= 1 - \prod_{j=1}^{\lceil \ln \frac{1}{\delta} \rceil} \mathbb{P}\Big[c_{j,h_j(i)} \geq F[i] + \epsilon t\Big] && (16)
\end{aligned}$$

By Markov's inequality:

$$\mathbb{P}\Big[c_{j,h_j(i)} \geq F[i] + \epsilon t\Big] \leq \frac{\mathbb{E}\Big[c_{j,h_j(i)} - F[i]\Big]}{\epsilon t} \tag{17}$$

and property from part 2(c):

$$\mathbb{E}[c_{j,h_j(i)}] \leq F[i] + \frac{\epsilon}{e} t \tag{18}$$

we have:

$$\mathbb{P}\Big[c_{j,h_j(i)} \geq F[i] + \epsilon t\Big] \leq \frac{\frac{\epsilon}{e} t}{\epsilon t} = \frac{1}{e} \tag{19}$$

Therefore we substitute the value back into Eq.(16):

$$\begin{aligned}
\mathbb{P}\Big[\tilde{F}[i] \leq F[i] + \epsilon t\Big] &= 1 - \prod_{j=1}^{\lceil \ln \frac{1}{\delta} \rceil} \mathbb{P}\Big[c_{j,h_j(i)} \geq F[i] + \epsilon t\Big] && (20)\\
&\geq 1 - \prod_{j=1}^{\lceil \ln \frac{1}{\delta} \rceil} \frac{1}{e} && (21)\\
&= 1 - \Big(\frac{1}{e}\Big)^{\lceil \ln \frac{1}{\delta} \rceil} && (22)\\
&\geq 1 - \delta && (23)
\end{aligned}$$

Answer to Question 2(e)

From the plot we can see that word with a frequency greater than 10^{-6} have a relative error below 1.

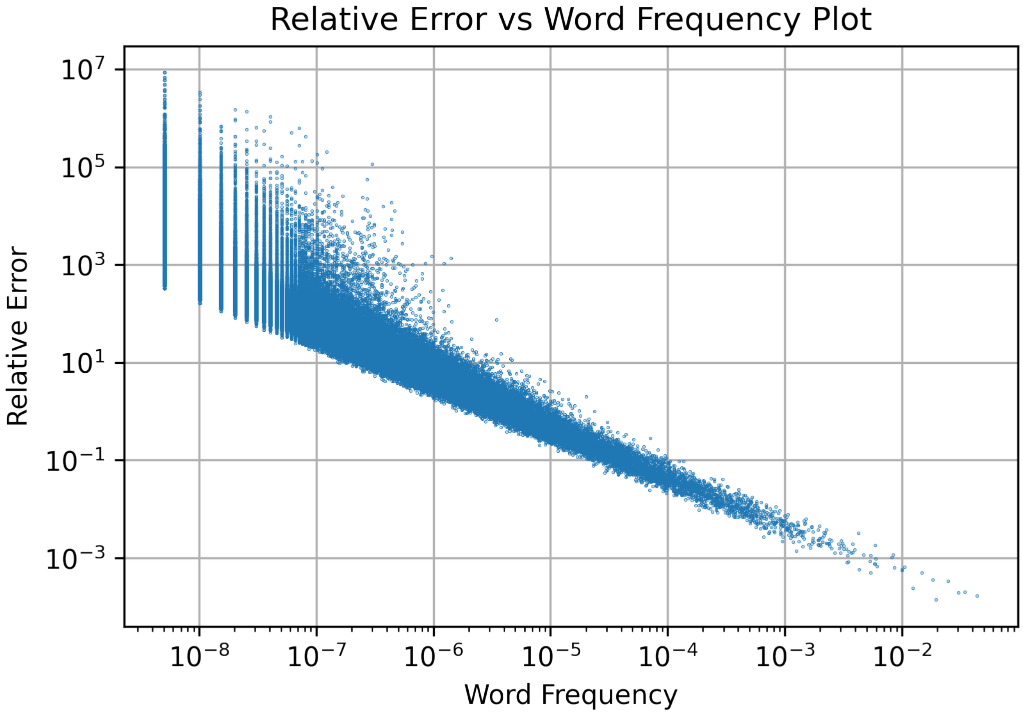

Figure 2: Relatvie error vs word frequency plot with $n_{hash} = 5$, $n_{buckets} = 10000$

Answer to Question 3(a)

By the estimate $X = Z^2$:

$$\mathbb{E}_h[X] = \mathbb{E}_h[Z^2] = \mathbb{E}_h\left[\left(\sum_{j=1}^{n} h(j)F(j)\right)^2\right] \tag{24}$$

$$= \mathbb{E}_h\left[\sum_{i,j=1}^{n} h(i)h(j)F(i)F(j)\right] \tag{25}$$

$$= \mathbb{E}_h\left[\sum_{j=1}^{n} h(j)^2F(j)^2\right] + \mathbb{E}_h\left[\sum_{i \neq j}^{n} h(i)h(j)F(i)F(j)\right] \tag{26}$$

with the given function $h[n] \Rightarrow \{\pm 1\}$, we have $h(j)^2 = 1$ and $\sum_{i<j}^{n} h(i)h(j)F(i)F(j) = 0$ due to the cancellation of the pairs, then we have:

$$\mathbb{E}_h[X] = \sum_{j=1}^{n} F(j)^2 = M \tag{27}$$

Answer to Question 3(b)

By definition of variance:

$$Var(X) = \mathbb{E}_h[X^2] - \mathbb{E}_h[X]^2 \tag{28}$$

Then by the estimate $X^2 = Z^4$:

$$\begin{aligned}
\mathbb{E}_h[X^2] = \mathbb{E}_h[Z^4] =& \mathbb{E}_h\left[\left(\sum_{j=1}^{n} h(j)F(j)\right)^4\right] && (29)\\
=& \mathbb{E}_h\left[\sum_{i,j,k,l=1}^{n} h(i)h(j)h(k)h(l)F(i)F(j)F(k)F(l)\right] && (30)\\
=& \mathbb{E}_h\left[\sum_{j=1}^{n} h(j)^4F(j)^4\right] + \mathbb{E}_h\left[\sum_{i\neq j}^{n} h(i)^2h(j)^2F(i)^2F(j)^2\right] && (31)\\
&+ \mathbb{E}_h\left[\sum_{i\neq j}^{n} h(i)h(j)^3F(i)F(j)^3\right] && (32)\\
&+ \mathbb{E}_h\left[\sum_{i\neq j\neq k}^{n} h(i)^2h(j)h(k)F(i)^2F(j)F(k)\right] && (33)\\
&+ \mathbb{E}_h\left[\sum_{i\neq j\neq k\neq l}^{n} h(i)h(j)h(k)h(l)F(i)F(j)F(k)F(l)\right] && (34)
\end{aligned}$$

similar to 3(a), we observe that terms in Ep. (9), Eq. (10) and Eq. (11) will have the value of 0 due to the cancellation of the pairs. Therefore we have:

$$\begin{aligned}
\mathbb{E}_h[X^2] &= \mathbb{E}_h\left[\sum_{j=1}^{n} h(j)^4F(j)^4\right] + \mathbb{E}_h\left[\sum_{i\neq j}^{n} h(i)^2h(j)^2F(i)^2F(j)^2\right] && (35)\\
&= \sum_{j=1}^{n} F(j)^4 + \binom{4}{2}\cdot\sum_{i\neq j}^{n} F(i)^2F(j)^2 && (36)\\
&\leq 3\cdot\sum_{j=1}^{n} F(j)^4 + \binom{4}{2}\cdot\sum_{i\neq j}^{n} F(i)^2F(j)^2 && (37)\\
&= 3\cdot\left(\sum_{j=1}^{n} F(j)^2\right)^2 = 3M^2 && (38)
\end{aligned}$$

As we already obtained $\mathbb{E}_h[X] = M$ in 3(a), we derive

$$Var(X) = \mathbb{E}_h[X^2] - \mathbb{E}_h[X]^2 = 3M^2 - M^2 = 2M^2 \leq 4M^2 \tag{39}$$