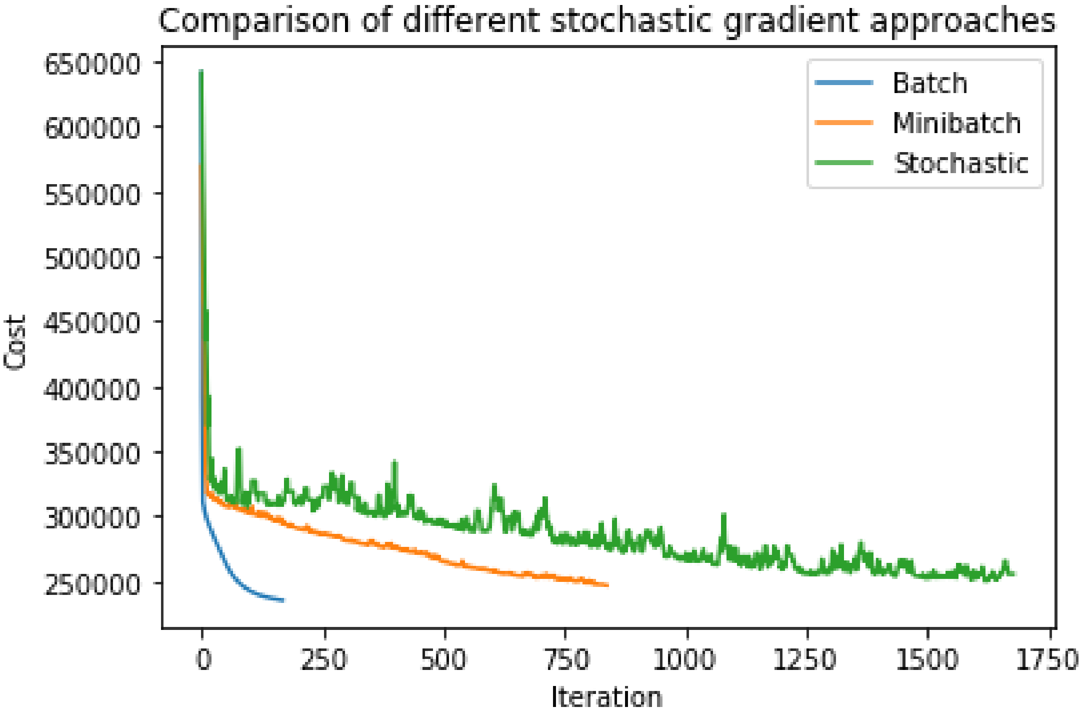
CSE547: Machine Learning for Big Data
Homework 4

Answer to Question 1(a)

The obtained convergence times are

- Batch Gradient: 1.12s
- Stochastic Gradient: 11.07s
- Minibatch Gradient: 5.16s

From Figure and the convergence times, we can see that batch gradient converges the fastest (both in terms of number of iterations as well as convergence time) and also converges to a cost that is lower than for the other two methods. The stochastic approach takes the longest and also converges to some local minimum that is worse than for batch gradient. In contrast to batch gradient, the cost function is not monotonic, but experiences a few upward jumps which might be caused by samples that are not representative of the actual underlying distribution, distorting our weights and offset. Minibatch is somewhere in between, both in terms of convergence time as well as final cost achieved. Note that minibatch is also not monotonically decreasing, but its jumps are a lot less extreme.



Answer to Question 2(a)

(i) Values of G:

- Ice Cream: G = 0
- Tea: G = 0.5
- Hiking: G = 5.36

(ii) Since we want to use the attribute that maximizes the gain G, i would pick the attribute hiking to split the data at the root.

Answer to Question 2(b)

(iii) The decision tree would split at the root using the attribute a_1 and from then on split at attributes based on maximum gain. If we would want to prevent overfitting, there would be no more decision branches after the initial split, since after the initial split at a_1 we are 99% confident about our label. If we keep branching out, we will end up either classifying nodes as '-' on the '+' side of a_1 or nodes as '+' on the '-' side of a_1, which with 99% probability is wrong (probability based on training data, of course).

Answer to Question 3(a)

The memory usage is not dependent on the number of datapoints, but rather on the number of hash functions and number of buckets per hash function, in other word the memory used will be $\lceil \log(1/\delta) \rceil \lceil e/\epsilon \rceil$. So the memory usage is in $\mathcal{O}(\frac{\log(1/\delta)}{\epsilon})$.

Answer to Question 3(b)

For any item i, each occurence of the item in the stream will be hashed to the same bucket $c_{j,h(j)}$ for each hash function j. Hence $F[i] \leq c_{j,h(j)}$ for all j. Then in particular, $F[i] \leq \min_j c_{j,h(j)} = \tilde{F}[i]$, which was to show.

Answer to Question 3(c)

Let j be arbitrary. We assume that the hash function h_j uniformly distributes the items across buckets, meaning the probability of some item $i \in \{1, \ldots, n\}$ to end up in a specific bucket is $\frac{1}{\lceil \frac{e}{\epsilon} \rceil}$.

$$\begin{aligned}
\mathbb{E}[c_{j,h_j(i)}] &= \mathbb{E}[F[i] + \sum_{\substack{k \leq n \\ k \neq i \\ h_j(k)=h_j(i)}} F[k]] \leq \mathbb{E}[F[i] + \sum_{\substack{k \leq n \\ h_j(k)=h_j(i)}} F[k]] \\
&= \mathbb{E}[F[i] + \sum_{k \leq t} \mathbb{1}_{\{h_j(a_k)=h_j(i)\}}] \\
&= F[i] + \sum_{k \leq t} \mathbb{E}[\mathbb{1}_{\{h_j(a_k)=h_j(i)\}}] \\
&= F[i] + t\mathbb{E}[\mathbb{1}_{\{h_j(a_1)=h_j(i)\}}] \\
&= F[i] + t\frac{1}{\lceil \frac{e}{\epsilon} \rceil} \leq F[i] + t\frac{1}{\frac{e}{\epsilon}} = F[i] + t\frac{\epsilon}{e}
\end{aligned}$$

Answer to Question 3(d)

We use the fact that the hash functions are independent and identicallty distributed.

$$\begin{aligned}
P[\tilde{F}[i] \leq F[i] + \epsilon t] &= 1 - P[\tilde{F}[i] \geq F[i] + \epsilon t] = 1 - P[\tilde{F}[i] - F[i] \geq \epsilon t] \\
&= 1 - P[\min_j c_{j,h_j(i)} - F[i] \geq \epsilon t] \\
&= 1 - P[c_{j,h_j(i)} - F[i] \geq \epsilon t \ \forall j \in \{1, \dots, \lceil \log(1/\delta) \rceil\}] \\
&= 1 - \prod_j^{\lceil \log \frac{1}{\delta} \rceil} P[c_{j,h_j(i)} - F[i] \geq \epsilon t] \\
&\overset{(1)}{\geq} 1 - \prod_j^{\lceil \log \frac{1}{\delta} \rceil} \frac{\mathbb{E}[c_{j,h_j(i)} - F[i]]}{\epsilon t} \\
&= 1 - \left(\frac{\mathbb{E}[c_{1,h_1(i)} - F[i]]}{\epsilon t} \right)^{\lceil \log(1/\delta) \rceil} \\
&= 1 - \left(\frac{\mathbb{E}[c_{1,h_1(i)}] - F[i]}{\epsilon t} \right)^{\lceil \log(1/\delta) \rceil} \\
&\geq 1 - \left(\frac{F[i] + \frac{\epsilon}{e} t - F[i]}{\epsilon t} \right)^{\lceil \log(1/\delta) \rceil} \\
&= 1 - \left(\frac{\frac{\epsilon}{e} t}{\epsilon t} \right)^{\lceil \log(1/\delta) \rceil} \\
&= 1 - \left(\frac{1}{e} \right)^{\lceil \log(1/\delta) \rceil} \\
&\overset{(2)}{\geq} 1 - \left(\frac{1}{e} \right)^{\log(1/\delta)} \\
&= 1 - (e)^{\log(\delta)} \\
&= 1 - \delta‘
\end{aligned}$$

where we assumed in (2) that $\log(1/\delta) \geq 0$ and used the fact that $\frac{1}{e}$ is monotonically decreasing and used the result of 3a), namely the fact that $c_{j,h_j(i)} - F[i] \geq 0$ as justification for applying Markov's inequality in (1) (non-negative random variable).

Answer to Question 3(e)

From Figure 1, we can eyeball that a relative error below 1 is achieved from a word frequency of around $1.6 \cdot 10^{-6}$

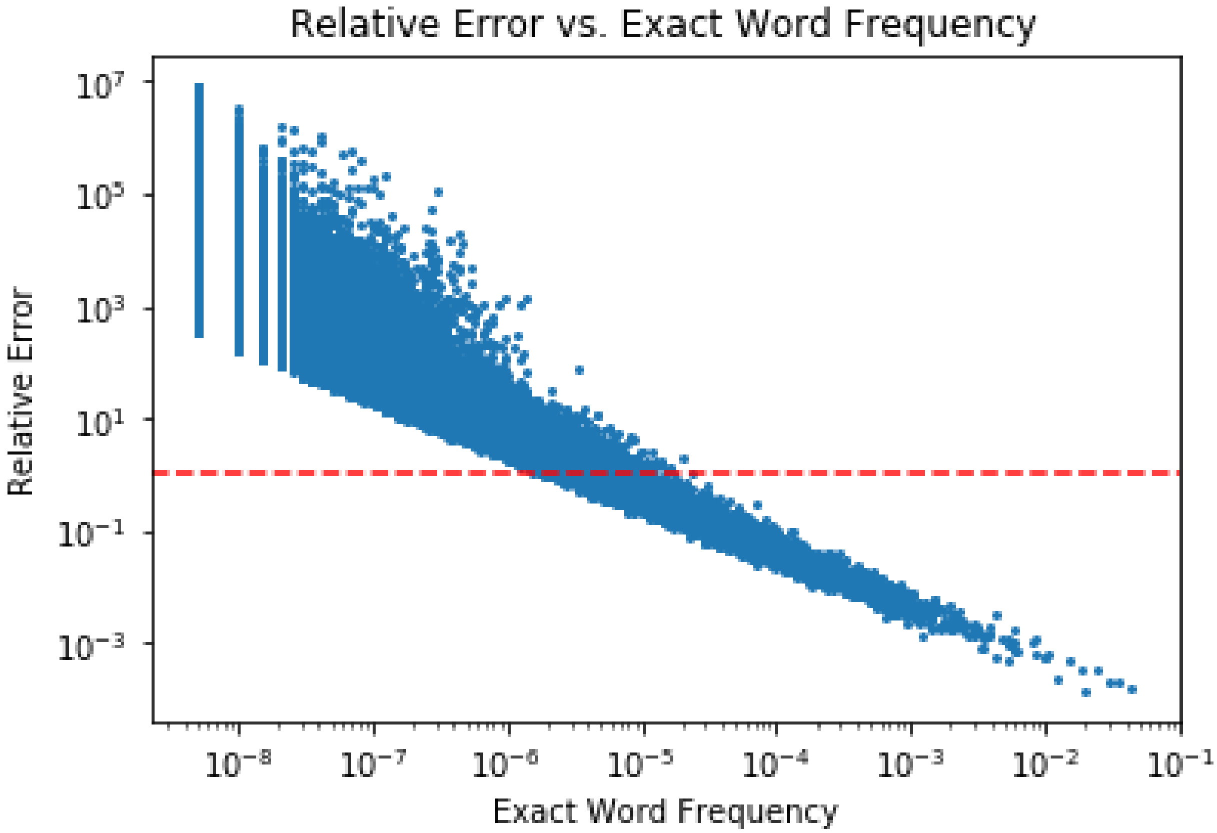


Figure 1: The red line indicates where the relative error is 1.

Answer to Question 4(a)

We assume that p(h(i)=-1)=p(h(i)=1)=.5 f.a. $i \leq n$, i.e. the probability of an item i being associated with a given sign is the probability of a coin flip (that is, before the sign decision for a given item is made, since we need to be consistent across occurrences of the same item. This means that $\mathbb{E}[h(j)] = .5 - .5 = 0$. Further note that for arbitrary $i, j \leq n$,

$$\begin{aligned}\mathbb{E}[h(j)h(i)] =& p(h(i) = 1)p(h(j) = 1) - p(h(i) = 1)p(h(j) = -1) \\ & - p(h(i) = -1)p(h(j) = 1) + p(h(i) = -1)p(h(j) = -1) \\ =& .25 - .5 + .25 = 0 \end{aligned} \tag{1}$$

$$\begin{aligned}\mathbb{E}[X] = \mathbb{E}[Z^2] = \mathbb{E}\left[\left(\sum_{i=1}^{n} h(i)F[i]\right)^2\right] &= \mathbb{E}\left[\sum_{i=1}^{n} h(i)^2 F[i]^2 + 2\sum_{i=1}^{n}\sum_{j\neq i} h(j)F[j]h(i)F[i]\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{n} F[i]^2\right] + 2\left[\sum_{i=1}^{n}\sum_{j\neq i} h(j)h(i)F[j]F[i]\right] \\ &= \sum_{i=1}^{n} F[i]^2 + 2F[j]F[i]\sum_{i=1}^{n}\sum_{j\neq i} E[h(j)h(i)] \\ &\overset{(1)}{=} M + 2F[j]F[i]\sum_{i=1}^{n}\sum_{j\neq i} 0 = M \end{aligned}$$

Answer to Question 4(b)

For reasons of readability, define $h(j)F(j) := \alpha_j$. Then

$$\begin{aligned}
\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[Z^4] - M^2 &= \mathbb{E}\left[\left(\sum_{i=1}^{n} \alpha_i\right)^4\right] - M^2 \\
&= \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} \alpha_i\alpha_j\alpha_k\alpha_l\right] - M^2 \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} \mathbb{E}\left[\alpha_i\alpha_j\alpha_k\alpha_l\right] - M^2 \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} F[i]F[j]F[k]F[l]\mathbb{E}[h(i)h(j)h(k)h(l)] - M^2
\end{aligned}$$

Now note that three different combinations of $m := h(i)h(l)h(k)h(l)$ can occur:
Case I: $i = j = k = l$

$$\mathbb{E}[h(i)h(j)h(k)h(l)] = \mathbb{E}[h(i)^4] = \mathbb{E}[1] = 1$$

Case II: $\exists g \in \{i, j, k, l\} : g \neq f$ for all $f \in \{i, j, k, l\} \setminus \{g\}$

$$\mathbb{E}[h(i)h(j)h(k)h(l)] = \mathbb{E}[h(g)h(f)^3] = \mathbb{E}[h(g)h(f)] = 0$$

where we used Equation 1 from the previous exercise.
Case III: $a = b \neq c = d$ where a, b, c, d are assigned to i, j, k, l by an arbitrary bijective mapping (in other words, in this case, each two of i, j, k, l are the same).

$$\mathbb{E}[h(i)h(j)h(k)h(l)] = \mathbb{E}[h(a)^2h(b)^2] = \mathbb{E}[1] = 1$$

Note that there are exactly three such cases: (i) $i = j \neq k = l$, (ii) $i = k \neq j = l$, (iii) $i = l \neq j = k$. Putting it all together, we can rewrite our variance as follows:

$$\begin{aligned}
Var(X) &= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} F[i]F[j]F[k]F[l]\mathbb{E}[h(i)h(j)h(k)h(l)] - M^2 \\
&= \sum_{i=j=k=l} F[i]^4 + \sum_{i=1}^{n}\sum_{k=1}^{n} F[i]^2F[k]^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} F[i]^2F[j]^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} F[i]^2F[k]^2 - M^2 \\
&= \sum_{i}^{n} F[i]^4 + 3\sum_{i=1}^{n}\sum_{j=1}^{n} F[i]^2F[j]^2 - M^2 \\
&\leq \sum_{i}\sum_{j} F[i]^2F[j]^2 + 3M^2 - M^2 \\
&= 3M^2 \leq 4M^2
\end{aligned}$$

We used the fact that $M^2 = \left(\sum_i F[i]^2\right)^2 = \sum_i \sum_j F[i]^2F[j]^2$

Discussion Group: Romain Camilleri, Andrew Wagenmaker, Nick Nuechterlein, Erin Wilson.

On-line or hardcopy documents used: None

I acknowledge and accept the Academic Integrity clause
Emil Azadian