1. **Minibatching by Sampling Without Replacement.**

Algorithm 1: Minibatched Stochastic Gradient Descent Sampled with Replacement

input: learning rate α, minibatch size B, initial parameters w_0, total number of iterations T
for $t = 0$ *to* $T-1$ **do**
  for $b = 1$ *to* B **do**
    sample $\tilde{i}_{b,t}$ independently and uniformly from $\{1, ..., n\}$
  end
  update model $w_t \leftarrow w_t - \alpha_t \frac{1}{B} \sum_{b=1}^{B} \nabla f_{\tilde{i}_{b,t}}(w_t)$.
end

We can tell this algorithm uses sampling with replacement because it is possible that the same example might be used twice (or more) in the same minibatch. We derived the mean squared error of the minibatch estimator is equal to the mean squared error of an individual gradient sample divided by the batch size, B. That is,

$$\mathbb{E}\left[\left\|\frac{1}{B}\sum_{i\in B}\nabla f_i(w) - \nabla f(w)\right\|^2\right] = \frac{1}{B}\cdot\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(w) - \nabla f(w)\|^2.$$

Thus, by increasing B, we can decrease this variance, and so improve the accuracy of SGD. Recall that for a sequence of pairwise uncorrelated random variable $\{X_i\}_{i=1}^{B}$ that $Var(\sum_i X_i) = \sum_i Var(X_i)$. As an alternative to this, we could sample *without replacement*, in which training examples are never reused in a single minibatch. This corresponds to the following algorithm.

Algorithm 2: Minibatched Stochastic Gradient Descent Sampled without Replacement

input: learning rate α, minibatch size B, initial parameters w_0, total number of iterations T
for $t = 0$ *to* $T-1$ **do**
  sample $\mathcal{B}$ uniformly at random from the set of subsets of $\{1, ..., n\}$ of size B
  update model $w_{t+1} \leftarrow w_t - \alpha_t \cdot \frac{1}{B} \sum_{i\in\mathcal{B}} \nabla f_i(w_t)$
end

(a) Consider some particular weight vector w, and suppose that the mean squared error (variance) of a single gradient sample at w is σ^2. That is,

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(w) - \nabla f(w)\|^2$$

Find an expression for the variance of the without-replacement minibatch estimator in terms of B, n, and σ^2. Note that explicitly, the variance in this case is defined by the expression

$$\mathbb{E}\left[\left\|\frac{1}{B}\sum_{i\in\mathcal{B}}\nabla f_i(w) - \nabla f(w)\right\|^2\right]$$

or, equivalently,

$$\binom{n}{B}^{-1} \sum_{\mathcal{B} \subseteq \{1,\ldots,n\}, |\mathcal{B}|=B} \left\| \frac{1}{B} \sum_{i \in B} \nabla f_i(w) - \nabla f(w) \right\|^2.$$

Is the variance higher or lower than the variance that we got from with-replacement mini batching? How do the two expressions compare asymptotically as the size of the training set n approaches infinity while the minibatch size B remains fixed?

Solution.

$$\mathbb{E}\left[\left\| \frac{1}{B} \sum_{i \in \mathcal{B}} \nabla f_i(w) - \nabla f(w) \right\|^2 \right] = \binom{n}{B}^{-1} \sum_{\mathcal{B} \subseteq \{1,\ldots,n\}, |B|=B} \left\| \frac{1}{B} \sum_{i \in \mathcal{B}} \nabla f_i(w) - \nabla f(w) \right\|^2$$

$$= \frac{1}{B^2} \binom{n}{B}^{-1} \sum_{\mathcal{B}} \left\| \sum_{i \in B} \nabla f_i(w) - \nabla f(w) \right\|^2$$

$$= \frac{1}{B^2} \binom{n}{B}^{-1} \sum_{\mathcal{B}} \left(\left(\sum_{i \in B} \nabla f_i(w) - \nabla f(w) \right)^T \left(\sum_{i \in B} \nabla f_i(w) - \nabla f(w) \right) \right)$$

$$= \frac{1}{B^2} \binom{n}{B}^{-1} \sum_{\mathcal{B}} \left(\sum_{i \in B} \sum_{j \in B} \left(\nabla f_i(w) - \nabla f(w) \right)^T \left(\nabla f_j(w) - \nabla f(w) \right) \right)$$

From here we now realize that if we fix *any* two indices i, j, such that $i \neq j$, then there are $\binom{n-2}{B-2}$ subsets $\mathcal{B}$ which contain indices i, j. Similarly, if we fix indices i, j, such that $i = j$, then there are $\binom{n-1}{B-1}$ subsets $\mathcal{B}$ which contain indices i, j. We can thus eliminate the outer sum based on the two conditions - (1) $i \neq j$ and (2) when $i = j$.

$$\frac{1}{B^2} \binom{n}{B}^{-1} \binom{n-2}{B-2} \left(\sum_{i \in \{1,\ldots,n\}} \sum_{j \neq i} \left(\nabla f_i(w) - \nabla f(w) \right)^T \left(\nabla f_j(w) - \nabla f(w) \right) \right) \tag{1}$$

$$\frac{1}{B^2} \binom{n}{B}^{-1} \binom{n-1}{B-1} \left(\sum_{i \in \{1,\ldots,n\}} \sum_{j = i} \left(\nabla f_i(w) - \nabla f(w) \right)^T \left(\nabla f_j(w) - \nabla f(w) \right) \right) \tag{2}$$

Our final result is the sum of (1) and (2),

$$\frac{1}{B^2}\binom{n}{B}^{-1}\binom{n-2}{B-2}\left(\sum_{i\in\{1,\dots,n\}}\sum_{j\neq i}\left(\nabla f_i(w)-\nabla f(w)\right)^T\left(\nabla f_j(w)-\nabla f(w)\right)\right) \quad (1)$$

$$+\frac{1}{B^2}\binom{n}{B}^{-1}\binom{n-1}{B-1}\left(\sum_{i\in\{1,\dots,n\}}\sum_{j=i}\left(\nabla f_i(w)-\nabla f(w)\right)^T\left(\nabla f_j(w)-\nabla f(w)\right)\right) \quad (2)$$

Analyzing (1) yields

$$\frac{(B-1)}{Bn(n-1)}\left(\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\left(\nabla f_i(w)-\nabla f(w)\right)^T\left(\nabla f_j(w)-\nabla f(w)\right)\right)$$

$$=\frac{(B-1)}{Bn(n-1)}\left(\sum_{i=1}^{n}\left(\nabla f_i(w)-\nabla f(w)\right)^T\sum_{j=1,j\neq i}^{n}\left(\nabla f_j(w)-\nabla f(w)\right)\right)$$

$$=\frac{(B-1)}{Bn(n-1)}\left(\sum_{i=1}^{n}\left(\nabla f_i(w)-\nabla f(w)\right)^T\left(-\left(\nabla f_i(w)-\nabla f(w)\right)\right)\right)$$

$$=\frac{-(B-1)}{Bn(n-1)}\sum_{i=1}^{n}\left\|\nabla f_i(w)-\nabla f(w)\right\|^2$$

$$=\frac{-\sigma^2(B-1)}{B(n-1)}$$

Analyzing (2) yields

$$\frac{1}{Bn}\sum_{i=1}^{n}\left\|\nabla f_i(w)-\nabla f(w)\right\|^2=\frac{\sigma^2}{B}.$$

Adding (1) and (2) together gives the final solution of

$$\mathbb{E}\left[\left\|\frac{1}{B}\sum_{i\in\mathcal{B}}\nabla f_i(w)-\nabla f(w)\right\|^2\right]=\frac{\sigma^2(n-B)}{B(n-1)}$$

and we note that

$$\lim_{n\to\infty}\frac{\sigma^2(n-B)}{B(n-1)}=\frac{\sigma^2}{B} \qquad \text{(by L'Hopitals)}$$

This variance is *higher* than sampling with replacement. As n approaches infinity, for a fixed B, sampling with and without replacement yield the same expression of $\frac{\sigma^2}{B}$. This

makes sense! Sampling with replacement from an infinite population size is just about guaranteeing you never sample the same thing twice which would yield the result from sampling with replacement. ■

Another variant of minibatched SGD is *random reshuffling*. Here, rather than selecting a new minibatch by random sampling at each iteration, the algorithm shuffles *the entire dataset* and then produces samples by just running linearly through the dataset in that order. Once it has used the whole dataset, then it re-shuffles the data for the next pass. Concretely, this corresponds to the following algorithm.

Algorithm 3: Minibatched Stochastic Gradient Descent Using Random Reshuffling

input: learning rate α, minibatch size B where B divides n evenly, initial parameters $w_{0,0}$
input: total number of epochs K
for $k = 0$ *to* $K - 1$ **do**
  sample ζ uniformly at random from the set of permutations of $\{1, ..., n\}$
  for $t = 0$ *to* $n/B - 1$ **do**
    update model $w_{k,t+1} \leftarrow w_{k,t} - \alpha_t \cdot \frac{1}{B} \sum_{i=1}^{B} \nabla f_{\sigma(i+Bt)}(w_{k,t})$
  end
  $w_{k+1,0} \leftarrow w_{k,n/B}$
end

For the above algorithm assume that $B \neq 1$ and $B \neq n$, but $B|n$. Also note that ζ is a bijection from $\{1, ..., n\}$ to $\{1, ..., n\}$.

(b) Consider the *first iteration* of minibatch SGD using this sampling strategy. Is the variance/squared error of its gradient estimator going to be:

i. equivalent to sampling with replacement,
ii. equivalent to sampling without replacement, or
iii. equivalent to neither?

Briefly justify.

Solution. **(ii)** There is a clear bijection between shuffling a list and then selecting the first k elements and selecting distinct elements. ■

(c) Now consider *all* the iterations of minibatched SGD in the first epoch using this sampling strategy, and compare this to running the same number of iterations of SGD using with- or without- replacement sampling as described above. Is this random sampling going to be

i. statistically equivalent to minibatch SGD using sampling with replacement
ii. statistically equivalent to minibatch SGD using sampling without replacement, or
iii. statistically equivalent to neither?

Here, "statistically equivalent" means that the distribution of the output weights will necessarily be equal between the two algorithms.

Solution. **(iii)** This clearly isn't the same as (i) and it's not the same as (ii) because we only reshuffle at the end of using all elements. ■

2. **Estimating Large Sums and Concentration Inequalities.** Suppose that we want to approximate the empirical risk of some hypothesis on the validation set

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$$

of size n, where each $x_i \in \mathbb{R}^d$ and each $y_i \in \{-1, 1\}$. We are interested in the empirical risk with two different loss functions, the $0-1$ loss (corresponds to error) and the hinge loss (the loss function of a soft-margin SVM). Recall that we can write empirical risk as

$$R(h) = \frac{1}{n}\sum_{i=1}^{n} L(h(x_i), y_i).$$

For the $0-1$ loss,

$$L_{0-1}(\hat{y}, y) = \begin{cases} 1 & \hat{y} \cdot y \leq 0 \\ 0 & \hat{y} \cdot y > 0 \end{cases}$$

and for the hinge loss

$$L_{hinge}(\hat{y}, y) = \max(0, 1 - \hat{y} \cdot y).$$

Suppose that the hypotheses we are trying to evaluate are linear predictors of the form

$$h_w(x) = w^T x,$$

and that the magnitude of the examples in the validation dataset are bounded by

$$\max_i \|x_i\| = X_{\max}.$$

We approximate this empirical risk using a sum of training examples sampled with replacement. Explicitly, we sample a sequence of identically distributed random variables, $\{Z_i\}_{i=1}^K$, each of which is sampled according to

$$Z_k = L(h(x_i), y_i) \text{ with probability } \frac{1}{n} \text{for all } i \in \{1, ..., n\}.$$

That is, each Z_k is a randomly sampled term of the empirical risk sum. We then approximate the empirical risk with the sum

$$S_K = \frac{1}{K}\sum_{k=1}^{K} Z_k.$$

(a) Use Hoeffding's inequality to prove that, for the $0-1$ loss, the error of the estimate for a single hypothesis h is bounded, for any $a \geq 0$, by

$$\mathbb{P}\left(|S_K - R(h)| \geq a\right) \leq 2 \cdot \exp(-2a^2 K)$$

Solution. Since $Z_1, \cdots Z_k$ are independent random variable, we can apply Hoeffding's inequality:

$$\mathbb{P}\left(S_k - \mathbb{E}[S_K]\right) \geq a) \leq 2\exp\left(-\frac{2Ka^2}{(z_{max} - z_{min})^2}\right)$$

where $z_{max} = 1$ and $z_{min} = 0$. Also $\mathbb{E}[S_k] = R(h)$ by the law of large numbers and linearity of expectation. so we get:

$$\mathbb{P}\left(S_k - R(h)\right) \geq a) \leq 2\exp\left(-2Ka^2\right)$$

■

(b) Use Hoeffding's inequality to prove that, for the hinge loss, the error of the estimate for a single hypothesis h is bounded, for any $a \geq 0$, by

$$\mathbb{P}(|S_K - R(h)| \geq a) \leq 2 \cdot \exp\left(-\frac{2a^2K}{(X_{\max} \cdot \|w\| + 1)^2}\right).$$

Solution. We proceed similar to part (a), except note that $z_{max} = \max(0, 1 - \hat{y} \cdot y)$. So we have:

$$\begin{aligned}\|z_{max}\| &= \|1 - \hat{y} \cdot y\| \\ &\leq 1 + \|\hat{y}\| \cdot \|y\| \\ &= 1 + X_{max} \cdot \|w\|\end{aligned}$$

where we get the last line since $\hat{y} = w^T x$ so $\max \hat{y} = \max w^T x = \|w\|\|x\| = \|w\| \cdot X_{max}$. Also we turn the subtraction into addition when y takes the label -1.
Also, we have that $z_{min} = 0$, so it follows that:

$$z_{max} - z_{min} = (X_{max} \cdot \|w\| + 1)$$

Plugging in, we get the resulting inequality:

$$\mathbb{P}(|S_K - R(h)| \geq a) \leq 2 \cdot \exp\left(-\frac{2a^2K}{(X_{\max} \cdot \|w\| + 1)^2}\right).$$

■

(c) Now suppose that $X_{\max} = \|w\| = 1$. Using the inequality derived in 1(b), how large does K need to be so that we can be 99% sure that our estimate for the hinge loss is accurate to within a maximum error of 0.1?

Solution. We set $a = 0.1$ and solve for when the right hand side of the expression is equal to 0.01. This way we are calculating that the probability that we differ from the expected by more than 0.1 is less than 1% which is equivalent to calculating that the expected value is within 0.1 with probability 99%:

$$\mathbb{P}(|S_K - R(h)| \geq a) \leq 2 \cdot \exp\left(-\frac{2a^2K}{(X_{\max} \cdot \|w\| + 1)^2}\right)$$

Setting this equal:

$$2 \cdot \exp\left(-\frac{2a^2K}{(X_{\max} \cdot \|w\| + 1)^2}\right) = 0.01$$

substituting:

$$2 \cdot \exp\left(-\frac{2(0.1^2)K}{(1 \cdot 1 + 1)^2}\right) = 0.01$$

we get $K = \boxed{1060}$ ■

(d) In some cases you may not know the values of things like $X_{\max}$ a priori. To address this, your data scientist friend Minerva suggests the following approach. First, sample all the values $Z_1, ..., Z_k$ that you are going to use in the sum. Second, compute the minimum and maximum value of those samples as

$$Z_{\min} = \min_k Z_k \quad \text{and} \quad Z_{\max} = \max_k Z_k.$$

Finally, get a bound by plugging them into Hoeffding's inequality, resulting in

$$\mathbb{P}(|S_K - R(h)| \geq a) \leq 2 \cdot \exp\left(-\frac{2a^2K}{(Z_{\max} - Z_{\min})^2}\right)$$

Is Minerva's approach valid?

Solution. This is not valid because the experimentally calculated $(Z_{max} - Z_{min})^2$ will be at most the actual difference squared. This means that the experimentally calculated right side of the inequality will be less than or equal to the actual expression (since the RHS is monotonically increasing with increasing value of $(Z_{\max} - Z_{\min})$. Concretely, let Z_{maxact} and Z_{minact} be the actual max and min of random variable Z_i. Then we have:

$$2 \cdot \exp\left(-\frac{2a^2K}{(Z_{\max} - Z_{\min})^2}\right) \leq 2 \cdot \exp\left(-\frac{2a^2K}{(Z_{\max\, act} - Z_{\min\, act})^2}\right)$$

This means that our experimental upper bound may be less than the actual upper bound, which could be incorrect. So this approach does not work.

■