8 APPENDIX

8.1 EMPIRICAL EVALUATION OF SAMPLING

We wish to evaluate the empirical accuracy of our sampling technique on concrete examples. We do this in two ways. First, we can sort the elements by probability and make events of drawing an element in the top 10, or the top 100, top 1000, etc. We show the results for two random θ with different distributions in Figure 8 for $50,000$ samples. Note that our method closely matches the histogram of the true distribution. For a more comprehensive evaluation, we sample 30 values of θ and compute the relative error for exact sampling and our approximate sampling. See Figure 8. We also present in Figure 7 the amortized speedups obtained by our method. The amortized cost is defined as the time needed to train the index, added to the runtime of sampling $10,000$ samples.

8.2 LEARNING TRAINING SET

For the learning experiment, we show the set of images $\mathcal{D}$ that we maximized the probability of. See Figure 9. The common theme of the images is the presence of water.

8.3 VALUE OF c

For all of the proofs, we will include the result with the approximate MIPS with an error of c. To recover the original results in the paper, set $c = 0$.

8.4 SAMPLING

Theorem 3.2. *For Algorithm 1,* $\mathbb{E}[m] \leq \frac{ne^c}{k}$

Proof. Note that m is the number of Gumbels that are larger than $B = M - S_{min} - c$.

Note that Gumbels can be defined by $-\ln(-\ln(U_i))$ where U_i is a uniform random variable on the interval $[0, 1]$. Thus, we can think of each point having a uniform sample U_i and finding places where

$$-\ln(-\ln(U_i)) > M - S_{min} - c \tag{17}$$

$$U_i > \exp(-\exp(S_{min} + c - M)) \tag{18}$$

Thus, if we can find places where $U_i > \exp(-\exp(S_{min}+c-M))$, then we have the value of m. The number of points where this occurs is distributed according to $Bin(n-|S|, 1-\exp(-\exp(S_{min}+c-M))$.

Thus,

$$\mathbb{E}[m|M] = (n - |S|)(1 - \exp(-\exp(S_{min} + c - M))) \tag{19}$$

$$\mathbb{E}[m|M] \leq n \exp(S_{min} + c - M) \tag{20}$$

Note that

$$\Pr[ne^{S_{min}+c-M} > x] = \Pr[M - S_{min} - c < \ln(n/x)] \tag{21}$$

$$= \Pr[(\max_{i \in S} y_i + G_i) - S_{min} - c < \ln(n/x)] \tag{22}$$

$$\leq \Pr[\max_{i \in S} G_i - c < \ln(n/x)] \tag{23}$$

$$\leq \Pr[\ln(|S|) + G - c < \ln(n/x)] \tag{24}$$

$$\leq \Pr[\frac{ne^c e^{-G}}{|S|} > x] \tag{25}$$

$$\leq \Pr[exponential(\frac{ne^c}{|S|}) > x] \tag{26}$$

And thus,

$$\mathbb{E}[ne^{S_{min}+c-M}] \leq \mathbb{E}[exponential(\frac{ne^c}{|S|})] = \frac{ne^c}{|S|} \tag{27}$$

Putting it all together,

$$\mathbb{E}[m] = \mathbb{E}[\mathbb{E}[m|M]] \leq \frac{ne^c}{|S|} \tag{28}$$

□

Theorem 3.3. *For Algorithm 2, the sample is an exact sample with probability* $1 - \delta$ *for* $\delta = \exp(-\frac{kl}{n}e^{-c})$.

Proof. Note that the elements not in $S \cup T$ have values $y_i \leq S_{min} + c$ and $G_i \leq B = -\ln(-\ln(1 - l/n))$. Further, there exists an element in S with $y_i \geq S_{min}$ and with $G_i = \max_{j=1}^{k} G_j$. As long as there is an element in S that exceeds all the elements not in $S \cup T$, the sample will be exact.

$$\Pr[\text{not exact sample}] \leq \tag{29}$$

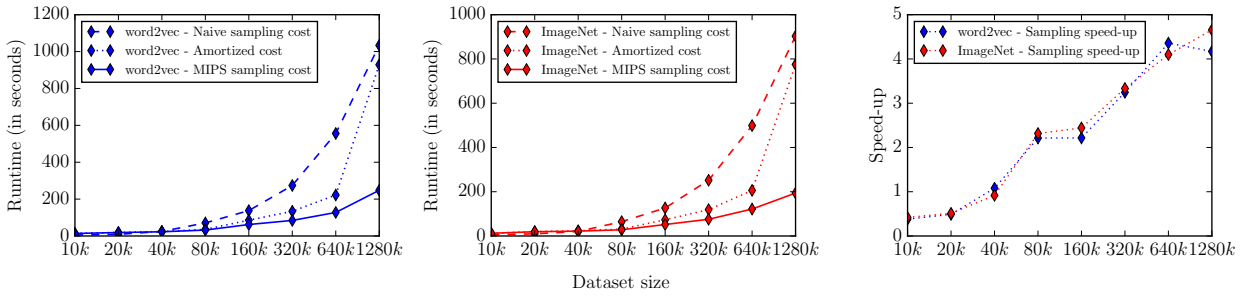

Figure 7: *Left and Center:* empirical comparison of the runtime of sampling for a $10,000$ randomly chosen θ on Word Embeddings and Image Net for varying fraction of the data. The amortized cost is defined as the time necessary for the sampling in addition to the training time of the index. *Right:* Evaluation of the sampling speed-up for both datasets for varying fraction of the data.

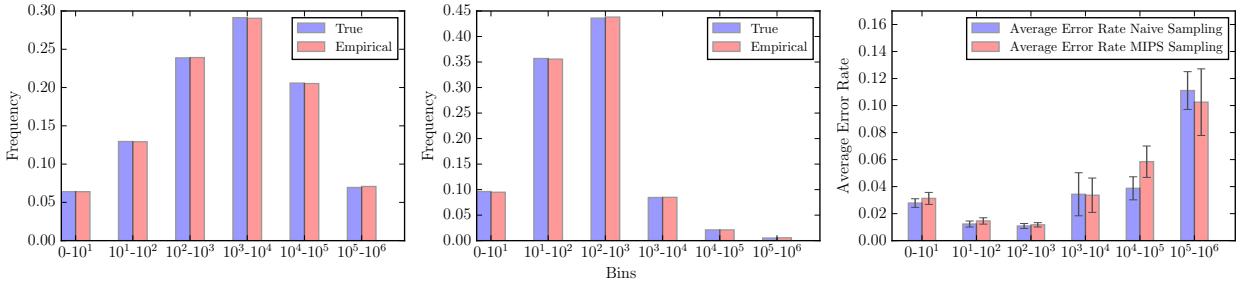

Figure 8: *Left and Center:* Two randomly chosen θ with different bin distributions. We see that the empirical sampling closely matches the true distribution for all the bins. *Right:* Evaluation of the relative error on 30 samples of θ for the exact sampling and our sampling technique. The error bars are for the average error rate between the empirical distribution and the true distribution for both exact sampling and our method. We see that the error rates are not statistically significantly different.

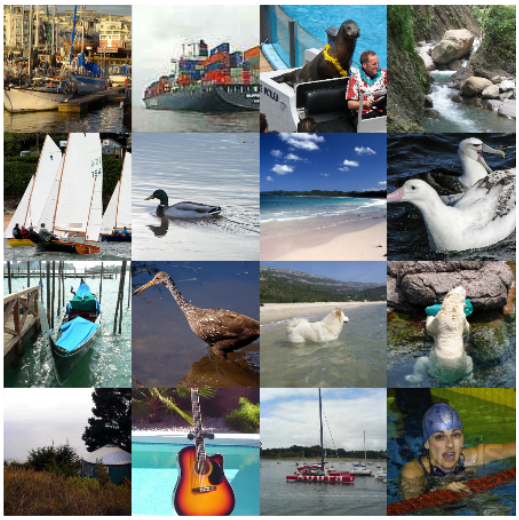

Figure 9: The set of images $\mathcal{D}$ used in the learning experiment. Note that all of the images contain water, though the content of the images is quite different.

$$\leq \Pr[\max_{j=1}^{k} G_j < -\ln(-\ln(1-l/n)) + c] \tag{30}$$

$$\leq \Pr[\ln(k) - \ln(-\ln(U)) < -\ln(-\ln(1-l/n)) + c] \tag{31}$$

$$\leq \Pr[k(-\ln(1-l/n))e^{-c} < -\ln(U)] \tag{32}$$

$$\leq \Pr[\frac{kl}{n}e^{-c} < -\ln(U)] \tag{33}$$

$$\leq \Pr[\exp(-\frac{kl}{n}e^{-c}) > U] \tag{34}$$

$$\leq \exp(-\frac{kl}{n}e^{-c}) \tag{35}$$

Thus, the probability of failure, δ, is bounded by $\exp(-\frac{kl}{n}e^{-c})$ □

8.5 PARTITION FUNCTION ESTIMATE

Theorem 3.4. *Algorithm 3 returns an unbiased estimate $\hat{Z}$ and for $kl \geq \frac{2}{3}\frac{1}{\epsilon^2}ne^c\ln(1/\delta)$, then with $1-\delta$ probability,*

$$\frac{|\hat{Z}-Z|}{Z} \leq \epsilon$$

Proof. Define $Z = \sum_i e^{y_i}$. Let S be the indices of the k largest elements of $\{y_i\}$ and $S' = [1,n] \setminus S$. Denote $||S'||_1 = \sum_{i\in S'} e^{y_i}$ and $||S||_1 = \sum_{i\in S} e^{y_i}$. Thus, the true partition function is $Z = \sum_{i\in S} e^{y_i} + \sum_{i\in S'} e^{y_i} = ||S||_1 + ||S'||_1$.

Let r be the largest value of e^{y_i} for elements in S'. Then for all $i \in S'$, $e^{y_i} \in (0,r]$ and further, for all $i \in S$, $e^{y_i} \geq re^{-c}$. We can scale these values of S' and denote them as $q_i = \frac{e^{y_i}}{r}$ where $q_i \in [0,1]$.

For the estimate $\hat{Z}$ we will draw l samples with replacement from S' and denote the set as T. Denote the samples as $y^{(j)}$ and the scaled versions as $q^{(j)} = \frac{e^{y^{(j)}}}{r}$.

We use the estimate:

$$\hat{Z} = \sum_{i\in S} e^{y_i} + \frac{|\mathcal{X}-S|}{|T|}\sum_{i\in T} e^{y_i} \tag{36}$$

$$\hat{Z} = ||S||_1 + \frac{(n-k)r}{l}\sum_{j=1}^{l} q^{(j)} \tag{37}$$

Note that

$$\mathbb{E}[\hat{Z}] = (n-k)r\mathbb{E}[q^{(1)}] + ||S||_1 \tag{38}$$

$$\mathbb{E}[\hat{Z}] = (n-k)r\sum_{i\in S'}\frac{1}{|S'|}\frac{e^{y_i}}{r} + ||S||_1 = ||S'||_1 + ||S||_1 = Z \tag{39}$$

This is because $|S'| = n-k$. Thus, $\hat{Z}$ is an unbiased estimator of Z. However, we are concerned if it is well concentrated about its mean.

Let Q be a random variable as the scaled sample from S' and $\bar{Q}$ be the empirical mean over l samples. Thus, $\mathbb{E}[Q] = \frac{||S'||_1}{(n-k)r}$.

Note that

$$|\hat{Z}-Z| = |\frac{(n-k)r}{l}\sum_{j=1}^{l} q^{(j)} - ||S'||_1| \tag{40}$$

$$|\hat{Z}-Z| = |\frac{(n-k)r}{l}\sum_{j=1}^{l} q^{(j)} - (n-k)r\mathbb{E}[Q]| \tag{41}$$

$$|\hat{Z}-Z| = (n-k)r|\frac{1}{l}\sum_{j=1}^{l} q^{(j)} - \mathbb{E}[Q]| \tag{42}$$

$$|\hat{Z}-Z| = (n-k)r|\bar{Q} - \mathbb{E}[Q]| \tag{43}$$

Therefore,

$$\Pr[|\hat{Z}-Z| > \epsilon Z] = \Pr[(n-k)r|\bar{Q}-\mathbb{E}[Q]| > \epsilon(||S'||_1 + ||S||_1)] \tag{44}$$

$$= \Pr[(n-k)r|\bar{Q}-\mathbb{E}[Q]| > \epsilon((n-k)r\mathbb{E}[Q] + ||S||_1)] \tag{45}$$

$$= \Pr[|\bar{Q} - \mathbb{E}[Q]| > \epsilon(\mathbb{E}[Q] + \frac{||S||_1}{(n-k)r})] \quad (46)$$

$$\leq \Pr[|\bar{Q} - \mathbb{E}[Q]| > \epsilon(\mathbb{E}[Q] + \frac{ke^{-c}}{n})] \quad (47)$$

If we use Chernoff (use a convexity argument to bound the MGF in terms of the mean as on page 22 of "Concentration of Measure for the Analysis of Randomised Algorithms" by Dubhashi and Panconesi)

$$\Pr[|\sum_j Q^{(j)} - l\mathbb{E}[Q]| > \delta l\mathbb{E}[Q]] \leq 2\exp(-\frac{1}{3}\delta^2 l\mathbb{E}[Q]) \quad (48)$$

$$\Pr[|\bar{Q} - \mathbb{E}[Q]| > \delta\mathbb{E}[Q]] \leq 2\exp(-\frac{1}{3}\delta^2 l\mathbb{E}[Q]) \quad (49)$$

$$\Pr[|\bar{Q} - \mathbb{E}[Q]| > a] \leq 2\exp(-\frac{1}{3}a^2 l\frac{1}{\mathbb{E}[Q]}) \quad (50)$$

Combining these two by setting $a = \epsilon(\mathbb{E}[Q] + \frac{ke^{-c}}{n})$,

$$\Pr[|\bar{Q} - \mathbb{E}[Q]| > \epsilon(\mathbb{E}[Q] + \frac{ke^{-c}}{n})] \leq \quad (51)$$

$$\leq 2\exp(-\frac{1}{3}\epsilon^2 l(\mathbb{E}[Q] + \frac{ke^{-c}}{n})^2\frac{1}{\mathbb{E}[Q]}) \quad (52)$$

$$\leq 2\exp(-\frac{2}{3}\epsilon^2\frac{kle^{-c}}{n}) \quad (53)$$

Thus, as long as $kl \geq \frac{2}{3}\frac{1}{\epsilon^2}ne^c\ln(1/\delta)$, then with $1-\delta$ probability, $\hat{Z} \in (1 \pm \epsilon)Z$ □

returns an unbiased estimate $\hat{Z}$ and for $kl = \frac{2}{3}\frac{1}{\epsilon^2}n\ln(1/\delta)$, then with $1-\delta$ probability, $\hat{Z} \in (1 \pm \epsilon)Z$

8.6 EXPECTATION ESTIMATE

Theorem 3.5. *Algorithm 4 returns an estimate $\hat{F}$ such that $|\hat{F} - F| \leq \epsilon C$ with probability δ if*

$$lk^2 \geq \frac{8n^2e^{2c}}{\epsilon^2}\ln(4/\delta)$$

and

$$kl \geq \frac{8}{3}\frac{1}{\epsilon^2}ne^c\ln(2/\delta)$$

Proof. Recall

$$J = \sum_i e^{y_i}f_i$$

$$\hat{J} = \sum_{i\in S} e^{y_i}f_i + \frac{n-k}{l}\sum_{i\in T} e^{y_i}f_i$$

Thus, $F = J/Z$ and $\hat{F} = \hat{J}/\hat{Z}$.

To show that

$$|\hat{F} - F| = |\frac{\hat{J}}{\hat{Z}} - \frac{J}{Z}| \leq \epsilon C$$

with probability $1-\delta$, we will show that

$$|\frac{\hat{J}}{\hat{Z}} - \frac{\hat{J}}{Z}| \leq \frac{\epsilon}{2}C$$

$$|\frac{\hat{J}}{Z} - \frac{J}{Z}| \leq \frac{\epsilon}{2}C$$

each with probability $1 - \delta/2$. These will be shown as two separate parts.

8.6.1 Part One

Because

$$kl \geq \frac{2}{3}\frac{4}{\epsilon^2}ne^c\ln(2/\delta)$$

from Theorem 3.4, with probability $1 - \delta/2$ then $\frac{|\hat{Z}-Z|}{Z} \leq \epsilon$.

$$|\frac{\hat{J}}{\hat{Z}} - \frac{\hat{J}}{Z}| = \frac{|\hat{J}|}{\hat{Z}}\frac{|\hat{Z} - Z|}{Z}$$

$$\leq \frac{|\hat{J}|}{\hat{Z}}\frac{\epsilon}{2}$$

$$\leq \frac{\epsilon}{2}C$$

8.6.2 Part Two

For the second one is written as the following lemma

Lemma 8.1. $|\frac{\hat{J}}{Z} - \frac{J}{Z}| \leq \frac{\epsilon}{2}C$ *with probability* $1 - \delta/2$ *for*

$$lk^2 \geq \frac{8n^2}{\epsilon^2}\ln(4/\delta)$$

Proof. Note that the smallest element in S has "probability" $e^{y_i}/Z \leq 1/k$. Thus, for the largest element not in S, $e^{y_i}/Z \leq e^c/k$. Define Q as the random variable of the value of sampling i uniformly from $[1, n] \setminus S$ and returning

$$Q = \frac{ke^{y_i}}{e^c Z}\frac{f_i}{C}$$

Note that $Q \in [-1, 1]$ and that $\mathbb{E}[Q] = \frac{1}{n-k}\frac{k}{ZCe^c}\sum_{i \notin S} e^{y_i} f_i$. The elements of T are samples $\{y^{(j)}\}_i$ and $\{f^{(i)}\}_i$ and thus we can define

$$Q^{(j)} = \frac{ke^{y^{(j)}}}{e^c Z}\frac{f^{(j)}}{C}$$

$$|\frac{J}{Z} - \frac{\hat{J}}{Z}| = \frac{1}{Z}|\sum_{i \notin S} e^{y_i} f_i + \frac{n-k}{l}\sum_{j=1}^{l} e^{y^{(j)}} f^{(j)}|$$

$$= \frac{(n-k)Ce^c}{k}|\mathbb{E}[Q] - \frac{1}{l}\sum_j Q^{(i)}|$$

From Hoeffding's Inequality,

$$\Pr[|\mathbb{E}[Q] - \frac{1}{l}\sum_j Q^{(i)}| > t] \leq 2\exp(-\frac{lt^2}{2})$$

Thus,

$$\Pr[|\frac{J}{Z} - \frac{\hat{J}}{Z}| > \frac{(n-k)Ce^c}{k}t] \leq 2\exp(-\frac{lt^2}{2})$$

Defining $t = \frac{k\epsilon}{2(n-k)e^c}$ we get that $\frac{(n-k)Ce^c}{k}t = \frac{\epsilon}{2}C$, so

$$\Pr[|\frac{J}{Z} - \frac{\hat{J}}{Z}| > \frac{\epsilon}{2}C] \leq 2\exp(-\frac{lk^2\epsilon^2}{8(n-k)^2e^{2c}})$$

$$\Pr[|\frac{J}{Z} - \frac{\hat{J}}{Z}| > \frac{\epsilon}{2}C] \leq 2\exp(-\frac{lk^2\epsilon^2}{8n^2e^{2c}})$$

Thus, the conclusion of the Lemma follows for

$$lk^2 \geq \frac{8n^2e^{2c}}{\epsilon^2}\ln(4/\delta)$$

□

With this lemma, the conclusion of the theorem follows. □

Theorem 3.6. *There exists a MIPS technique that returns the approximate top k elements in sublinear time.*

Proof. For the data structure, we create a sequence of LSH instances that are tuned to values that are $c/2$ apart. Thus, if $\|\theta\| \leq M_1$ and $\|\phi(x)\| \leq M_2$, then $|\theta \cdot \phi(x)| \leq M_1M_2$. And we create $n_{LSH} = \frac{4M_1M_2}{c}$ instances.

Call the LSH instances $\{L_i\}_i$ and for the i^{th} instance, set the lower tuned value to be $S_{i,2} = (c/2)(i-1) - M_1M_2$ and the higher tuned value to be $S_{i,1} = (c/2)i - M_1M_2$. Thus, $S_1 - S_2 = c/2$. Further, set the failure probability of each LSH instance to be $\delta' = \delta k n_{LSH}$ so that with high probability, each of the LSH instances will not fail to find each of the top k values.

At query time, hash the query θ and let B_i be the buckets of neighbors from L_i. From the LSH guarantee, there are a small constant number of elements in B_i that are smaller than $S_{i,2}$ and with high probability, all elements larger than $S_{i,1}$ will be in B_i. Find the neighboring pair of LSH instances L_i and L_{i+1} where $|B_{i+1}| \leq k$ and $|B_i| \geq k$. Collect $k' = k - |B_{i+1}|$ elements $B' \subseteq B_i - B_{i+1}$ where the elements are larger than $S_{i,2}$ (all but a constant number will be larger than $S_{i,2}$). Then return the elements $S = B' \cup B_{i+1}$.

Note that any elements larger than $S_{i+1,1}$ will be in S with high probability because they will be contained in B_{i+1}. So $\max_{x \notin S} \theta \cdot \phi(x) \leq S_{i+1,1}$. Further, by construction, $\min_{x \in S} \theta \cdot \phi(x) \geq S_{i,2}$. Thus, the technique returns the approximate top k elements with high probability with a gap of $S_{i+1,1} - S_{i,2} = 2(c/2) = c$.

This technique will have a total runtime of

$$O(k + (\log(k) + \log(1/\delta))\log(n)n^\rho)$$

where $\rho < 1$. Thus, we have a sublinear approximate top k element MIPS technique. □