Adaptive Sampled Softmax with Kernel Based Sampling Supplementary Material

Guy Blanc[1] Steffen Rendle[2]

1. Sampled Softmax is biased

Theorem 1.1. *The gradient of sample softmax is an unbiased estimator of the full softmax gradient iff* $q_i = p_i \propto \exp(o_i)$

Proof. Note that the random variable in eq. (6) is the sample $\boldsymbol{s}$. Remember that the sample consists of one positive label that is chosen with probability 1 and m negatives that are sampled from $\boldsymbol{q}$. Our proof analyzes both parts separately. For notational convenience and without loss of generality, we assume that the positive class has index 1, $y_1 = 1$, and that the positive is at the first position in the sample, i.e., $s_1 = 1$.

Case 1: First we analyze the output o_i of the positive label, i.e., $y_i = 1$. Under our notational assumptions $i = 1$.

$$E\left[\sum_{j=1}^{m+1} I(s_j = i)p'_j\right] = E[p'_1]$$

$$= E\left[\frac{\exp(o_1)}{\exp(o_1) + \sum_{k=2}^{m+1}\exp(o'_k)}\right]$$

$$\overset{(A)}{\geq} \frac{\exp(o_1)}{\exp(o_1) + E\left[\sum_{k=2}^{m+1}\exp(o'_k)\right]}$$

$$\overset{eq.(12)}{=} \frac{\exp(o_1)}{\exp(o_1) + \sum_{k=2}^{n}\exp(o_k)} = p_i$$

The last step is based on:

$$E\left[\sum_{k=2}^{m+1}\exp(o'_k)\right] = \sum_{k=2}^{m+1} E\left[\frac{\exp(o_{s_k})}{mq_{s_k}}\right]$$

$$= \sum_{k=2}^{m+1}\sum_{j=2}^{n} q_j \frac{\exp(o_j)}{mq_j} = \sum_{j=2}^{n}\exp(o_j) \tag{12}$$

Step (A) is Jensen's inequality. This turns into an equality if and only if $\sum_{k=2}^{m+1}\exp(o'_{s_k})$ is constant which means $\exp(o'_l) = \exp(o_{s_l} - \ln mq_{s_l}) = \frac{\exp(o_{s_l})}{mq_{s_l}}$ has to be constant. This is true iff $q_{s_l} \propto \exp(o_{s_l})$. In other words, for a positive label, the softmax distribution is the only choice to create an unbiased sampled softmax.

Case 2: Let i be a negative class such that $y_i = 0$. Under our notational assumptions, $i > 1$. For the second case, we only show that softmax is unbiased, i.e., the sufficient condition of the theorem. We don't show that softmax is the only distribution that is unbiased. This necessary condition is already covered by case 1. Let the negative sampling distribution be the softmax $q_i := \frac{\exp(o_i)}{\sum_{l=2}^{n}\exp(o_j)}$.

$$E\left[\sum_{j=1}^{m+1} I(s_j = i)p'_j\right]$$

$$= E\left[\sum_{j=2}^{m+1} I(s_j = i)\frac{\exp(o'_j)}{\exp(o_1) + \sum_{k=2}^{m+1}\exp(o'_k)}\right]$$

$$\overset{eq.\ (13)}{=} \frac{1}{m}E\left[\sum_{j=2}^{m+1} I(s_j = i)\frac{1}{q_i}\frac{\exp(o_i)}{\exp(o_1) + \sum_{k=2}^{n}\exp(o_k)}\right]$$

$$= \frac{1}{m}E\left[\sum_{j=2}^{m+1} I(s_j = i)\frac{p_i}{q_i}\right] = p_i$$

This proof uses that for the softmax sampling distribution, the sum of corrected sampled outputs is equal to the sum of all outputs.

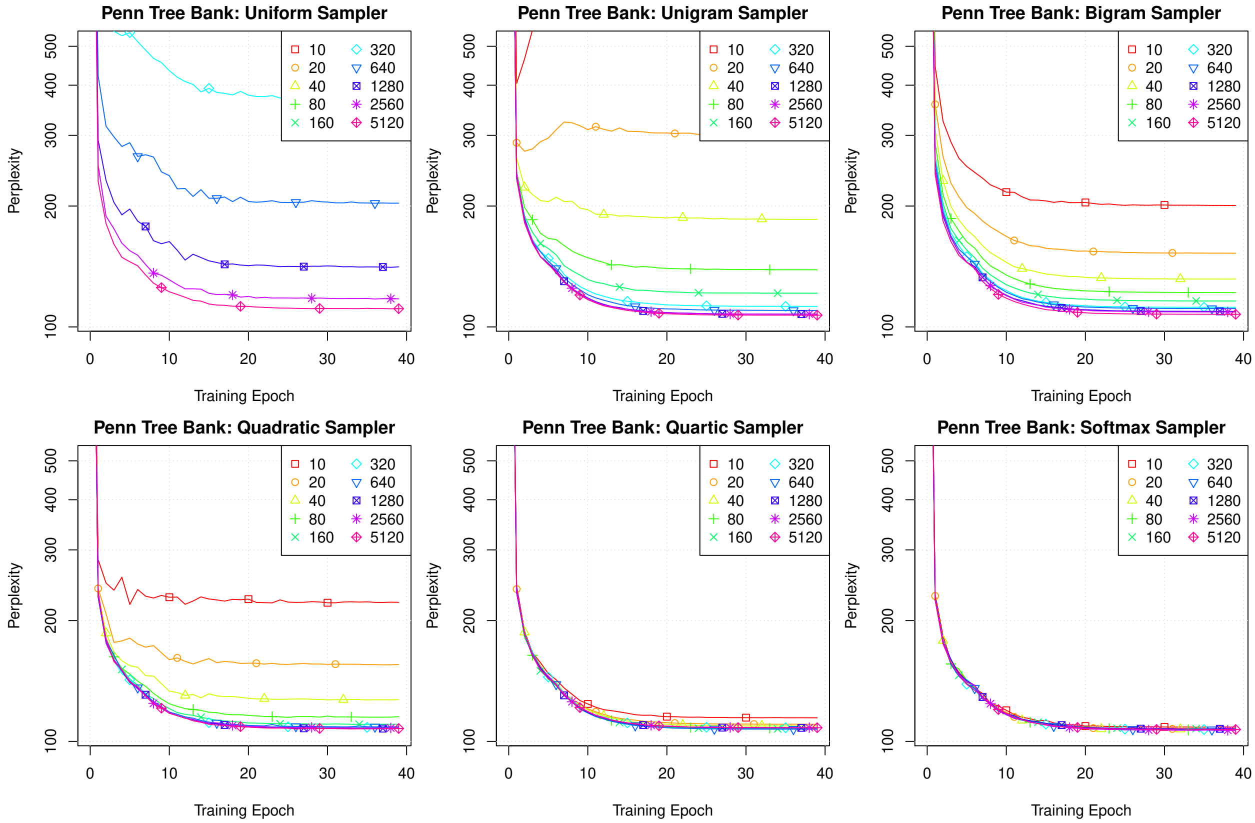
$$\sum_{k=2}^{m+1}\exp(o'_k) = \sum_{k=2}^{m+1}\frac{\exp(o_{s_k})}{m\,q_{s_k}} \tag{13}$$

$$= \sum_{k=2}^{m+1}\frac{\exp(o_{s_k})}{m}\frac{\sum_{l=2}^{n}\exp(o_l)}{\exp(o_{s_k})} = \sum_{l=2}^{n}\exp(o_l)$$

This equality is related to eq. (12). While eq. (12) holds for any sampling distribution but only in expectation, eq. (13) holds only for softmax but for any sample. □

[1] Work done during internship at Google, Mountain View, USA [2] Google, Mountain View, USA. Correspondence to: Guy Blanc <guy.blanc@gmail.com>, Steffen Rendle <srendle@google.com>.

Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).



Penn Tree Bank: Uniform Sampler

Penn Tree Bank: Unigram Sampler

Penn Tree Bank: Bigram Sampler

Penn Tree Bank: Quadratic Sampler

Penn Tree Bank: Quartic Sampler

Penn Tree Bank: Softmax Sampler

Figure 1. Convergence speed of different sampling distributions (*uniform*, *unigram*, *bigram*, *quadratic*, *quartic*, *softmax*) for a varying sample size $m \in \{10, 20, 40, \dots\}$ on the *Penn Tree Bank* dataset. Once enough samples are taken to remove the bias, adding more samples does not increase convergence speed considerably.

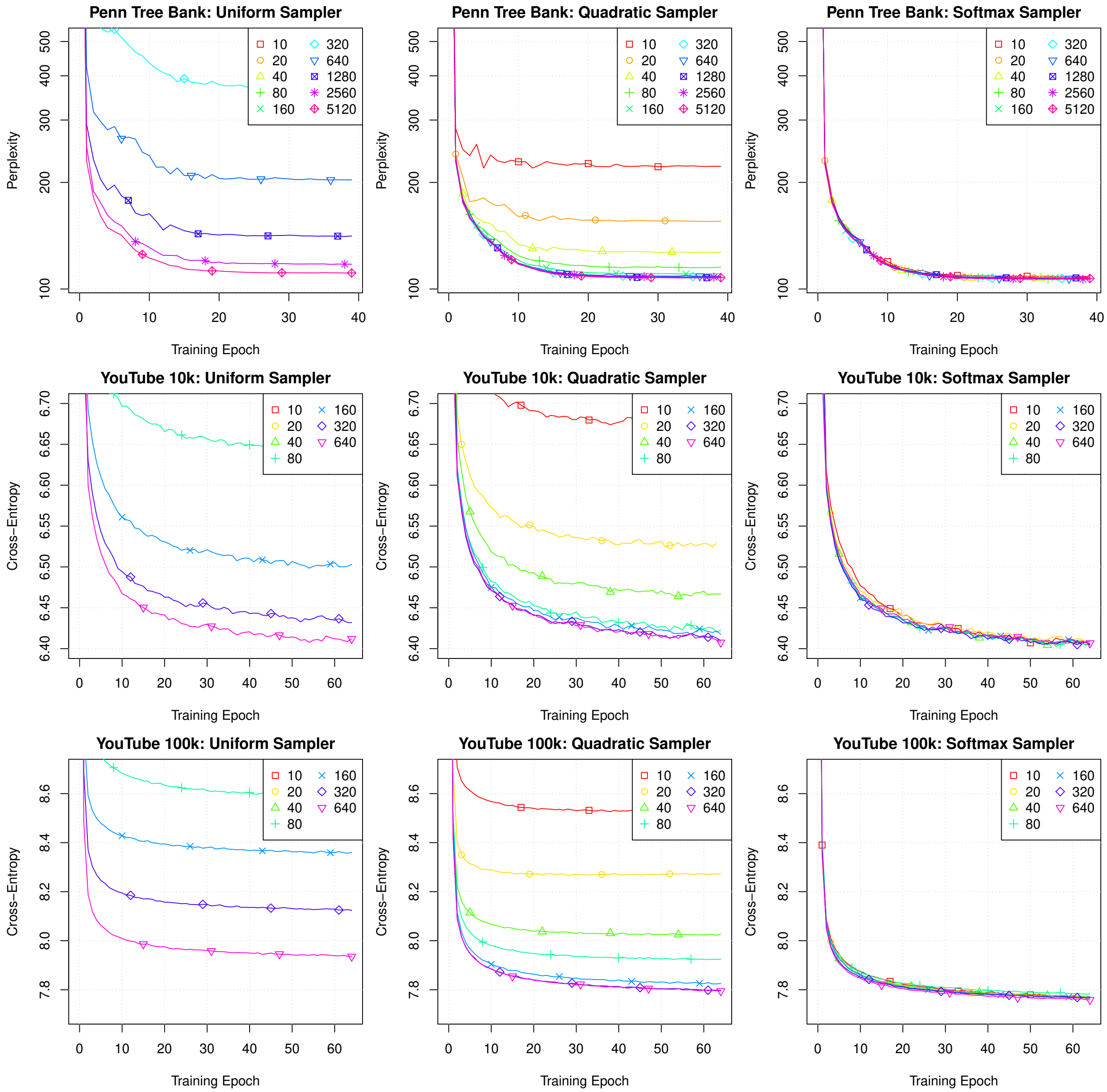


Figure 2. Convergence speed for three sampling distributions (*uniform*, *quadratic*, *softmax*) for a varying sample size $m \in \{10, 20, 40, \ldots\}$ on three datasets. Once enough samples are taken to remove the bias, adding more samples does not increase convergence speed considerably.

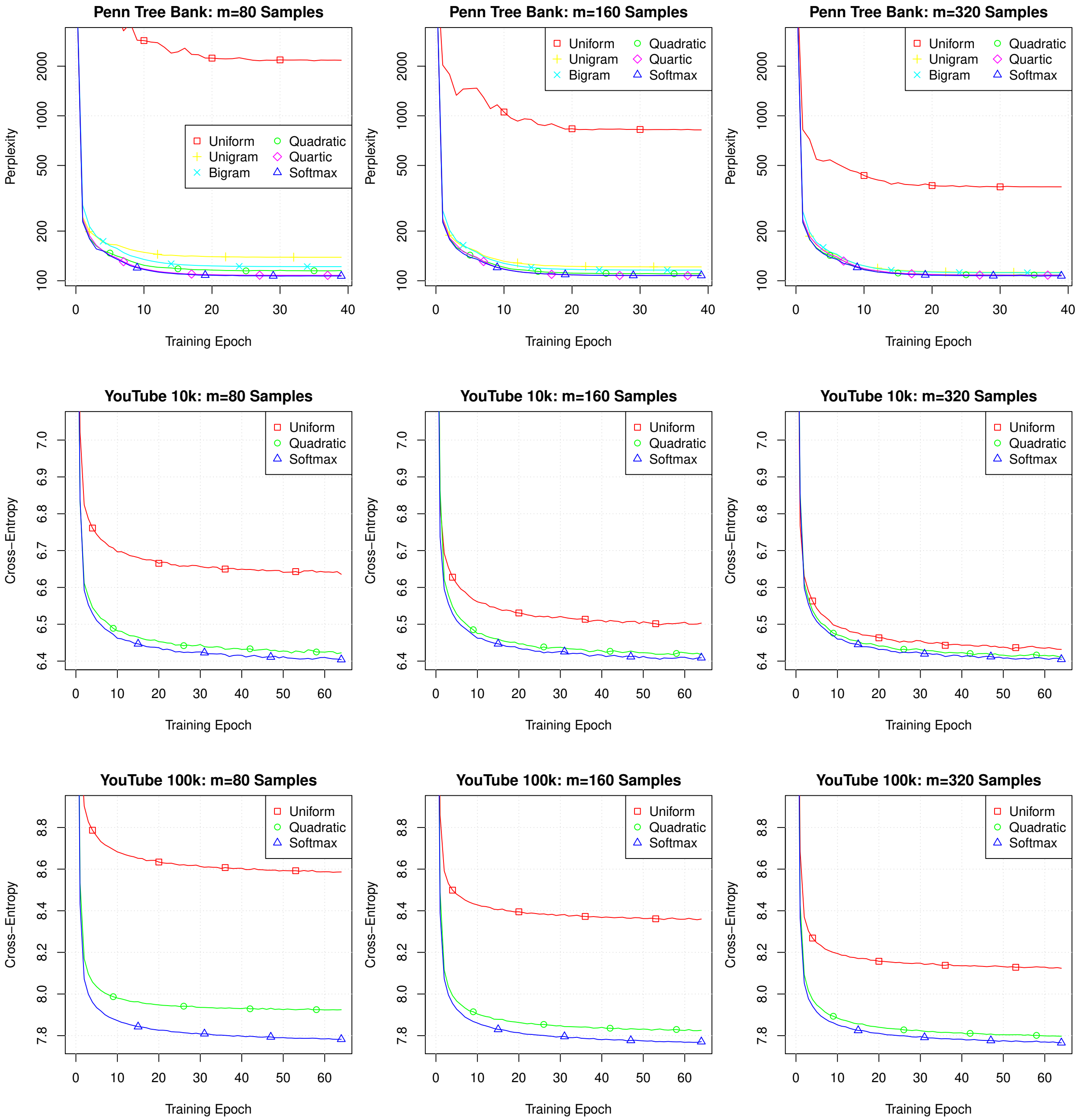


Figure 3. Convergence speed of different sampling distributions for a fixed sampling size. The convergence speed of all distributions is similar only the bias is different.