A Variance of `LogEstimator`

We now bound the variance of our estimator by $O(\log^2 k)$. Recall that the output of `LogEstimator` is given by $\log(\boldsymbol{X}/t) - g(\boldsymbol{B}_1, \dots, \boldsymbol{B}_r)$, where the function g is bounded. Since the variance we seek is $O(\log^2 k)$, it suffices to show that the variance of $\log(\boldsymbol{X}/t)$ is $O(\log^2 k)$ with $i \sim \mathcal{D}$, since subtracting g changes the estimate by at most a constant (see Lemma 2.3).

Lemma A.1. *Let $\boldsymbol{i} \sim \mathcal{D}$ and $\boldsymbol{X}$ denote the number of independent trials from $\mathrm{Ber}(p_{\boldsymbol{i}})$ before we see t successes. Then,* $\mathbf{Var}[\log(\boldsymbol{X}/t)] = O(\log^2 k)$.

Proof. Let $X_{\max} = 2kt$, and consider the random variable $\boldsymbol{X}' = \min\{\boldsymbol{X}, X_{\max}\}$. Then

$$\begin{aligned}\mathbf{Var}[\log(\boldsymbol{X}/t)] &\leq \mathbf{E}\left[\left(\log(\boldsymbol{X}/t) - \log(\boldsymbol{X}'/t) + \log(\boldsymbol{X}'/t)\right)^2\right] \\ &\leq 2 \cdot \mathbf{E}\left[\left(\log(\boldsymbol{X}/t) - \log(\boldsymbol{X}'/t)\right)^2\right] + 2 \cdot \mathbf{E}\left[\log^2(\boldsymbol{X}'/t)\right] \\ &\leq 2 \cdot \mathbf{E}\left[\log^2(\boldsymbol{X}/\boldsymbol{X}')\right] + 2\log^2(2k) \\ &\leq \frac{4}{\ln^2(2)} \cdot \mathbf{E}\left[\left(\sqrt{\frac{\boldsymbol{X}}{\boldsymbol{X}'} - 1}\right)^2\right] + 2\log^2(2k),\end{aligned}$$

where we used that $\log(\boldsymbol{X}'/t) \leq \log(2k)$ always, and that $\log(z) \leq \sqrt{z-1}/\ln(2)$ for all $z \geq 1$. Then,

$$\mathbf{E}\left[\frac{\boldsymbol{X}}{\boldsymbol{X}'} - 1\right] \leq \mathbf{E}\left[\frac{\boldsymbol{X}}{X_{\max}}\right] = \frac{1}{X_{\max}} \sum_{i=1}^{k} p_i \cdot \frac{t}{p_i} = \frac{tk}{X_{\max}} = 2.$$

□

B Omitted Details from Section 2

Proof of Claim 2.5. Notice that $\boldsymbol{X}$ is the number of trials from $\mathrm{Ber}(p_i)$ until we see t successes. We now have the following string of equalities:

$$\begin{aligned}\mathbf{E}_{\boldsymbol{X},\boldsymbol{B}_1,\dots,\boldsymbol{B}_r}\left[\boldsymbol{\eta} - \log\left(\frac{1}{p_i}\right)\right] &= \mathbf{E}_{\boldsymbol{X}}[\log \boldsymbol{Y}] - \mathbf{E}_{\boldsymbol{B}_1,\dots,\boldsymbol{B}_r}[g(\boldsymbol{B}_1, \boldsymbol{B}_2, \dots, \boldsymbol{B}_r)] \\ &= \mathbf{E}_{\boldsymbol{X}}[f(\boldsymbol{Y}) + h(\boldsymbol{Y})] - g(p_i, p_i^2, \dots, p_i^r) = \mathbf{E}_{\boldsymbol{X}}[h(\boldsymbol{Y})],\end{aligned}$$

where we used the fact that g is a linear function, and that $\mathbf{E}[\boldsymbol{B}_\ell] = p_i^\ell$ in order to substitute

$$\mathbf{E}_{\boldsymbol{B}_1,\dots,\boldsymbol{B}_r}[g(\boldsymbol{B}_1, \dots, \boldsymbol{B}_r)] = g(p_i, p_i^2, \dots, p_i^r).$$

Furthermore, we divide $\log \boldsymbol{Y} = f(\boldsymbol{Y}) + h(\boldsymbol{Y})$, where $f(z)$ is the degree-r Taylor expansion of $\log z$ at 1, and $h(z) = \log z - f(z)$ is the error in the degree-r Taylor expansion of $\log(z)$, i.e.,

$$h(z) = \log(z) - f(z).$$

Finally, by construction of g, $\mathbf{E}[f(\boldsymbol{Y})] = g(p_i, p_i^2, \dots, p_i^r)$, which gives the desired equality. □

Verifying $\boldsymbol{Y}$ is subgamma. Recall that $\boldsymbol{X}$ is the number of independent draws from a $\mathrm{Ber}(p)$ distribution until we see t successes. In other words, we may express $\boldsymbol{X} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_t$, where $\boldsymbol{X}_i$ is the number of draws of $\mathrm{Ber}(p)$ before we get a single success. Then, we always satisfy

$$\mathbf{E}[\boldsymbol{X}_i] = \frac{1}{p} \qquad \mathbf{Pr}[\boldsymbol{X}_i > \ell] = (1-p)^{\lceil \ell \rceil} < e^{-p\ell}.$$

This, in turn, implies that for any $r \geq 1$

$$\left(\mathbf{E}[|\boldsymbol{X}_i - 1/p|^r]\right)^{1/r} \leq \left(\mathbf{E}_{\boldsymbol{X}_i,\boldsymbol{X}_i'}\left[|\boldsymbol{X}_i - \boldsymbol{X}_i'|^r\right]\right)^{1/r} \leq 2\left(\mathbf{E}[|\boldsymbol{X}_i|^r]\right)^{1/r} = O(r/p),$$

where the first line is by Jensen's inequality, and the second is by the triangle inequality and Hölder inequality. Finally, we use the tail bound on $\boldsymbol{X}_i$ to upper bound the expectation of $|\boldsymbol{X}_i|^r$. Then, we have

$$\begin{aligned}\mathbf{E}\left[e^{\lambda(\boldsymbol{X}_i-1/p)}\right] &= 1+\lambda\mathbf{E}[\boldsymbol{X}_i-1/p]+\sum_{k=2}^{\infty}\frac{\lambda^k}{k!}\cdot\mathbf{E}[|\boldsymbol{X}_i-1/p|] \\ &= 1+\sum_{k=2}^{\infty}\frac{\lambda^k}{k!}\left(O(k/p)\right)^k \le 1+O(\lambda^2/p^2), \qquad \text{when } |\lambda| \text{ sufficiently smaller than } p \\ &\le \exp\left(O(\lambda^2/p^2)\right)\end{aligned}$$

Then, since $\boldsymbol{X}_1,\dots,\boldsymbol{X}_t$ are all independent, we have

$$\mathbf{E}\left[e^{\lambda(\boldsymbol{X}-t/p)}\right] \le \exp\left(O(\lambda^2 t/p^2)\right) \Longrightarrow \mathbf{E}\left[e^{\lambda(\boldsymbol{Y}-1)}\right] \le \exp\left(O(\lambda^2/t)\right),$$

and this bound is valid whenever $|\lambda|$ is sufficiently smaller than t.

C Omitted Proofs from Section 3

Proof of Lemma 3.1. The approach is to estimate

$$\mathbf{E}_{\boldsymbol{i}\sim\mathcal{D}}[h_t(p_{\boldsymbol{i}})] = \mathbf{E}_{\boldsymbol{i}\sim\mathcal{D}}\left[g(p_{\boldsymbol{i}},p_{\boldsymbol{i}}^2,\dots,p_{\boldsymbol{i}}^r)\right]. \tag{10}$$

There exists an algorithm using $O(\log(1/\epsilon)/\epsilon^2)$ samples to estimate the above quantity: for $j \in \{0,\dots,O(1/\epsilon^2)\}$, one takes a sample $\boldsymbol{i}_j \sim \mathcal{D}$ and uses $r = O(\log(1/\epsilon))$ additional samples $\boldsymbol{s}_1,\dots,\boldsymbol{s}_r \sim \mathcal{D}$ to define

$$\boldsymbol{B}_m^{(j)} \stackrel{\text{def}}{=} \mathbb{1}\{\boldsymbol{s}_1=\dots=\boldsymbol{s}_m=\boldsymbol{i}_j\} \quad \text{and} \quad \boldsymbol{Z}_j = g(\boldsymbol{B}_1^{(j)},\dots,\boldsymbol{B}_r^{(j)}).$$

Then, let $\boldsymbol{Z}$ be the average of all $\boldsymbol{Z}_j$'s, which is an unbiased estimate to $\mathbf{E}_{\boldsymbol{i}\sim\mathcal{D}}\left[g(p_{\boldsymbol{i}},p_{\boldsymbol{i}}^2,\dots,p_{\boldsymbol{i}}^r)\right]$. Since g is bounded (from Lemma 2.3), the variance of $O(1/\epsilon^2)$ such values is a large constant factor smaller than ϵ^2. By Chebyshev's inequality, we estimate (10) to error $\pm\epsilon$ with probability at least 0.9. With that estimate, we will now use Lemma 2.4. Specifically, the entropy of $\mathcal{D}$ is exactly $\mathbf{E}_{\boldsymbol{i}\sim\mathcal{D}}[\log(1/p_{\boldsymbol{i}})]$, and we have

$$\begin{aligned}\left|\mathbf{E}_{\boldsymbol{i}\sim\mathcal{D}}[\log(1/p_{\boldsymbol{i}})]-\left(\hat{H}-\boldsymbol{Z}\right)\right| &\le \epsilon+\left|\mathbf{E}_{\boldsymbol{i}\sim\mathcal{D}}[\log(1/p_{\boldsymbol{i}})]-\left(\hat{H}-\boldsymbol{Z}\right)\right| \\ &\le \epsilon+\mathbf{E}_{\boldsymbol{i}\sim\mathcal{D}}\left[\left|\log\left(\frac{1}{p_{\boldsymbol{i}}}\right)-\mathbf{E}[\boldsymbol{\eta}_{\boldsymbol{i}}]\right|\right] \le 2\epsilon,\end{aligned}$$

where $\boldsymbol{\eta}_{\boldsymbol{i}}$ is the result of running $\texttt{LogEstimator}(\mathcal{D},\boldsymbol{i})$. □

Proof of Lemma 3.2. We note that since $\log(\cdot)$ is monotone increasing, we must have $H \ge \tilde{H}$. To see that it is not much larger, note that we always have $\log z = \ln(z)/\ln(2) \le (z-1)/\ln(2)$, which means

$$\begin{aligned}H-\tilde{H} &= \mathbf{E}_{\boldsymbol{i},\boldsymbol{X}}\left[\log(\boldsymbol{X}/\boldsymbol{X}')\right] \le \frac{1}{\ln(2)}\mathbf{E}_{\boldsymbol{i},\boldsymbol{X}}\left[\frac{\boldsymbol{X}}{\min\{\boldsymbol{X},X_{\max}\}}-1\right] \le \frac{1}{\ln(2)}\mathbf{E}_{\boldsymbol{i},\boldsymbol{X}}\left[\frac{\boldsymbol{X}}{X_{\max}}\right] \\ &= \frac{1}{X_{\max}\cdot\ln(2)}\sum_{i=1}^{k}p_i\cdot\frac{t}{p_i} = \frac{tk}{X_{\max}\cdot\ln(2)} = \epsilon.\end{aligned}$$

□

Proof of Lemma 3.4. Substituting the r_ℓ values into Lemma 3.3 ensures $\mathbf{E}\left[\mathsf{Error}^2\right] \le \epsilon^2/10$. Hence the estimator is within $\pm\epsilon$ of $\tilde{H}$ with probability 0.9 by Chebyshev's inequality.

For the intervals $\ell = \{1,\dots,L-1\}$, we always spend r_ℓ tries to determine whether a sample falls within a particular interval. Note that we take one sample to determine $\boldsymbol{i} \sim \mathcal{D}$, and then we take at most b_ℓ samples. Therefore, the sample complexity for these is

$$\sum_{\ell=1}^{L-1} r_\ell \cdot b_\ell = \frac{80tk}{\epsilon^2} \cdot \sum_{\ell=1}^{L-1} \frac{\log^2(\log^{(\ell-1)}(k)/\epsilon)}{(\log^{(\ell)} k)^3} = \frac{80tk}{\epsilon^2} \cdot \sum_{\ell=1}^{L-1} \frac{(3\log^{(\ell)}(k) + \log(1/\epsilon))^2}{(\log^{(\ell)} k)^3}$$
$$\leq kt \cdot O(\log^2(1/\epsilon)/\epsilon^2),$$

where we used the fact that

$$\sum_{\ell=1}^{L-1} \frac{1}{(\log^{(\ell)} k)} \leq \frac{1}{1} + \frac{1}{\exp(1)} + \frac{1}{\exp(\exp(1))} + \frac{1}{\exp(\exp(\exp(1)))} + \ldots = O(1)\ .$$

Finally, it remains to bound the expected sample complexity of the bucket L. Here, we note

$$r_L = \frac{O(1)}{\epsilon^2} \cdot \log^2\left(\frac{\log^{(L-1)} k}{\epsilon}\right) \leq O\left(\frac{\log^2(1/\epsilon)}{\epsilon^2}\right).$$

Therefore, the expected sample complexity for interval L is $r_L \cdot \sum_{i=1}^{k} p_i \cdot \dfrac{t}{p_i} = O(k\log^4(1/\epsilon)/\epsilon^2)$.

□

D Conjectured Lower Bound

Recall that without a memory constraint the sample complexity is known to be $n = \Theta(\max\{\epsilon^{-1} \cdot k/\log(k/\epsilon), \epsilon^{-2}\log^2 k\})$ [VV17, VV11, JVHW15, WY16]. To prove a $\Omega(k/\epsilon^2)$ lower bound for the memory constrained version, we conjecture the following randomized process can be used to generate distributions over $[2k]$ that look alike to any constant space algorithm that uses $o(k/\epsilon^2)$ samples but they have *different* entropies.

Suppose we have k Bernoulli random variables with parameter α: $Y_1, \ldots, Y_k$. And, we have k Rademacher random variables $Z_1, \ldots, Z_k$ (that are $+1$ or -1 with probability $1/2$). We construct distribution p in such a way that it is uniform over k pairs of elements $(1, 2), (3, 4), \ldots, (2k-1, 2k)$. However, conditioning on pair $(2i-1, 2i)$, we may have a constant bias based on the random variable Y_i. And, we decide about the direction of the bias based on Z_i. More precisely, we set the probabilities in p as follows:

$$p_{2i-1} = \frac{1 + Y_i \cdot Z_i/4}{2k}\,, \qquad p_{2i} = \frac{1 - Y_i \cdot Z_i/4}{2k} \qquad\qquad \forall i \in [k]\,.$$

Now, it is not hard to show that if we generate two distributions as above with $\alpha = (1+\epsilon)/2$ and $\alpha = (1-\epsilon)/2$, then their entropies are $\Theta(\epsilon)$ separated with a constant probability. Thus, any algorithm that can estimate the entropy has to *distinguish* $\alpha = (1+\epsilon)/2$ from $\alpha = (1-\epsilon)/2$. Intuitively, to learn α, we would require to *determine* $\Omega(1/\epsilon^2)$ many of Y_i's. Since we have only a constant words of memory, we cannot perform the estimation of the Y_i's in parallels. Thus, any natural algorithm would require to draw $\Omega(k/\epsilon^2)$ samples.