Supplementary Information for: Estimating the number of unseen species: A bird in the hand is worth $\log n$ in the bush

Alon Orlitsky
UCSD
alon@ucsd.edu

Ananda Theertha Suresh
Google Research
s.theertha@gmail.com

Yihong Wu
Yale University
yihong.wu@yale.edu

August 26, 2016

Contents

1 Preliminaries

Throughout the paper, we use standard asymptotic notation. For positive sequences $\{a_n\}$ and $\{b_n\}$, denote $a_n = \Theta(b_n)$ or $a_n \asymp b_n$ if $1/c \le a_n/b_n \le c$ for some universal constant $c > 0$. Let $\mathbb{1}_A$ denote the indicator random variable of an event A. Let $\mathrm{Bin}(n, p)$ denote the binomial distribution with n trials and success probability p and let $\mathrm{poi}(\lambda)$ denote the Poisson distribution with mean λ. All logarithms are with respect to the natural base unless otherwise specified.

1.1 The Poisson model

Let p be a probability distribution over a discrete set $\mathcal{X}$, namely $p_x \ge 0$ for all $x \in \mathcal{X}$ and $\sum_{x\in\mathcal{X}} p_x = 1$. Recall that for the Poisson model, the sample sizes are Poisson distributed: $N \sim \mathrm{poi}(n)$, $M \sim \mathrm{poi}(m)$, and $t = \frac{m}{n}$. We abbreviate the number of unseen symbols by

$$U \triangleq U(X^N, X_{N+1}^{N+M}),$$

and denote an estimator by $U^{\mathrm{E}} \triangleq U^{\mathrm{E}}(X^N, t)$.

Let N_x and N'_x denote the multiplicity of a symbol x in the current and future samples, respectively. Let $\lambda_x \triangleq np_x$. Then a symbol x appears $N_x \sim \mathrm{poi}(\lambda_x)$ times, and for any $i \ge 0$,

$$\mathbb{E}[\mathbb{1}_{N_x=i}] = e^{-\lambda_x}\frac{\lambda_x^i}{i!}.$$

Hence

$$\mathbb{E}[\Phi_i] = \mathbb{E}\left[\sum_x \mathbb{1}_{N_x=i}\right] = \sum_x e^{-\lambda_x}\frac{\lambda_x^i}{i!}.$$

A helpful property of Poisson sampling is that the multiplicities of different symbols are independent of each other. Therefore, for any function $f(x, i)$,

$$\mathrm{Var}\left(\sum_x f(x, N_x)\right) = \sum_x \mathrm{Var}(f(x, N_x)).$$

Many of our derivations rely on these three equations. For example,

$$\mathbb{E}[U] = \sum_x \mathbb{E}[\mathbb{1}_{N_x=0}]\cdot\mathbb{E}[\mathbb{1}_{N'_x>0}] = \sum_x e^{-\lambda_x}\cdot(1-e^{-t\lambda_x}),$$

and

$$\begin{aligned}\mathrm{Var}(U) &= \mathrm{Var}\left(\sum_x \mathbb{1}_{N_x=0}\cdot\mathbb{1}_{N'_x>0}\right) = \sum_x \mathrm{Var}\left(\mathbb{1}_{N_x=0}\cdot\mathbb{1}_{N'_x>0}\right)\\ &\le \sum_x \mathbb{E}\left[\mathbb{1}_{N_x=0}\cdot\mathbb{1}_{N'_x>0}\right] = \mathbb{E}\left[U\right].\end{aligned}$$

Note that these equations imply that the standard deviation of U is at most $\sqrt{\mathbb{E}[U]} \ll \mathbb{E}[U]$, hence U highly concentrates around its expectation, and estimating U and $\mathbb{E}[U]$ are essentially the same.

1.2 Properties of the Good-Toulmin estimator

Before proceeding with general estimators, we prove a few properties of U^{GT}. Under the Poisson model, U^{GT} is in fact the *unique* unbiased estimator for U.[1]

Lemma 3 ((8)). *For any distribution,*

$$\mathbb{E}[U] = \mathbb{E}[U^{\mathrm{GT}}].$$

Proof.

$$\begin{aligned}\mathbb{E}[U] &= \mathbb{E}\left[\sum_x \mathbb{1}_{N_x=0} \cdot \mathbb{1}_{N'_x>0}\right] = \sum_x e^{-\lambda_x} \cdot \left(1 - e^{-t\lambda_x}\right) \\ &= -\sum_x e^{-\lambda_x} \cdot \sum_{i=1}^{\infty} \frac{(-t\lambda_x)^i}{i!} = -\sum_{i=1}^{\infty} (-t)^i \cdot \sum_x e^{-\lambda_x} \frac{\lambda_x^i}{i!} \\ &= -\sum_{i=1}^{\infty} (-t)^i \cdot \mathbb{E}[\Phi_i] = \mathbb{E}[U^{\mathrm{GT}}].\end{aligned}$$

□

Even though U^{GT} is unbiased for all t, for $t > 1$ it has high variance and hence does not estimate U well even for the simplest distributions.

Lemma 4. *For any $t > 1$,*

$$\lim_{n\to\infty} \mathcal{E}_{n,t}(U^{\mathrm{GT}}) = \infty.$$

Proof. Let p be the uniform distribution over two symbols a and b, namely, $p_a = p_b = 1/2$. First consider even n. Since $(U^{\mathrm{GT}} - U)^2$ is always nonnegative,

$$\mathbb{E}[(U^{\mathrm{GT}} - U)^2] \geq \mathbb{P}(N_a = N_b = n/2)(2(-t)^{n/2})^2 = \left(e^{-n/2}\frac{(n/2)^{n/2}}{(n/2)!}\right)^2 4t^n \geq \frac{4t^n}{e^2 n},$$

where we used the fact that $k! \leq (\frac{k}{e})^k \sqrt{k}e$. Hence for any $t > 1$,

$$\lim_{n\to\infty} \frac{\mathbb{E}[(U^{\mathrm{GT}} - U)^2]}{(nt)^2} \geq \lim_{n\to\infty} \frac{4t^n}{e^2 n(nt)^2} = \infty.$$

The case of odd n can be shown similarly by considering the event $N_a = \lfloor n/2 \rfloor, N_b = \lceil n/2 \rceil$. □

2 Proofs for the Poisson model

In this section, we provide a performance guarantee for SGT estimators under the Poisson sampling model. We first prove Lemma 1 and then show that the truncated GT estimators incur a high bias. We then introduce the class of smoothed GT estimators obtained by averaging several truncated GT estimators and bound their mean squared error in Theorem 3 for an arbitrary smoothing distribution. We then apply this result to obtain NMSE bounds for Poisson and Binomial smoothing in Corollaries 1 and 2 respectively, which imply the main result Theorem 1 for the Poisson model.

[1] To establish the uniqueness, suppose $\hat{U} : \mathbb{N}^{\mathcal{X}} \to \mathbb{R}$ is an unbiased estimator for U. Then $\mathbb{E}[\hat{U}(N_1, \ldots, N_k)] = \sum_{i\in\mathbb{N}^{\mathcal{X}}} \hat{U}(i) \prod_{x\in\mathcal{X}} \frac{e^{-\lambda_x}\lambda_x^{i_x}}{i_x!} = \sum_x e^{-\lambda_x}(1-e^{-t\lambda_x}) = \sum_x e^{-\lambda_x} \sum_{i\geq 1} -(-t)^i \frac{\lambda_x^i}{i!} = \sum_{i\in\mathbb{N}^{\mathcal{X}}} \sum_x -(t)^{i_x} \prod_{x\in\mathcal{X}} \frac{e^{-\lambda_x}}{i_x!}$. Since this holds for any $\lambda_x \geq 0$, we have $\hat{U}(i) = \sum_x -(t)^{i_x}$, that is, $\hat{U}$ is the GT estimator.

2.1 Bounds for general linear estimators

We provide the proof of Lemma 1 in the main article.

Proof of Lemma 1. Note that

$$\begin{aligned} U^{\mathrm{h}} - U &= \sum_{i=1}^{\infty} \Phi_i h_i - \sum_x \mathbb{1}_{N_x=0} \cdot \mathbb{1}_{N'_x>0} \\ &= \sum_{i=1}^{\infty} \sum_x \mathbb{1}_{N_x=i} \cdot h_i - \sum_x \mathbb{1}_{N_x=0} \cdot \mathbb{1}_{N'_x>0} \\ &= \sum_x \left(\sum_{i=1}^{\infty} \mathbb{1}_{N_x=i} \cdot h_i - \mathbb{1}_{N_x=0} \cdot \mathbb{1}_{N'_x>0} \right). \end{aligned}$$

For every symbol x,

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^{\infty} \mathbb{1}_{N_x=i} \cdot h_i - \mathbb{1}_{N_x=0} \cdot \mathbb{1}_{N'_x>0} \right] &= \sum_{i=1}^{\infty} e^{-\lambda_x} \frac{\lambda_x^i}{i!} \cdot h_i - e^{-\lambda_x} \cdot (1 - e^{-t\lambda_x}) \\ &= e^{-\lambda_x} \left(\sum_{i=1}^{\infty} \frac{\lambda_x^i h_i}{i!} - (1 - e^{-t\lambda_x}) \right) \\ &= e^{-\lambda_x} \left(h(\lambda_x) - (1 - e^{-t\lambda_x}) \right), \end{aligned}$$

from which the bias result follows. For the variance, observe that for every symbol x,

$$\begin{aligned} \mathrm{Var}\left(\sum_{i=1}^{\infty} \mathbb{1}_{N_x=i} \cdot h_i - \mathbb{1}_{N_x=0} \cdot \mathbb{1}_{N'_x>0} \right) &\le \mathbb{E}\left[\left(\sum_{i=1}^{\infty} \mathbb{1}_{N_x=i} \cdot h_i - \mathbb{1}_{N_x=0} \cdot \mathbb{1}_{N'_x>0} \right)^2 \right] \\ &\overset{(a)}{=} \mathbb{E}\left[\sum_{i=1}^{\infty} \mathbb{1}_{N_x=i} h_i^2 \right] + \mathbb{E}[\mathbb{1}_{N_x=0}] \cdot \mathbb{E}[\mathbb{1}_{N'_x>0}] \\ &= \sum_{i=1}^{\infty} \mathbb{E}[\mathbb{1}_{N_x=i}] \cdot h_i^2 + \mathbb{E}[\mathbb{1}_{N_x=0}] \cdot \mathbb{E}[\mathbb{1}_{N'_x>0}], \end{aligned}$$

where (a) follows as for every $i \neq j$, $\mathbb{E}[\mathbb{1}_{N_x=i}\mathbb{1}_{N_x=j}] = 0$. Since the variance of a sum of independent random variables is the sum of their variances,

$$\begin{aligned} \mathrm{Var}(U^{\mathrm{h}} - U) &\le \sum_x \sum_{i=1}^{\infty} \mathbb{E}[\mathbb{1}_{N_x=i}] h_i^2 + \sum_x \mathbb{E}[\mathbb{1}_{N_x=0}] \cdot \mathbb{E}[\mathbb{1}_{N'_x>0}] \\ &= \sum_{i=1}^{\infty} \mathbb{E}[\Phi_i] \cdot h_i^2 + \mathbb{E}[U] \\ &\le \mathbb{E}[\Phi_+] \cdot \sup_{i \ge 1} h_i^2 + \mathbb{E}[U]. \end{aligned}$$

□

2.2 Negative result for truncated Good-Toulmin estimator

We provide the proof of Lemma 2 in the main article.

Proof of Lemma 2. To rigorously prove an impossibility result for the truncated GT estimator, we demonstrate a particular distribution under which the bias is large. Consider the uniform distribution over $n/(\ell+1)$ symbols, where ℓ is a non-zero even integer. By Lemma 1, for this distribution the bias is

$$\begin{aligned}
\mathbb{E}[U - U^\ell] &= \sum_x e^{-\lambda_x}(1 - e^{-\lambda_x t} - h(\lambda_x)) \\
&= \frac{n}{\ell+1} e^{-(\ell+1)} \left(1 - e^{-(\ell+1)t} + \sum_{i=1}^{\ell} \frac{(-(\ell+1)t)^i}{i!}\right) \\
&\geq \frac{n}{\ell+1} e^{-(\ell+1)} \left(\sum_{i=1}^{\ell} \frac{(-(\ell+1)t)^i}{i!}\right) \\
&\overset{(a)}{\geq} \frac{n}{\ell+1} e^{-(\ell+1)} \left(\frac{((\ell+1)t)^\ell}{\ell!} - \frac{((\ell+1)t)^{\ell-1}}{(\ell-1)!}\right) \\
&\geq \frac{n}{(\ell+1)} e^{-(\ell+1)} \frac{((\ell+1)t)^\ell}{\ell!} \cdot \frac{(t-1)}{t} \\
&\geq \frac{n}{3(\ell+1)^{3/2}} t^\ell \frac{(t-1)}{t} \geq \frac{n}{3 \cdot 2^{3/2}} \frac{t^\ell}{\ell^{3/2}} \frac{(t-1)}{t},
\end{aligned}$$

where (a) follows from the fact that $\frac{(-(\ell+1)t)^i}{i!}$ for $i = 1, \ldots, \ell$ is an alternating series with increasing magnitude of terms. Hence

$$\mathbb{E}[U - U^\ell] \geq \frac{n}{3 \cdot 2^{3/2}} \frac{(t-1)}{t} \min_{\ell \in \{2,4,\ldots\}} \frac{t^\ell}{\ell^{3/2}}.$$

For $t \geq 2$, the above minimum occurs at $\ell = 2$ and hence $\min_{\ell \in \{2,4,\ldots\}} \frac{t^\ell}{\ell^{3/2}} \geq \frac{(t-1)^{3/2}}{2^{3/2}}$. For $1 < t < 2$, using the fact that $e^y \geq ey$ for $y > 0$ and $\log t \geq (t-1)\log 2$ for $1 < t < 2$, we have $\min_{\ell \in \{2,4,\ldots\}} \frac{t^\ell}{\ell^{3/2}} \geq (\frac{2e \log t}{3})^{3/2} \geq (\frac{2e \log 2(t-1)}{3})^{3/2}$. Thus for any even value of $\ell > 0$,

$$\mathbb{E}[U - U^\ell] \geq \frac{n(t-1)^{5/2}}{6.05t}.$$

A similar argument holds for odd values of ℓ and $\ell = 0$, showing that $|\mathbb{E}[U - U^\ell]| \gtrsim \frac{n(t-1)^{5/2}}{t}$ and hence the desired NMSE bound. □

2.3 Bounds on SGT estimators: arbitrary smoothing

Here we prove Theorem 3 in the main article on the NMSE of SGT estimator for an arbitrary smoothing distribution. The proof consists of bounds on bias (Lemma 7) and variance (Lemma 5).

Lemma 5. *For a random variable L over $\mathbb{Z}_+$ and $t \geq 1$,*

$$\mathrm{Var}(U^{\mathrm{L}} - U) \leq \mathbb{E}[\Phi_+] \cdot \mathbb{E}^2[t^L] + \mathbb{E}[U].$$

Proof. By Lemma 1, to bound the variance it suffices to bound the highest coefficient in h^{L}.

$$|h_i^{\mathrm{L}}| \le t^i \mathbb{P}(L \ge i) = t^i \sum_{j=i}^{\infty} \mathbb{P}(L=j) \le \sum_{j=i}^{\infty} \mathbb{P}(L=j) t^j \le \mathbb{E}[t^L]. \qquad [7]$$

The above bound together with Lemma 1 yields the result. □

To bound the bias, we need few definitions. For any random variable L over $\mathbb{Z}_+$, let

$$g(y) \triangleq -\sum_{i=1}^{\infty} \frac{\mathbb{P}(L \ge i)}{i!} (-y)^i. \qquad [8]$$

Under this definition, $h^{\mathrm{L}}(y) = g(yt)$. The following auxiliary lemma bounds the bias.

Lemma 6. *For any random variable L over $\mathbb{Z}_+$,*

$$g(y) - (1 - e^{-y}) = -e^{-y} \int_0^y \mathbb{E}\left[\frac{(-s)^L}{L!}\right] e^s \, ds.$$

Proof. Subtracting [8] from the Taylor series expansion of $1 - e^{-y}$,

$$\begin{aligned} g(y) - (1 - e^{-y}) &= \sum_{i=1}^{\infty} \frac{\mathbb{P}(L < i)}{i!} (-y)^i \\ &= \sum_{i=1}^{\infty} \sum_{j=0}^{i-1} \frac{(-y)^i}{i!} \mathbb{P}(L=j) \\ &= \sum_{j=0}^{\infty} \left(\sum_{i=j+1}^{\infty} \frac{(-y)^i}{i!} \right) \mathbb{P}(L=j). \end{aligned}$$

By the incomplete Gamma function,

$$\sum_{i=j+1}^{\infty} \frac{z^i}{i!} = \frac{e^z}{j!} \int_0^z \tau^j e^{-\tau} \mathrm{d}\tau.$$

Thus by Fubini's theorem,

$$\begin{aligned} g(y) - (1 - e^{-y}) &= \sum_{j=0}^{\infty} \frac{e^{-y}}{j!} \int_0^{-y} \tau^j e^{-\tau} \mathrm{d}\tau \mathbb{P}(L=j) \\ &= e^{-y} \int_0^{-y} e^{-\tau} \mathrm{d}\tau \left(\sum_{j=0}^{\infty} \frac{\tau^j}{j!} \mathbb{P}(L=j) \right) \\ &= -e^{-y} \int_0^{y} e^{s} \mathrm{d}s \left(\sum_{j=0}^{\infty} \frac{(-s)^j}{j!} \mathbb{P}(L=j) \right) \\ &= -e^{-y} \int_0^{y} \mathbb{E}\left[\frac{(-s)^L}{L!}\right] e^s \mathrm{d}s. \end{aligned}$$

□

To bound the bias, we need one more definition. For a random variable L over $\mathbb{Z}_+$, let

$$\xi_L(t) \triangleq \max_{0\le s<\infty} \left|\mathbb{E}\left[\frac{(-s)^L}{L!}\right]\right| e^{-s/t},$$

Lemma 7. *For a random variable L over $\mathbb{Z}_+$,*

$$|\mathbb{E}[U^{\mathrm{L}} - U]| \le (\mathbb{E}[\Phi_+] + \mathbb{E}[U]) \cdot \xi_L(t).$$

Proof. By Lemma 6,

$$\begin{aligned}
|g(y) - (1 - e^{-y})| &\le e^{-y} \int_0^y \left|\mathbb{E}\left[\frac{(-s)^L}{L!}\right]\right| e^s \mathrm{d}s \\
&\le \max_{s\le y} \left|\mathbb{E}\left[\frac{(-s)^L}{L!}\right]\right| e^{-y} \int_0^y e^s \mathrm{d}s \\
&= \max_{s\le y} \left|\mathbb{E}\left[\frac{(-s)^L}{L!}\right]\right| (1 - e^{-y}).
\end{aligned}$$

For a symbol x,

$$e^{-\lambda_x}\left(h^{\mathrm{L}}(\lambda_x) - (1 - e^{-\lambda_x t})\right) = e^{-\lambda_x}\left(g(\lambda_x t) - (1 - e^{-\lambda_x t})\right).$$

Hence,

$$\begin{aligned}
|e^{-\lambda_x}\left(h^{\mathrm{L}}(\lambda_x) - (1 - e^{-\lambda_x t})\right)| &\le (1 - e^{-\lambda_x t}) \max_{0\le y\le\infty} e^{-y} \max_{0\le s\le yt} \left|\mathbb{E}\left[\frac{(-s)^L}{L!}\right]\right| \\
&\le (1 - e^{-\lambda_x t}) \max_{0\le s\le\infty} \left|\mathbb{E}\left[\frac{(-s)^L}{L!}\right]\right| e^{-s/t}.
\end{aligned}$$

The lemma follows by summing over all the symbols and substituting $\sum_x 1 - e^{-\lambda_x t} \le \sum_x 1 - e^{-\lambda_x (t+1)} = \mathbb{E}[\Phi_+] + \mathbb{E}[U]$. □

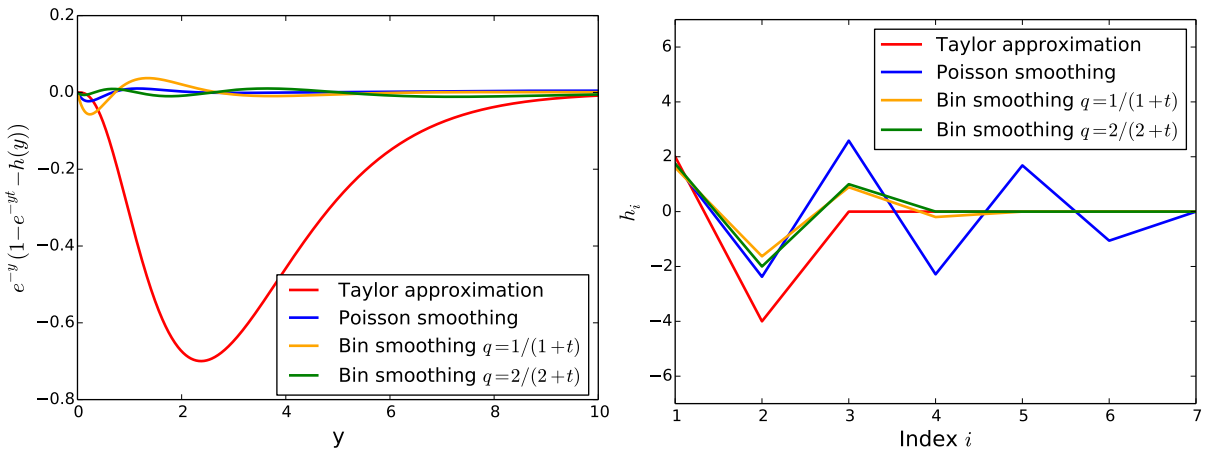
The effectiveness of SGT estimators can also be demonstrated in terms of their approximation performance. As shown in Figure 5(a), the Poisson and Binomial smoothing have significantly smaller approximation error compared to the Taylor series approximation, leading to reduced bias. The coefficients of the resulting estimator is plotted in Figure 5(b). It is easy to see that the maximum magnitude of the coefficients is also lower for the smoothed estimators, resulting in smaller variance. In the following sections, we particularize the main theorem for Poisson and binomial smoothings.

2.4 Poisson smoothing

Corollary 1. *For $t \ge 1$, $L \sim \mathrm{poi}(r)$ with $r = \frac{1}{2t}\log\left(\frac{n(t+1)^2}{t-1}\right)$,*

$$\mathcal{E}_{n,t}(U^{\mathrm{L}}) \le \frac{c_t}{n^{1/t}},$$

where $0 \le c_t \le 3$ and $\lim_{t\to\infty} c_t = 1$.



(a) (b)

Figure 5: Comparisons of approximations of $h^{\mathrm{L}}(\cdot)$ with $\mathbb{E}[L] = 2$ and $t = 2$. (a) $e^{-y}(1-e^{-yt}-h^{\mathrm{L}}(y))$ as a function of y. (b) Coefficients h_i^{L} as a function of index i.

Proof. For $L \sim \mathrm{poi}(r)$,

$$\mathbb{E}[t^L] = e^{-r} \sum_{\ell=0}^{\infty} \frac{(rt)^\ell}{\ell!} = e^{r(t-1)}. \qquad [9]$$

Furthermore,

$$\mathbb{E}\left[\frac{(-s)^L}{L!}\right] = e^{-r} \sum_{j=0}^{\infty} \frac{(-sr)^j}{(j!)^2} = e^{-r} J_0(2\sqrt{sr}),$$

where J_0 is the Bessel function of first order which takes values in $[-1, 1]$ cf. (1, 9.1.60). Therefore

$$\xi_L(t) \le e^{-r}. \qquad [10]$$

Equations [9] and [10] together with Theorem 3 yields

$$\mathbb{E}[(U^{\mathrm{L}} - U)^2] \le \mathbb{E}[\Phi_+] \cdot e^{2r(t-1)} + \mathbb{E}[U] + (\mathbb{E}[\Phi_+] + \mathbb{E}[U])^2 \cdot e^{-2r}.$$

Since $\mathbb{E}[\Phi_+] \le n$ and $\mathbb{E}[U] \le nt$,

$$\mathbb{E}[(U^{\mathrm{L}} - U)^2] \le ne^{2r(t-1)} + nt + (n + nt)^2 e^{-2r}.$$

Choosing $r = \frac{1}{2t} \log \frac{n(t+1)^2}{t-1}$,

$$\mathcal{E}_{n,t}(U^{\mathrm{L}}) \le \frac{1}{(nt)^{1/t}} \cdot \left(\frac{t(t-1)}{(t+1)^2}\right)^{\frac{1-t}{t}} + \frac{1}{nt},$$

and the lemma with $c_t \triangleq \frac{1}{t^{1/t}} \cdot \left(\frac{t(t-1)}{(t+1)^2}\right)^{\frac{1-t}{t}} + \frac{1}{t}$. □

2.5 Binomial smoothing

We now prove the results when $L \sim \mathrm{Bin}(k, q)$. Our analysis holds for all $q \in [0, 2/(2+t)]$ and in this range, the performance of the estimator improves as q increases, and hence the NMSE bounds are strongest for $q = 2/(2+t)$. Therefore, we consider binomial smoothing for two cases: the Efron-Thisted suggested value $q = 1/(1+t)$ and the optimized value $q = 2/(2+t)$.

Corollary 2. *For $t \geq 1$ and $L \sim \mathrm{Bin}(k, q)$, if $k = \left\lfloor \frac{1}{2} \log_2 \frac{nt^2}{t-1} \right\rfloor$ and $q = \frac{1}{t+1}$, then*

$$\mathcal{E}_{n,t}(U^{\mathrm{L}}) \leq \frac{c_t}{n^{\log_2(1+1/t)}},$$

where c_t satisfies $0 \leq c_t \leq 4$ and $\lim_{t\to\infty} c_t = 1$; if $k = \left\lfloor \frac{1}{2} \log_3 \frac{nt^2}{t-1} \right\rfloor$ and $q = \frac{2}{t+2}$, then

$$\mathcal{E}_{n,t}(U^{\mathrm{L}}) \leq \frac{c'_t}{(nt)^{\log_3(1+2/t)}},$$

where c'_t satisfies $0 \leq c'_t \leq 7$ and $\lim_{t\to\infty} c'_t = 1$.

Proof. If $L \sim \mathrm{Bin}(k, q)$,

$$\mathbb{E}[t^L] = \sum_{\ell=0}^{k} \binom{k}{\ell} (tq)^\ell (1-q)^{k-\ell} = (1 + q(t-1))^k.$$

Furthermore,

$$\mathbb{E}\left[\frac{(-s)^L}{L!}\right] = \sum_{j=0}^{k} \frac{(-s)^j}{j!} \binom{k}{j} (q)^j (1-q)^{k-j} = (1-q)^k L_k\left(\frac{qs}{1-q}\right),$$

where

$$L_k(y) = \sum_{j=0}^{k} \frac{(-y)^j}{j!} \binom{k}{j} \tag{11}$$

is the Laguerre polynomial of degree k. If $\frac{tq}{2(1-q)} \leq 1$, for any $s \geq 0$,

$$e^{-\frac{s}{t}} \left| \mathbb{E}\left[\frac{(-s)^L}{L!}\right] \right| \leq (1-q)^k e^{-\frac{s}{t}} e^{\frac{qs}{2(1-q)}} \leq (1-q)^k,$$

where the first inequality follows from the fact cf. (1, 22.14.12) that for all $y \geq 0$ and all $k \geq 0$,

$$|L_k(y)| \leq e^{y/2}. \tag{12}$$

Hence for $q \leq 2/(t+2)$,

$$\mathbb{E}[(U^{\mathrm{L}} - U)^2] \leq \mathbb{E}[\Phi_+] \cdot (1 + q(t-1))^{2k} + \mathbb{E}[U] + (\mathbb{E}[\Phi_+] + \mathbb{E}[U])^2 \cdot (1-q)^{2k}.$$

Since $\mathbb{E}[U] \leq nt$ and $\mathbb{E}[\Phi_+] \leq n$,

$$\mathbb{E}[(U^{\mathrm{L}} - U)^2] \leq n \cdot (1 + q(t-1))^{2k} + nt + (nt+n)^2 \cdot (1-q)^{2k}. \tag{13}$$

Substituting the Efron-Thisted suggested $q = \frac{1}{t+1}$ results in

$$\mathcal{E}_{n,t}(U^{\mathrm{L}}) \le \left(\frac{2^{2k}}{nt^2} + \frac{(t+1)^2}{t^2}\right)\left(\frac{t}{t+1}\right)^{2k} + \frac{1}{nt}.$$

Choosing $k = \left\lfloor \frac{1}{2}\log_2 \frac{nt^2}{t-1} \right\rfloor$ yields the first result with $c_t \triangleq \left(\frac{1}{t-1} + \left(\frac{t+1}{t}\right)^4\right) \cdot \left(\frac{t-1}{t^2}\right)^{\log_2(1+1/t)} + \frac{1}{t}$. For the second result, substituting $q = \frac{2}{t+2}$ in [13] results in

$$\mathcal{E}_{n,t}(U^{\mathrm{L}}) \le \left(\frac{3^{2k}}{nt^2} + \frac{(t+1)^2}{t^2}\right)\left(\frac{t}{t+2}\right)^{2k} + \frac{1}{nt}.$$

Choosing $k = \left\lfloor \frac{1}{2}\log_3 \frac{nt^2}{t-1} \right\rfloor$ yields the result with $c'_t \triangleq \left(\frac{1}{t-1} + \frac{(t+1)^2}{t^2}\left(\frac{t+2}{t}\right)^2\right) \cdot \left(\frac{t-1}{t^2}\right)^{\log_3(1+2/t)} + \frac{1}{t}$. □

In terms of the exponent, the result is strongest for $L \sim \mathrm{Bin}(k, 2/(t+2))$. Hence, we state the following asymptotic result, which is a direct consequence of Corollary 2:

Corollary 3. *For $L \sim \mathrm{Bin}(k,q)$, $q = \frac{2}{t+2}$,$k = \lfloor \log_3(\frac{nt^2}{t-1}) \rfloor$, and any fixed δ, the maximum t till which U^{L} incurs a NMSE of δ is*

$$\lim_{n\to\infty} \frac{\max\{t : \mathcal{E}_{n,t}(U^{\mathrm{L}}) < \delta\}}{\log n} \ge \frac{2}{\log 3 \cdot \log \frac{1}{\delta}}.$$

Proof. By Corollary 2, if $t \to \infty$, then

$$\mathcal{E}_{n,t}(U^{\mathrm{L}}) \le (1+o(1))n^{-\frac{2+o(1)}{t\log 3}}.$$

where $o(1) = o_t(1)$ is uniform in n. Consequently, if $t = (\alpha + o(1))\log n$ and $n \to \infty$, then

$$\limsup_{n\to\infty} \mathcal{E}_{n,t}(U^{\mathrm{L}}) \le e^{-\frac{2}{\alpha \log 3}}.$$

Thus for any fixed δ, the maximum t till which U^{L} incurs a NMSE of δ is

$$\lim_{n\to\infty} \frac{\max\{t : \mathcal{E}_{n,t}(U^{\mathrm{L}}) < \delta\}}{\log n} \ge \frac{2}{\log 3 \cdot \log \frac{1}{\delta}}.$$ □

Corollaries 1 and 2 imply Theorem 1 for the Poisson model.

3 Extensions to other models

Our results so far have been developed for the Poisson model. Next we extend them to the multinomial model (fixed sample size), the Bernoulli-product model, and the hypergeometric model (sampling without replacement) (4), for which upper bounds of NMSE for general smoothing distributions that are analogous to Theorem 3 are presented in Theorem 4, 5 and 6, respectively. Using these results, we obtain the NMSE for Poisson and Binomial smoothings similar to Corollaries 1 and 2. We remark that up to multiplicative constants, the NMSE under multinomial and Bernoulli-product model are similar to those of Poisson model; however, the NMSE under hypergeometric model is slightly larger.

3.1 The multinomial model

The multinomial model corresponds to the setting described in the introduction, where upon observing n i.i.d. samples, the objective is to estimate the expected number of new symbols $U(X^n, X_{n+1}^{n+m})$ that would be observed if we took m more samples. We can write the expected number of new symbols as

$$U(X^n, X_{n+1}^{n+m}) = \sum_x \mathbb{1}_{N_x=0} \cdot \mathbb{1}_{N'_x>0}.$$

As before we abbreviate

$$U \triangleq U(X^n, X_{n+1}^{n+m})$$

and similarly $U^{\mathrm{E}} \triangleq U^{\mathrm{E}}(X^n, t)$ for any estimator E. The difficulty in handling multinomial distributions is that, unlike the Poisson model, the number of occurrences of symbols are correlated; in particular, they sum up to n. This dependence renders the analysis cumbersome. In the multinomial setting each symbol is distributed according to $\mathrm{Bin}(n, p_x)$ and hence

$$\mathbb{E}[\mathbb{1}_{N_x=i}] = \binom{n}{i} p_x^i (1-p_x)^{n-i}.$$

As an immediate consequence,

$$\mathbb{E}[\Phi_i] = \mathbb{E}\left[\sum_x \mathbb{1}_{N_x=i}\right] = \sum_x \binom{n}{i} p_x^i (1-p_x)^{n-i}.$$

We now bound the bias and variance of an arbitrary linear estimator U^{h}. We first show that the bias $\mathbb{E}[U^{\mathrm{h}} - U]$ under the multinomial model is close to that under the Poisson model, which is $\sum_x e^{-\lambda_x}(h(\lambda_x) - (1 - e^{-t\lambda_x}))$ as given in Lemma 1.

Lemma 8. *The bias of $U^{\mathrm{h}} = \sum_{i=1}^{\infty} \Phi_i h_i$ satisfies*

$$\left|\mathbb{E}[U^{\mathrm{h}} - U] - \sum_x e^{-\lambda_x}\left(h(\lambda_x) - (1 - e^{-t\lambda_x})\right)\right| \le 2 \sup_i |h_i| + 2.$$

Proof. First we recall a result on Poisson approximation: For $X \sim \mathrm{Bin}(n, p)$ and $Y \sim \mathrm{poi}(np)$,

$$|\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]| \le 2p \sup_i |f(i)|, \qquad [14]$$

which follows from the total variation bound $d_{\mathrm{TV}}(\mathrm{Bin}(n,p), \mathrm{poi}(np)) \le p$ (2, Theorem 1) and the fact that $d_{\mathrm{TV}}(\mu, \nu) = \frac{1}{2}\sup_{\|f\|_\infty \le 1} \int f d\mu - \int f d\nu$. In particular, taking $f(x) = \mathbb{1}_{x=0}$ gives

$$0 \le e^{-np} - (1-p)^n \le 2p.$$

Note that the linear estimator can be expressed as $U^{\mathrm{h}} = \sum_x h_{N_x}$. Under the multinomial model,

$$\mathbb{E}[U^{\mathrm{h}} - U] = \sum_x \mathbb{E}_{N_x \sim \mathrm{Bin}(n,p_x)}[h_{N_x}] - \sum_x (1-p_x)^n (1 - (1-p_x)^m).$$

Under the Poisson model,

$$\sum_x e^{-\lambda_x}\left(h(\lambda_x) - (1 - e^{-t\lambda_x})\right) = \sum_x \mathbb{E}_{N_x \sim \mathrm{poi}(np_x)}[h_{N_x}] - \sum_x e^{-np_x}(1 - e^{-mp_x}).$$

Then

$$\left|\sum_x \mathbb{E}_{N_x\sim\mathrm{Bin}(n,p_x)}[h_{N_x}] - \sum_x \mathbb{E}_{N_x\sim\mathrm{poi}(np_x)}[h_{N_x}]\right| \overset{[14]}{\leq} 2\sup_i |h_i| \sum_x p_x = 2\sup_i |h_i|.$$

Furthermore,

$$\begin{aligned}&\sum_x (1-p_x)^n(1-(1-p_x)^m) - \sum_x e^{-np_x}(1-e^{-mp_x})\\ \leq &\sum_x e^{-np_x}(e^{-mp_x} - (1-p_x)^m) \overset{[14]}{\leq} \sum_x e^{-np_x} 2p_x \leq 2.\end{aligned}$$

Similarly, $\sum_x (1-p_x)^n(1-(1-p_x)^m) - \sum_x e^{-np_x}(1-e^{-mp_x}) \geq -2$. Assembling the above proves the lemma. □

The next result bounds the variance.

Lemma 9. *For any linear estimator U^{h},*

$$\mathrm{Var}(U^{\mathrm{h}} - U) \leq 8n \max\left\{\sup_{i\geq 1} h_i^2, 1\right\} + 8m.$$

Proof. Recognizing that $U^{\mathrm{h}} - U$ is a function of $n+m$ independent random variables, namely, $X_1, \ldots, X_{n+m}$ drawn i.i.d. from p, we apply Steele's variance inequality (11) to bound its variance. Similar to [3.1],

$$U^{\mathrm{h}} - U = \sum_x h_{N_x} + \mathbb{1}_{N_x=0}\mathbb{1}_{N'_x>0}$$

Changing the value of any one of the first n samples changes the multiplicities of two symbols, and hence the value of $U^{\mathrm{h}} - U$ can change by at most $4\max(\max_{i\geq 1}|h_i|, 1)$. Similarly, changing any one of the last m samples changes the value of $U^{\mathrm{h}} - U$ by at most four. Applying Steele's inequality gives the lemma. □

Lemmas 8 and 9 are analogous to Lemma 1. Together with [7] and Lemma 7, we obtain the main result for the multinomial model.

Theorem 4. *For $t \geq 1$ and any random variable L over $\mathbb{Z}_+$,*

$$\mathbb{E}[(U^{\mathrm{L}} - U)^2] \leq 8n\,\mathbb{E}^2[t^L] + 8m + \left((n(t+1)\xi_L(t) + 2\mathbb{E}[t^L] + 2\right)^2.$$

Similar to Corollaries 1 and 2, one can compute the NMSE for Binomial and Poisson smoothings. We remark that up to multiplicative constants the results are identical to those for the Poisson model.

3.2 Bernoulli-product model

Consider the following species assemblage model. There are k distinct species and each one can be found in one of n independent sampling units. Thus every species can be present in multiple sampling units simultaneously and each sampling unit can capture multiple species. For example species x can be found in sampling units $1, 3$ and 5 and species y can be found in units $2, 3$, and 4.

Given the data collected from n sampling units, the objective is to estimate the expected number of new species that would be observed if we placed m more units.

The aforementioned problem is typically modeled as by the *Bernoulli-product model*. Since, in this model each sample only has presence-absence data, it is often referred to as incidence model (6). For notational simplicity, we use the same notation as the other three models. In Bernoulli-product model, for a symbol x, N_x denotes the number of sampling units in which x appears and Φ_i denotes the number of symbols that appeared in i sampling units. Given a set of distinct symbols (potentially infinite), each symbol x is observed in each sampling unit independently with probability p_x and the observations from each sampling unit are independent of each other. To distinguish from the multinomial and Poisson sampling models where each sample can be only one symbol, we refer to samples here as sampling units. Given the results of n sampling units, the goal is to estimate the expected number of new symbols that would appear in the next m sampling units. Let $p_{_S} = \sum_x p_x$. Note that $p_{_S}$ is also the expected number of symbols that we observe for each sampling unit and need not sum to 1. For example, in the species application, probability of catching bumble bee can be 0.5 and honey bee be 0.7.

This model is significantly different from the multinomial model in two ways. Firstly, here given n sampling units the number of occurrences of symbols are independent of each other. Secondly, $p_{_S} \triangleq \sum_x p_x$ need not be 1. In the Bernoulli-product model, the probability observing each symbol at a particular sample is p_x and hence in n samples, the number of occurrences is distributed $\text{Bin}(n, p_x)$. Therefore the probability that x is be observed in i sampling units is

$$\mathbb{E}[\mathbb{1}_{N_x=i}] = \binom{n}{i} p_x^i (1-p_x)^{n-i},$$

and an immediate consequence on the number of distinct symbols that appear i sampling units is

$$\mathbb{E}[\Phi_i] = \mathbb{E}\left[\sum_x \mathbb{1}_{N_x=i}\right] = \sum_x \binom{n}{i} p_x^i (1-p_x)^{n-i}.$$

Furthermore, the expected total number of symbols is $np_{_S}$ and hence

$$\sum_{i=1}^{n} \mathbb{E}[\Phi_i] i = np_{_S}.$$

Under the Bernoulli-product model the objective is to estimate the number of new symbols that we observe in m more sampling units and is

$$U(X^n, X_{n+1}^{n+m}) = \sum_x \mathbb{1}_{N_x=0} \cdot \mathbb{1}_{N'_x>0}.$$

As before, we abbreviate

$$U \triangleq U(X^n, X_{n+1}^{n+m})$$

and similarly $U^{\text{E}} \triangleq U^{\text{E}}(X^n, t)$ for any estimator E. Since the probabilities need not add up to 1, we redefine our definition of $\mathcal{E}_{n,t}(U^{\text{E}})$ as

$$\mathcal{E}_{n,t}(U^{\text{E}}) \triangleq \max \mathbb{E}_p \left(\frac{U - U^{\text{E}}}{ntp_{_S}}\right)^2.$$

Under this model, the SGT estimator satisfy similar results to that of Corollaries 1 and 2, up to multiplicative constants. The main ingredient is to bound the bias and variance (like Lemma 1). We note that since the marginal of N_x is $\mathrm{Bin}(n, p_x)$ under both the multinomial and the Bernoulli-product model, the bias bound follows entirely analogously as in Lemma 8. The proof of variance bound is very similar to that of Lemma 1 and hence is omitted.

Lemma 10. *The linear estimator U^{h} has bias*

$$\left|\mathbb{E}[U^{\mathrm{h}} - U] - \sum_x e^{-\lambda_x}\left(h(\lambda_x) - (1 - e^{-t\lambda_x})\right)\right| \le 2p_{\mathcal{S}}\left(\sup_i |h_i| + 1\right),$$

and the variance

$$\mathrm{Var}(U^{\mathrm{h}} - U) \le np_{\mathcal{S}} \cdot \left(t + \sup_{i \ge 1} h_i^2\right).$$

The above lemma together with [7] and Lemma 7 yields the main result for the Bernoulli-product model.

Theorem 5. *For any random variable L over $\mathbb{Z}_+$ and $t \ge 1$,*

$$\mathbb{E}[(U^{\mathrm{L}} - U)^2] \le np_{\mathcal{S}} \cdot \left(t + \mathbb{E}^2[t^L]\right) + (n(t+1)p_{\mathcal{S}}\xi_L(t) + 2p_{\mathcal{S}}(\mathbb{E}[t^L] + 1))^2.$$

Similar to Corollaries 1 and 2, one can compute the normalized mean squared loss for Binomial and Poisson smoothings. We remark that up to multiplicative constants the results would be similar to that for the Poisson model.

3.3 The hypergeometric model

The hypergeometric model considers the population estimation problem with samples drawn without replacement. Given n samples drawn uniformly at random, without replacement from a set $\{y_1, \ldots, y_R\}$ of R symbols, the objective is to estimate the number of new symbols that would be observed if we had access to m more random samples without replacement, where $n + m \le R$. Unlike the Poisson, multinomial, and Bernoulli-product models we have considered so far, where the samples are independently and identically distributed, in the hypergeometric model the samples are *dependent* hence a modified analysis is needed.

Let $r_x \triangleq \sum_{i=1}^{R} \mathbb{1}_{y_i = x}$ be the number of occurrences of symbol x in the R symbols, which satisfies $\sum_x r_x = R$. Denote by N_x the number of times x appears in the n samples drawn without replacements, which is distributed according to the hypergeometric distribution $\mathrm{Hyp}(R, r_x, n)$ with the following probability mass function:[2]

$$\mathbb{P}(N_x = i) = \frac{\binom{r_x}{i}\binom{R - r_x}{n - i}}{\binom{R}{n}}.$$

We also denote the joint distribution of $\{N_x\}$, which is multivariate hypergeometric, by $\mathrm{Hyp}(\{r_x\}, n)$. Consequently,

$$\mathbb{E}[\Phi_i] = \sum_x \mathbb{P}(N_x = i) = \sum_x \frac{\binom{r_x}{i}\binom{R - r_x}{n - i}}{\binom{R}{n}}.$$

[2] We adopt the convention that $\binom{n}{k} = 0$ for all $k < 0$ and $k > n$ throughout.

Furthermore, conditioned on $N_x = 0$, N'_x is distributed as $\mathrm{Hyp}(R-n, r_x, m)$ and hence

$$\mathbb{E}[U] = \sum_x \mathbb{E}[\mathbb{1}_{N_x=0}] \cdot \mathbb{E}[\mathbb{1}_{N'_x>0} | \mathbb{1}_{N_x=0}] = \sum_x \frac{\binom{R-r_x}{n}}{\binom{R}{n}} \cdot \left(1 - \frac{\binom{R-n-r_x}{m}}{\binom{R-n}{m}}\right). \tag{15}$$

As before, we abbreviate

$$U \triangleq U(X^n, X_{n+1}^{n+m})$$

which we want to estimate and similarly for any estimator $U^{\mathrm{E}} \triangleq U^{\mathrm{E}}(X^n, t)$. We now bound the variance and bias of a linear estimator U^{h} under the hypergeometric model.

Lemma 11. *For any linear estimator U^{h},*

$$\mathrm{Var}(U^{\mathrm{h}} - U) \le 12n \sup_i h_i^2 + 6n + 3m.$$

Proof. We first note that for a random variable Y that lies in the interval $[a, b]$,

$$\mathrm{Var}(Y) \le \frac{(a-b)^2}{4}.$$

For notational convenience define $h_0 = 0$. Then $U^{\mathrm{h}} = \sum_x h_{N_x}$. Let $Z = \sum \mathbb{1}_{N_x=0}$ and $Z' = \sum \mathbb{1}_{N_x=N'_x=0}$ denote the number of unobserved symbols in the first n samples and the total $n+m$ samples, respectively. Then $U = Z - Z'$. Since the collection of random variables $\mathbb{1}_{N_x=0}$ indexed by x are negatively correlated, we have

$$\mathrm{Var}\big(Z\big) \le \sum_x \mathrm{Var}(\mathbb{1}_{N_x=0}\big) = \sum_x \mathbb{E}[\mathbb{1}_{N_x=0}(1 - \mathbb{1}_{N_x=0})] \le \sum_x \mathbb{E}\left[\mathbb{1}_{N_x>0}\right] \le n.$$

Analogously, $\mathrm{Var}(Z') \le n + m$ and hence

$$\mathrm{Var}(U^{\mathrm{h}} - U) = \mathrm{Var}(U^{\mathrm{h}} - Z + Z') \le 3\mathrm{Var}(U^{\mathrm{h}}) + 3\mathrm{Var}(Z') + 3\mathrm{Var}(Z) \le 3\mathrm{Var}(U^{\mathrm{h}}) + 6n + 3m.$$

Thus it remains to show

$$\mathrm{Var}(U^{\mathrm{h}}) \le 4n \sup_i h_i^2. \tag{16}$$

By induction on n, we show that for any $n \in \mathbb{N}$, any set of nonnegative integers $\{r_x\}$ and any function $(x, k) \mapsto f(x, k)$ with $k \in \mathbb{Z}_+$ satisfying $f(x, 0) = 0$,

$$\mathrm{Var}\left(\sum_x f(x, N_x)\right) \le 4n \|f\|_\infty^2, \tag{17}$$

where $\{N_x\} \sim \mathrm{Hyp}(\{r_x\}, n)$ and $\|f\|_\infty = \sup_{x,k} |f(x, k)|$. Then the desired Equation [16] follows from [17] with $f(x, k) = h_k$.

We first prove [17] for $n = 1$, in which case exactly one of N_x's is one and the rest are zero. Hence, $|\sum_x f(x, N_x)| \le \|f\|_\infty$ and $\mathrm{Var}(\sum_x f(x, N_x)) \le \|f\|_\infty^2$.

Next assume the induction hypothesis holds for $n-1$. Let X_1 denote the first sample and let $\tilde{N}_x$ denote the number of occurrences of symbol x in samples $X_2, \ldots, X_n$. Then $N_x = \tilde{N}_x + \mathbb{1}_{X_1=x}$.

Furthermore, conditioned on $X_1 = y$, $\{\tilde{N}_x\} \sim \mathrm{Hyp}(\{\tilde{r}_x\}, n-1)$, where $\tilde{r}_x = r_x - \mathbb{1}_{x=y}$. By the law of total variance, we have

$$\mathrm{Var}\left(\sum_x f(x, N_x)\right) = \mathbb{E}\left[V(X_1)\right] + \mathrm{Var}\left(g(X_1)\right). \tag{18}$$

where

$$V(y) \triangleq \mathrm{Var}\left(\sum_x f(x, N_x)\middle| X_1 = y\right), \quad g(y) \triangleq \mathbb{E}\left[\sum_x f(x, N_x)\middle| X_1 = y\right]$$

For the first term in [18], note that

$$V(y) = \mathrm{Var}\left(\sum_x f(x, \tilde{N}_x + \mathbb{1}_{x=y})\middle| X_1 = y\right) = \mathrm{Var}\left(\sum_x f_y(x, \tilde{N}_x)\middle| X_1 = y\right).$$

where we defined $f_y(x,k) \triangleq f(x, k + \mathbb{1}_{x=y})$. Hence, by the induction hypothesis, $V(y) \le 4(n-1)\|f_y\|_\infty^2 \le 4(n-1)\|f\|_\infty^2$ and $\mathbb{E}\left[V(X_1)\right] \le 4(n-1)||f||_\infty^2$.

For the second term in [18], observe that for any $y \neq z$

$$g(y) = \mathbb{E}[f(y, \tilde{N}_y + 1)|X_1 = y] + \mathbb{E}[f(z, \tilde{N}_z)|X_1 = y] + \mathbb{E}\left[\sum_{x \neq y,z} f(x, \tilde{N}_x)\middle| X_1 = y\right],$$

and

$$g(z) = \mathbb{E}[f(z, \tilde{N}_z + 1)|X_1 = z] + \mathbb{E}[f(y, \tilde{N}_y)|X_1 = z] + \mathbb{E}\left[\sum_{x \neq y,z} f(x, \tilde{N}_x)\middle| X_1 = z\right],$$

Observe that $\{N_x\}_{x \neq y,z}$ have the same joint distribution conditioned on either $X_1 = y$ or $X_1 = z$ and hence $\mathbb{E}[\sum_{x \neq y,z} f(x, \tilde{N}_x)|X_1 = y] = \mathbb{E}[\sum_{x \neq y,z} f(x, \tilde{N}_x)|X_1 = z]$. Therefore $|g(y) - g(z)| \le 4\|f\|_\infty$ for any $y \neq z$. This implies that the function g takes values in an interval of length at most $4\|f\|_\infty$. Therefore $\mathrm{Var}(g(X_1)) \le \frac{1}{4}(4\|f\|_\infty)^2 = 4\|f\|_\infty^2$. This completes the proof of [17] and hence the lemma. □

Let

$$B(h, r_x) \triangleq \sum_{i=1}^{r_x} \binom{r_x}{i} \left(\frac{n}{R}\right)^i \left(1 - \frac{n}{R}\right)^{r_x - i} h_i - \left(1 - \frac{n}{R}\right)^{r_x} \left(1 - \left(1 - \frac{m}{R-n}\right)^{r_x}\right).$$

To bound the bias, we first prove an auxiliary result.

Lemma 12. *For any linear estimator U^{h},*

$$\left|\mathbb{E}[U^{\mathrm{h}} - U] - \sum_x B(h, r_x)\right| \le 4 \max\left(\sup_i |h_i|, 1\right) + \frac{2R}{R-n}.$$

Proof. Recall that $N_x \sim \mathrm{Hyp}(R, r_x, n)$. Let $\tilde{N}_x$ be a random variable distributed as $\mathrm{Bin}(r_x, n/R)$. Since $\mathrm{Hyp}(R, r_x, n)$ coincides with $\mathrm{Hyp}(R, n, r_x)$, we have

$$d_{\mathrm{TV}}(\mathrm{Bin}(r_x, n/R), \mathrm{Hyp}(R, r_x, n)) = d_{\mathrm{TV}}(\mathrm{Bin}(r_x, n/R), \mathrm{Hyp}(R, n, r_x)) \le \frac{2r_x}{R},$$

where the last inequality follows from (7, Theorem 4). Since $d_{\mathrm{TV}}(\mu,\nu) = \frac{1}{2}\sup_{\|f\|_\infty\le 1}\int f d\mu - \int f d\nu = \sup_E \mu(E) - \nu(E)$, we have

$$\left|\mathbb{E}[f(N_x)] - \mathbb{E}[f(\tilde{N}_x)]\right| \le \frac{4r_x}{R}\sup_i |f(i)|, \tag{19}$$

and

$$\left|\frac{\binom{R-n-r_x}{m}}{\binom{R-n}{m}} - \left(1-\frac{m}{R-n}\right)^{r_x}\right| \le d_{\mathrm{TV}}(\mathrm{Bin}(r_x, m/(R-n)), \mathrm{Hyp}(R-n,m,r_x)) \le \frac{2r_x}{R-n}. \tag{20}$$

Define $f_x(i) = h_i - \mathbb{1}_{i=0}\left(1-\left(1-\frac{m}{R-n}\right)^{r_x}\right)$. In view of [15] and the fact that $\sum r_x = R$, we have

$$\left|\mathbb{E}[U^{\mathrm{h}} - U] - \sum_x \mathbb{E}[f_x(N_x)]\right| \le \frac{2R}{R-n}.$$

Applying [19] yields

$$\sum_x \left|\mathbb{E}[f_x(\tilde{N}_x)] - \mathbb{E}\left[f_x(N_x)\right]\right| \le 4\sup_i |f_x(i)| \le 4\max\left(\sup_i |h_i|, 1\right).$$

The above equation together with [20] results in the lemma since $B(h, r_x) = \mathbb{E}[f_x(\tilde{N}_x)]$. □

Note that to upper bound the bias, we need to bound $\sum_x B(h, r_x)$. It is easy to verify for the GT coefficients $h_i^{\mathrm{GT}} = -(-t)^i$ with $t = m/n$, $B(h^{\mathrm{GT}}, r_x) = 0$. Therefore, if we choose $h = h^{\mathrm{L}}$ based on the tail of random variable L with $h_i^{\mathrm{L}} = h_i^{\mathrm{GT}}\mathbb{P}(L \ge i)$ as defined in [6], we have

$$\begin{aligned} B(h^{\mathrm{L}}, r_x) &= \sum_{i=1}^{r_x}\binom{r_x}{i}\left(\frac{n}{R}\right)^i\left(1-\frac{n}{R}\right)^{r_x-i}(-t)^i\mathbb{P}(L<i) \\ &= \left(1-\frac{n}{R}\right)^{r_x}\sum_{i=1}^{r_x}\binom{r_x}{i}\left(-\frac{m}{R-n}\right)^i\mathbb{P}(L<i). \end{aligned} \tag{21}$$

Similar to Lemma 6, our strategy is to find an integral presentation of the bias. This is done in the following lemma.

Lemma 13. *For any $y \ge 0$ and any $k \in \mathbb{N}$,*

$$\sum_{i=1}^{k}\binom{k}{i}(-y)^i\mathbb{P}(L<i) = -k(1-y)^k\int_0^y \mathbb{E}\left[\binom{k-1}{L}(-s)^L\right](1-s)^{-k-1}ds. \tag{22}$$

Remark 1. *For the special case of $y = 1$, [22] is understood in the limiting sense: Letting $\delta = 1-y$ and $\beta = \frac{1-s}{\delta}$, we can rewrite the right-hand side as*

$$-k\int_1^{1/\delta}\mathbb{E}\left[\binom{k-1}{L}(\beta\delta-1)^L\right]k\beta^{-k-1}d\beta.$$

For all $|\delta| \le 1$ and hence $0 \le 1 - \beta\delta \le 2$, we have

$$\left|\mathbb{E}\left[\binom{k-1}{L}(\beta\delta-1)^L\right]\right| = \left|\mathbb{E}\left[\binom{k-1}{L}(\beta\delta-1)^L \mathbb{1}_{L<k}\right]\right| \le 4^k.$$

By dominated convergence theorem, as $\delta \to 0$, the right-hand side converges to $-\mathbb{E}\left[\binom{k-1}{L}(-1)^L\right]$ and coincides with the left-hand side, which can be easily obtained by applying $\binom{k}{i} = \binom{k-1}{i} + \binom{k-1}{i-1}$.

Proof. Denote the left-hand side of [22] by $F(y)$. Using $i\binom{k}{i} = k\binom{k-1}{i-1}$, we have

$$\begin{aligned} F'(y) &= \sum_{i=1}^{k}\binom{k}{i}(-i)(-y)^{i-1}\mathbb{P}(L<i) = -k\sum_{i=1}^{k}\binom{k-1}{i-1}(-y)^{i-1}\mathbb{P}(L<i) \\ &= -k\sum_{i=1}^{k}\binom{k-1}{i-1}(-y)^{i-1}\mathbb{P}(L<i-1) - k\sum_{i=1}^{k}\binom{k-1}{i-1}(-y)^{i-1}\mathbb{P}(L=i-1). \end{aligned} \tag{23}$$

The second term is simply $-k\mathbb{E}\left[\binom{k-1}{L}(-y)^L\right] \triangleq G(y)$. For the first term, since $L \ge 0$ almost surely and $\binom{k}{i} = \binom{k-1}{i} + \binom{k-1}{i-1}$, we have

$$\begin{aligned} k\sum_{i=1}^{k}\binom{k-1}{i-1}(-y)^{i-1}\mathbb{P}(L<i-1) &= k\sum_{i=1}^{k}\binom{k-1}{i}(-y)^{i}\mathbb{P}(L<i) \\ &= k\sum_{i=1}^{k}\binom{k}{i}(-y)^{i}\mathbb{P}(L<i) - k\sum_{i=1}^{k}\binom{k-1}{i-1}(-y)^{i}\mathbb{P}(L<i) \\ &= kF(y) - yF'(y). \end{aligned} \tag{24}$$

Combining [23] and [24] yields the following ordinary differential equation:

$$F'(y)(1-y) + kF(y) = G(y), \quad F(0) = 0,$$

whose solution is readily obtained as $F(y) = (1-y)^k \int_0^y (1-s)^{-k-1}G(s)ds$, *i.e.,* the desired Equation [22]. □

Combining Lemma 12–13 yields the following bias bound:

Lemma 14. *For any random variable L over $\mathbb{Z}_+$ and $t = m/n \ge 1$,*

$$|\mathbb{E}[U^{\mathrm{L}} - U]| \le nt \cdot \max_{0\le s\le 1}\left|\mathbb{E}\left[\binom{r_x-1}{L}(-s)^L\right]\right| + 4\mathbb{E}[t^L] + \frac{2R}{R-n}.$$

Proof. Recall the coefficient bound [7] that $\sup_i |h_i| \le \mathbb{E}[t^L]$. By Lemma 12 and the assumption that $t \ge 1$,

$$\left|\mathbb{E}[U^{\mathrm{h}} - U] - \sum_x B(h^{\mathrm{L}}, r_x)\right| \le 4\mathbb{E}[t^L] + \frac{2R}{R-n}.$$

Thus it suffices to bound $\sum_x B(h^{\mathrm{L}}, r_x)$. For every x, using [21] and applying Lemma 13 with $y = \frac{m}{R-n}$ and $k = r_x$, we obtain

$$B(h^{\mathrm{L}}, r_x) = -\left(1 - \frac{n+m}{R}\right)^{r_x} \int_0^{\frac{m}{R-n}} \mathbb{E}\left[\binom{r_x - 1}{L}(-s)^L\right] r_x (1-s)^{-r_x-1} ds.$$

Since $0 \le \frac{m}{R-n} \le 1$, letting $K = \max_{0\le s\le 1} \left|\mathbb{E}\left[\binom{r_x-1}{L}(-s)^L\right]\right|$, we have

$$\begin{aligned}|B(h^{\mathrm{L}}, r_x)| &\le \left(1 - \frac{n+m}{R}\right)^{r_x} K \int_0^{\frac{m}{R-n}} r_x (1-s)^{-r_x-1} ds. \\ &= K\left(\left(1 - \frac{n}{R}\right)^{r_x} - \left(1 - \frac{n+m}{R}\right)^{r_x}\right) \le K\left(1 - \frac{n}{R}\right)^{r_x - 1} \frac{m r_x}{R},\end{aligned}$$

where the last inequality follows from the convexity of $x \mapsto (1-x)^{r_x}$. Summing over all symbols x results in the lemma. □

Combining Lemma 14 and Lemma 11 gives the following NMSE bound:

Theorem 6. *Under the assumption of Lemma 14,*

$$\mathbb{E}[(U^{\mathrm{L}} - U)^2] \le 12(n+1)\mathbb{E}^2[t^L] + 6n + 3m + \frac{12R^2}{(R-n)^2} + 3m^2 \max_{1\ge\alpha>0} \left|\mathbb{E}\left[\binom{r_x - 1}{L}(-\alpha)^L\right]\right|^2.$$

As before, we can choose various smoothing distribution and obtain upper bounds on the mean squared error.

Corollary 4. *If $L \sim \mathrm{poi}(r)$ and $R - n \ge m \ge n$, then*

$$\mathbb{E}[(U^{\mathrm{L}} - U)^2] \le 12(n+1)e^{2r(t-1)} + 3m^2 e^{-r} + 9m + 48.$$

Furthermore, if $r = \frac{1}{2t-1} \cdot \log(nt^2)$,

$$\mathcal{E}_{n,t}(U^{\mathrm{L}}) \le \frac{27}{(nt^2)^{\frac{1}{2t-1}}} + \frac{9nt + 48}{(nt)^2}.$$

Proof. For $L \sim \mathrm{poi}(r)$, $\mathbb{E}[t^L] = e^{r(t-1)}$ and

$$\max_{0\le\alpha\le 1} \left|\mathbb{E}\left[\binom{r_x - 1}{L}(-\alpha)^L\right]\right| = e^{-r} \max_{0\le\alpha\le 1} |L_{r_x-1}(\alpha r)| \le e^{-r/2},$$

where L_{r_x-1} is the Laguerre polynomial of degree $r_x - 1$ defined in [11] and the last equality follows the bound [12]. Furthermore, $R/(R-n) = 1 + n/(R-n) \le 1 + n/m \le 2$ and $n \le m$, and hence the first part of the lemma. The second part follows by substituting the value of r. □

4 Lower bounds

Under the multinomial model (i.i.d. sampling), we lower bound the risk $\mathcal{E}_{n,t}(U^{\mathrm{E}})$ for any estimator U^{E} using the support size estimation lower bound in (15). Since the lower bound in (15) also holds for the Poisson model, so does our lower bound.

Recall that $S(p) = \sum_x \mathbb{1}_{p_x>0}$ denotes the support size of a distribution p. It is shown that given n i.i.d. samples drawn from a distribution p whose minimum non-zero mass $p^+_{\min}$ is at least $1/k$, the minimax mean-square error for estimating $S(p)$ satisfies

$$\min_{\hat{S}} \max_{p:p^+_{\min}\geq 1/k} \mathbb{E}[(\hat{S} - S(p))^2] \geq c'k^2 \cdot \exp\left(-c\max\left(\sqrt{\frac{n\log k}{k}}, \frac{n}{k}\right)\right). \tag{25}$$

where c, c' are universal positive constants with $c > 1$. We prove Theorem 2 under the multinomial model with c being the universal constant from [25].

Suppose there is an estimator $\hat{U}$ for U that can accurately predict the number of new symbols arising in the next m samples, we can then produce a support size estimator by adding the number of symbols observed, Φ_+, in the current n samples, namely,

$$\hat{S} = \hat{U} + \Phi_+. \tag{26}$$

Note that $U = \sum_x \mathbb{1}_{N_x=0}\mathbb{1}_{N'_x>0}$. For $m = \infty$, U is the total number of unseen symbols and we have $S(p) = U + \Phi_+$. Consequently, if $\hat{U}$ can foresee far into the future (*i.e.,* for too large an m), then [26] will constitute a support size estimator that is too good to be true.

Combining Theorem 2 with the positive result (Corollary 1 or 2) yields the following characterization of the minimax risk:

Corollary 5. *For all $t \geq c$,*

$$\inf_{U^{\mathrm{E}}} \mathcal{E}_{n,t}(U^{\mathrm{E}}) = \exp\left(-\Theta\left(\max\left\{\frac{\log n}{t}, 1\right\}\right)\right)$$

Consequently, as $n \to \infty$, the minimax risk $\inf_{U^{\mathrm{E}}} \mathcal{E}_{n,t}(U^{\mathrm{E}}) \to 0$ if and only if $t = o(\log n)$.

Proof of Theorem 2. Recall that $m = nt$. Let $\hat{U}$ be an arbitrary estimator for U. The support size estimator $\hat{S} = \hat{U} + \Phi_+$ defined in [26] must obey the lower bound [25]. Hence for some p satisfying $p^+_{\min} \geq 1/k$,

$$\mathbb{E}[(S(p) - \hat{S})^2] \geq c'k^2 \cdot \exp\left(-c\max\left(\sqrt{\frac{n\log k}{k}}, \frac{n}{k}\right)\right). \tag{27}$$

Let $S = S(p)$ denote the support size, which is at most k. Let $\tilde{U} \triangleq \mathbb{E}_{X_{n+1}^{n+m}}[U]$ be the expectation of U over the unseen samples X_{n+1}^{n+m} conditioned on the available samples X_1^n. Then $\tilde{U} = \sum_x \mathbb{1}_{N_x=0}\left(1 - (1-p_x)^{nt}\right)$. Since $\hat{U}$ is independent of X_{n+1}^{n+m}, by convexity,

$$\mathbb{E}_{X_1^{n+m}}[(U - \hat{U})^2] \geq \mathbb{E}_{X_1^n}[(\mathbb{E}_{X_{n+1}^{n+m}}[U - \hat{U}])^2] = \mathbb{E}[(\tilde{U} - \hat{U})^2]. \tag{28}$$

Notice that with probability one,

$$|S - \tilde{U} - \Phi_+| \leq Se^{-nt/k} \leq ke^{-nt/k}, \tag{29}$$

which follows from

$$\tilde{U} + \Phi_+ = \sum_{x:p_x>0} \mathbb{1}_{N_x=0} \left(1 - (1-p_x)^{nt}\right) + \mathbb{1}_{N_x>0} \le S,$$

and, on the other hand,

$$\begin{aligned}\tilde{U} + \Phi_+ &= \sum_{x:p_x\ge 1/k} \mathbb{1}_{N_x=0} \left(1 - (1-p_x)^{nt}\right) + \mathbb{1}_{N_x>0} \\ &\ge \sum_{x} \mathbb{1}_{N_x=0} \left(1 - (1-1/k)^{nt}\right) + \mathbb{1}_{N_x>0} \ge S(1-(1-1/k)^{nt}) \ge S(1-e^{-nt/k}).\end{aligned}$$

Expanding the left hand side of [27],

$$\begin{aligned}\mathbb{E}[(S-\hat{S})^2] = \mathbb{E}\left[\left(S - \tilde{U} - \Phi_+ + \tilde{U} - \hat{U}\right)^2\right] &\le 2\mathbb{E}[(S-\tilde{U}-\Phi_+)^2] + 2\mathbb{E}[(\tilde{U}-\hat{U}))^2] \\ &\overset{[29]}{\le} 2k^2 e^{-2nt/k} + 2\mathbb{E}[(\tilde{U}-\hat{U}))^2] \overset{[28]}{\le} 2k^2 e^{-2nt/k} + 2\mathbb{E}[(U-\hat{U}))^2]\end{aligned}$$

Let

$$k = \min\left\{\frac{nt^2}{c^2 \log \frac{nt^2}{c^2}}, \frac{nt}{\log \frac{4}{c'}}\right\},$$

which ensures that

$$c' k^2 \cdot \exp\left(-c \max\left\{\sqrt{\frac{n\log k}{k}}, \frac{n}{k}\right\}\right) \ge 4k^2 e^{-2nt/k}. \qquad [30]$$

Then

$$\mathbb{E}[(U-\hat{U})^2] \ge k^2 e^{-2nt/k},$$

establishes the following lower bound with $\alpha \triangleq \frac{c'^2}{4\log^2(4/c')}$ and $\beta \triangleq c^2$:

$$\min_E \mathcal{E}_{n,t}(U^{\mathrm{E}}) \ge \min\left\{\alpha, \frac{4t^2}{\beta^2 \log^2 \frac{nt^2}{\beta}} \left(\frac{\beta}{nt^2}\right)^{2\beta/t}\right\}.$$

To verify [30], since $t \ge c$ by assumption, we have $\exp(\frac{2tn}{k} - \frac{cn}{k}) \ge \exp(\frac{nt}{k}) \ge \frac{4}{c'}$. Similarly, since $k\log k \le \frac{nt^2}{c^2}$ by definition, we have $\frac{2nt}{k} \ge 2c'\sqrt{\frac{n\log k}{k}}$ and hence $\exp\left(\frac{2tn}{k} - c\sqrt{\frac{n\log k}{k}}\right) \ge \exp(\frac{nt}{k}) \ge \frac{4}{c'}$, completing the proof of [30].

Thus we have shown that there exist universal positive constants α, β such that

$$\min_E \mathcal{E}_{n,t}(U^{\mathrm{E}}) \ge \min\left\{\alpha, \frac{4t^2}{\beta^2 \log^2 \frac{nt^2}{\beta}} \left(\frac{\beta}{nt^2}\right)^{2\beta/t}\right\}.$$

Let $y = \left(\frac{nt^2}{\beta}\right)^{2\beta/t}$, then

$$\min_E \mathcal{E}_{n,t}(U^{\mathrm{E}}) \ge \min\left\{\alpha, 16\frac{1}{y\log^2 y}\right\}.$$

Since $y > 1$, $y^3 \ge y\log^2 y$ and hence for some constants $c_1, c_2 > 0$,

$$\min_E \mathcal{E}_{n,t}(U^{\mathrm{E}}) \ge \min\left\{\alpha, 16\frac{1}{y^3}\right\} \ge \min\left\{\alpha, \left(\frac{\beta}{nt^2}\right)^{6\beta/t}\right\} \ge c_1 \min\left\{1, \left(\frac{1}{n}\right)^{c_2/t}\right\} \ge \frac{c_1}{n^{c_2/t}}. \qquad \square$$

5 Connections to support size estimation

Approximation-theoretic techniques for estimating norms and other properties such as support size and entropy have been successfully used in statistics. For example, estimating the L_p norms in Gaussian models (10, 5) and estimating entropy (14, 9) and support size (15) of discrete distributions. Among the aforementioned problems, support size estimation is closest to ours. Hence, we now discuss the difference between the approximation technique we use and the those used for support size estimation.

The support size of a discrete distribution p is

$$S(p) = \sum_x \mathbb{1}_{p_x>0}. \qquad [31]$$

At first glance, estimating $S(p)$ may appear similar to species estimation as one can convert a support size estimator $\hat{S}$ to $\hat{U}$ by

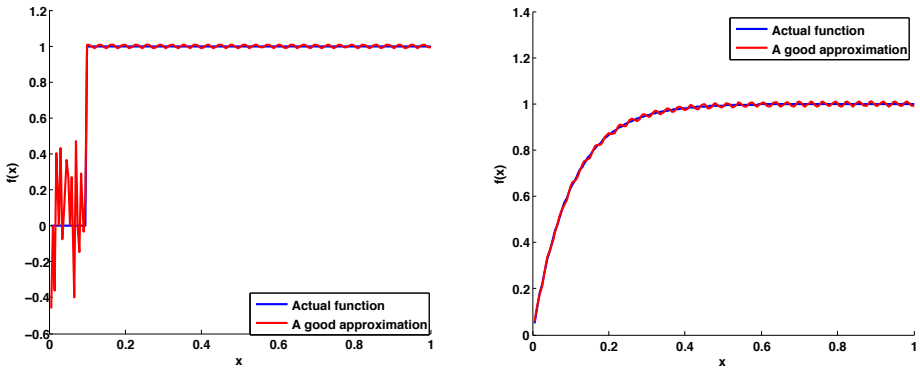
$$\hat{U} = \hat{S} - \sum_{i=1}^{\infty} \Phi_i.$$

However, without any assumption on the distribution it is impossible to estimate the support size. For example, regardless of the number of samples collected, there could be infinitely many symbols with arbitrarily small combined probabilities that have not been observed. A possible assumption is therefore that the lowest non-zero probability of the underlying distribution p, denoted by $p^+_{\min}$, is at least $1/k$, for some known k. Under this assumption (12) applied a linear programming estimator similar to the one in (8), to estimate the support size within an additive error of $k\epsilon$ with constant probability using $\Omega(\frac{k}{\log k}\frac{1}{\epsilon^2})$ samples. Based on best polynomial approximations, recently (15) showed that the minimax risk of support size estimation satisfies

$$\min_{\hat{S}} \max_{p:p^+_{\min}\geq 1/k} \mathbb{E}_p[(\hat{S} - S(p))^2] = k^2 \exp\left(-\Theta\left(\max\left\{\sqrt{\frac{k\log k}{n}}, \frac{k}{n}, 1\right\}\right)\right),$$

and therefore the optimal sample complexity of for estimating $S(p)$ within an additive error of $k\epsilon$ with constant probability is $\Theta(\frac{k}{\log k}\log^2\frac{1}{\epsilon})$. Note that the assumption $p^+_{\min} \geq 1/k$ is crucial for this result to hold as otherwise estimation is impossible. By contrast, we show later that for species estimation no such assumptions are necessary. The intuition is that if there exist a large number of very improbable symbols, most likely they will not appear in the new samples either.

To estimate the support size, in view of [31] and the assumption $p^+_{\min} \geq 1/k$, the technique of (15) is to approximate the indicator function $y \mapsto \mathbb{1}_{y\geq 1/k}$ in the range $\{0\} \cup [1/k, \log k/n]$ using Chebyshev polynomials. Since by assumption no p_x lies in $(0, \frac{1}{k})$, the approximation error in this interval is irrelevant. For example, in Figure 6(a), the red curve is a useful approximation for $S(p)$, even though it behaves badly over $(0, 1/k)$. To estimate the average number of unseen symbols U, in view of Lemma 1, we need to approximate $y \mapsto 1 - e^{-yt}$ over the entire $[0, \infty)$ as in, *e.g.*, Figure 6(b). Concurrent to this work, (13) proposed a linear programming algorithm to estimate U. However, their NMSE is $O(\frac{t}{\log n})$ compared to the optimal result $O(n^{-1/t})$ in Theorem 1, thus exponentially weaker for $t = o(\log n)$. Furthermore, the computational cost far exceeds those of our linear estimators.



(a) (b)

Figure 6: (a) a good approximation for support size; (b) a good approximation for species estimation.

6 Monotone-concave modification

As argued in (3), it is often useful for species estimators to be nonnegative as well as monotone and concave in the range of extrapolation m;[3] however, linear estimators including Good-Toulmin and SGT need not satisfy these properties. To address this issue, we can apply to any estimator sequence $\{U_m^{\mathrm{E}}\}_{m\geq 0}$ the following transform: for $m = 0$, let $\tilde{U}_0^{\mathrm{E}} = 0$, for $m = 1$

$$\tilde{U}_m^{\mathrm{E}} = \max\left(\tilde{U}_0^{\mathrm{E}}, U_1^{\mathrm{E}})\right),$$

and for every $m > 0$,

$$\tilde{U}_m^{\mathrm{E}} = \min(\max(\tilde{U}_{m-1}^{\mathrm{E}}, U_m^{\mathrm{E}}), 2\tilde{U}_{m-1}^{\mathrm{E}} - \tilde{U}_{m-2}^{\mathrm{E}}),$$

resulting in $\tilde{U}^{\mathrm{E}}$ that is always non-negative. Furthermore it is both monotone and concave in m.

Lemma 15. *$\tilde{U}_m^{\mathrm{E}}$ is always non-negative and it is both monotone and concave in m.*

Proof. For concavity, observe that

$$\begin{aligned}\tilde{U}_m^{\mathrm{E}} - \tilde{U}_{m-1}^{\mathrm{E}} &= \min(\max(\tilde{U}_{m-1}^{\mathrm{E}}, U_m^{\mathrm{E}}), 2\tilde{U}_{m-1}^{\mathrm{E}} - \tilde{U}_{m-2}^{\mathrm{E}}) - \tilde{U}_{m-1}^{\mathrm{E}} \\ &\leq 2\tilde{U}_{m-1}^{\mathrm{E}} - \tilde{U}_{m-2}^{\mathrm{E}} - \tilde{U}_{m-1}^{\mathrm{E}} \\ &= \tilde{U}_{m-1}^{\mathrm{E}} - \tilde{U}_{m-2}^{\mathrm{E}},\end{aligned}$$

since the difference between consecutive terms is decreasing the sequence is concave in m.

The proof of monotonicity is by induction. Observe that $\tilde{U}_1^{\mathrm{E}} \geq \tilde{U}_0^{\mathrm{E}}$ by construction. For any $m \geq 2$, we prove by induction that $\tilde{U}_m^{\mathrm{E}} \geq \tilde{U}_{m-1}^{\mathrm{E}}$. Suppose $\tilde{U}_{m-1}^{\mathrm{E}} \geq \tilde{U}_{m-2}^{\mathrm{E}})$, then

$$\tilde{U}_m^{\mathrm{E}} = \min(\max(\tilde{U}_{m-1}^{\mathrm{E}}, U_m^{\mathrm{E}}), 2\tilde{U}_{m-1}^{\mathrm{E}} - \tilde{U}_{m-2}^{\mathrm{E}}) \geq \min(\max(\tilde{U}_{m-1}^{\mathrm{E}}, U_m^{\mathrm{E}}), \tilde{U}_{m-1}^{\mathrm{E}}) \geq \tilde{U}_{m-1}^{\mathrm{E}},$$

where the first inequality follows by the inductive hypothesis.

Finally, since the sequence is monotone and $\tilde{U}_0^{\mathrm{E}} = 0$, it is always nonnegative. □

[3] A sequence a_m is said to be concave if the successive difference is non-increasing, i.e., $a_{m+1} - a_m \leq a_m - a_{m-1}$

In Figure 7, we compare the performance of the SGT estimator (with Binomial smoothing of parameter $q = 2/(2+t)$) and its monotone-concave version on the premise as Fig. 3 in the main paper. As before, the true value is shown in black, and the estimators are colored, with the solid line representing their means and the shaded band corresponding to one standard deviation. For computational purposes we apply a variation of this transform, relating estimates that are $0.2n$ apart instead of 1 apart. Note that the performance of the original and modified versions are similar.

References

[1] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Wiley-Interscience, New York, NY, 1964.

[2] A. D. Barbour and P. Hall. On the rate of poisson convergence. *Mathematical Proceedings of the Cambridge Philosophical Society*, 95:473–480, 5 1984. ISSN 1469-8064.

[3] S. Boneh, A. Boneh, and R. J. Caron. Estimating the prediction function and the number of unseen species in sampling with replacement. *Journal of the American Statistical Association*, 93(441):372–379, 1998.

[4] J. Bunge and M. Fitzpatrick. Estimating the number of species: a review. *Journal of the American Statistical Association*, 88(421):364–373, 1993.

[5] T. Cai and M. G. Low. Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2):1012–1041, 2011.

[6] R. K. Colwell, A. Chao, N. J. Gotelli, S.-Y. Lin, C. X. Mao, R. L. Chazdon, and J. T. Longino. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Jour. of Plant Ecology*, 5(1):3–21, 2012.

[7] P. Diaconis and D. Freedman. Finite exchangeable sequences. *Ann. Probab.*, 8(4):745–764, 08 1980. doi: 10.1214/aop/1176994663.

[8] B. Efron and R. Thisted. Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63(3):435–447, 1976.

[9] J. Jiao, K. Venkat, Y. Han, and T. Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2015.

[10] O. Lepski, A. Nemirovski, and V. Spokoiny. On estimation of the L_r norm of a regression function. *Probability Theory and Related Fields*, 113(2):221–253, 1999.

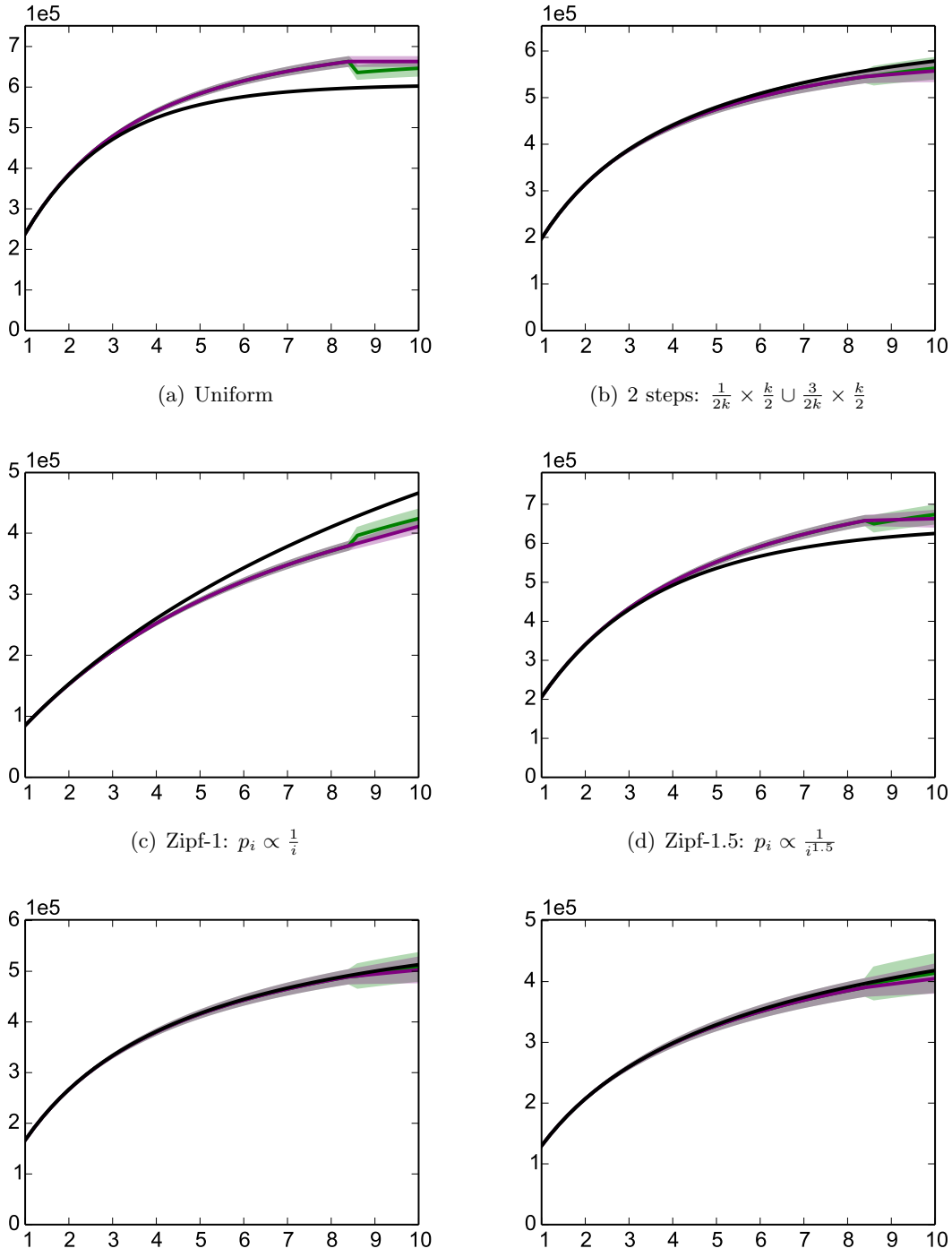
[11] J. M. Steele. An Efron-Stein inequality for nonsymmetric statistics. *Ann. Statist.*, 14(2): 753–758, 06 1986. doi: 10.1214/aos/1176349952.

[12] G. Valiant and P. Valiant. Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the 43rd annual ACM symposium on Theory of computing*, pages 685–694, 2011.

[13] G. Valiant and P. Valiant. Instance optimal learning. *arXiv preprint arXiv:1504.05321*, 2015.

[14] Y. Wu and P. Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *to appear in IEEE Transactions on Information Theory, arxiv:1407.0381*, Jul 2015.

[15] Y. Wu and P. Yang. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *preprint arxiv:1504.01227*, Apr. 2015.



(a) Uniform

(b) 2 steps: $\frac{1}{2k} \times \frac{k}{2} \cup \frac{3}{2k} \times \frac{k}{2}$

(c) Zipf-1: $p_i \propto \frac{1}{i}$

(d) Zipf-1.5: $p_i \propto \frac{1}{i^{1.5}}$

(e) Dirichlet-1 prior

(f) Dirichlet-1/2 prior

True value	Proposed
–	Binomial smoothing $q = \frac{2}{t+2}$ —
	Monotone-concave Binomial smoothing $q = \frac{2}{t+2}$ —

Figure 7: Comparisons of the estimated number of unseen species as a function of t. All experiments have distribution support size 10^6, $n = 5 \cdot 10^5$, and are averaged over 100 iterations.