The Bethe and Sinkhorn Approximations of the Pattern Maximum Likelihood Estimate and their Connections to the Valiant–Valiant Estimate

Pascal O. Vontobel

***Abstract*—For estimating a source's distribution histogram, Orlitsky and co-workers have proposed the pattern maximum likelihood (PML) estimate, which says that one should choose the distribution histogram that has the largest likelihood of producing the pattern of the observed symbol sequence. It can be shown that finding the PML estimate is equivalent to finding the distribution histogram that maximizes the permanent of a certain non-negative matrix.**

However, in general this optimization problem appears to be intractable and so one has to compute suitable approximations of the PML estimate. In this paper, we discuss various efficient PML estimate approximation algorithms, along with their connections to the Valiant–Valiant estimate of the distribution histogram. These connections are established by associating an approximately doubly stochastic matrix with the Valiant–Valiant estimate and comparing this approximately doubly stochastic matrix with the doubly stochastic matrices that appear in the free energy descriptions of the PML estimate and its approximations.

I. Introduction

Consider a memoryless source with alphabet $\mathcal{X}$ of finite size $|\mathcal{X}|$ and with probability mass function $\boldsymbol{\pi} = (\pi_x)_{x \in \mathcal{X}}$, *i.e.*, the probability that the source outputs a sequence $\boldsymbol{x} \triangleq x_1, x_2, \ldots, x_n$ of length n equals

$$P(\boldsymbol{x} \,|\, \boldsymbol{\pi}) = \pi_{x_1} \cdot \pi_{x_2} \cdots \pi_{x_n}.$$

Because this probability depends only on how often every symbol appears in $\boldsymbol{x}$, we can write

$$P(\boldsymbol{x} \,|\, \boldsymbol{\pi}) = \prod_{x \in \mathcal{X}} \pi_x^{|\{\ell \,|\, x_\ell = x\}|}.$$

Assume that $\boldsymbol{\pi}$ is unknown. If n is relatively large compared to $|\mathcal{X}|$, then estimating $\boldsymbol{\pi}$ via the maximum likelihood (ML) estimate

$$\hat{\pi}_x \triangleq \frac{\big|\{\ell \,|\, x_\ell = x\}\big|}{n}, \quad x \in \mathcal{X},$$

works very well. In particular, this estimate is asymptotically consistent in the limit $n \to \infty$.

However, there are quite a few applications where $|\mathcal{X}|$ is fairly large and where n is on the order of $|\mathcal{X}|$, and one wonders if there are better estimates than the above estimate. Such scenarios frequently occur when $\mathcal{X}$ is a set of words, a set of sequences, *etc.*, and one is not able, or, due to cost constraints, not willing to collect many samples. Note that when n is on the order of $|\mathcal{X}|$ then the variance of $\hat{\pi}_x$ is large and it can easily happen that $\hat{\pi}_x$ significantly over- or under-estimates π_x, and, in some cases even estimates $\hat{\pi}_x = 0$ because the symbol x has never been observed. The latter fact is particularly troublesome when one wants, for example, to find how many $x \in \mathcal{X}$ have $\pi_x > 0$ and how many $x \in \mathcal{X}$ have $\pi_x = 0$.

Solutions to this problem have been proposed at least since the 1940s, in particular in the work by Fisher, Corbet, and Williams [1] and by Good and Turing [2]. For more details, the interested reader is referred to the historical remarks in the papers [3], [4] and the bibliography available at [5]. Solutions to this problem have recently also gained attention because of their relevance for the analysis of big data (see, *e.g.*, [6]), where one tries to obtain a good understanding of the statistical properties of data despite the fact that, because of the sheer volume of the data, one can only sample a relatively small fraction of it.

A. The Distribution Histogram

One is often not interested in the distribution $\boldsymbol{\pi}$ itself, but only in the distribution histogram (also known as the distribution's probability multiset), *i.e.*, the function

$$\begin{aligned} v : [0,1] &\to \mathbb{Z}_{\geqslant 0} \\ q &\mapsto \big|\{x \in \mathcal{X} \mid \pi_x = q\}\big|. \end{aligned}$$

The reason for this is that for many quantities of interest the distribution histogram is a sufficient statistic. For example, the support size and entropy of $\boldsymbol{\pi}$ are given by, respectively,

$$|\operatorname{supp}(\boldsymbol{\pi})| = \sum_{q \in (0,1]} v(q), \tag{1}$$

$$H(\boldsymbol{\pi}) = -\sum_{q \in (0,1]} v(q) \cdot q \log(q). \tag{2}$$

Example 1 *Let* $\mathcal{X} \triangleq \{\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e}, \mathrm{f}\}$ *and let* $\pi_\mathrm{a} = 0.20$, $\pi_\mathrm{b} = 0.35$, $\pi_\mathrm{c} = 0.20$, $\pi_\mathrm{d} = 0$, $\pi_\mathrm{e} = 0.10$, *and* $\pi_\mathrm{f} = 0.15$. *The distribution histogram of* $\boldsymbol{\pi}$ *is* $v(0) = 1$, $v(0.10) = 1$, $v(0.15) = 1$, $v(0.20) = 2$, $v(0.35) = 1$, *and* $v(q) = 0$ *for all other* $q \in [0,1]$. *The distribution* $\boldsymbol{\pi}$ *has support size* $|\operatorname{supp}(\boldsymbol{\pi})| = 5$ *and entropy* $H(\boldsymbol{\pi}) \approx 1.526$ nats. □

In the rest of the paper, we will be interested in estimating the source's distribution histogram based on an observed

P. O. Vontobel is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA, and the Department of Information Technology and Electrical Engineering, ETH Zurich, 8092 Zurich, Switzerland (e-mail: pascal.vontobel@ieee.org).

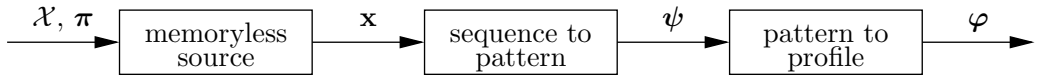

Fig. 1. The main setup of this paper: a memoryless source with alphabet $\mathcal{X}$ of size $|\mathcal{X}| = k$ and distribution $\boldsymbol{\pi}$ produces a symbol sequence $\boldsymbol{x}$ of length n. The task is to estimate the source's distribution histogram $\boldsymbol{p}$ based on the sequence $\boldsymbol{x}$. Towards this end, the PML estimate makes use of the pattern $\boldsymbol{\psi}$ (see Definition 7) and the profile $\boldsymbol{\varphi}$ (see Definition 17) of the sequence $\boldsymbol{x}$. (See the main text for further details.)

sequence $\boldsymbol{x} \triangleq x_1, x_2, \ldots, x_n$, where $x_\ell \in \mathcal{X}$. The distribution histogram estimate $\hat{v}$, besides being of interest by itself, can be used towards estimating the support size or the entropy of $\boldsymbol{\pi}$ by plugging it into (1) and (2), respectively, *i.e.*,

$$|\widehat{\operatorname{supp}(\boldsymbol{\pi})}| = \sum_{q \in (0,1]} \hat{v}(q), \tag{3}$$

$$\widehat{H(\boldsymbol{\pi})} = -\sum_{q \in (0,1]} \hat{v}(q) \cdot q \log(q). \tag{4}$$

Interestingly, many properties of the distribution $\boldsymbol{\pi}$ can be expressed by *linear* functionals of the distribution histogram v (like the support size and the entropy in (1) and (2), respectively); therefore these properties may be estimated by *linear* estimators based on $\hat{v}$ (like the support size and the entropy estimates in (3) and (4), respectively). Of course, other estimators of these quantities are also possible. (For a more detailed discussion on the topic of linear estimators based on $\hat{v}$ *vs.* other possible types of estimators, the interested reader is referred to, *e.g.*, [7].)

Before continuing, we need some notation. For any positive integer L we define $[L]$ to be the set $[L] \triangleq \{1, \ldots, L\}$ and we let $\mathcal{V}_L$ be the set of distribution histograms of size L, *i.e.*,

$$\mathcal{V}_L \triangleq \left\{ \text{distribution histogram } v \;\middle|\; \sum_{q \in [0,1]} v_q = L \right\}.$$

For a finite set $\mathcal{S}$, we define $\Pi_{\mathcal{S}}$ to be the set of probability mass functions over $\mathcal{S}$, *i.e.*,

$$\Pi_{\mathcal{S}} \triangleq \left\{ \boldsymbol{\pi} = (\pi_s)_{s \in \mathcal{S}} \;\middle|\; \pi_s \geqslant 0 \text{ for all } s \in \mathcal{S},\ \sum_{s \in \mathcal{S}} \pi_s = 1 \right\}.$$

If there is a total order on the set $\mathcal{S}$, then we define $\Pi_{\mathcal{S}}^{\downarrow}$ to be the set of monotone probability mass functions over $\mathcal{S}$, *i.e.*,

$$\Pi_{\mathcal{S}}^{\downarrow} \triangleq \left\{ \boldsymbol{p} \in \Pi_{\mathcal{S}} \;\middle|\; \begin{array}{c} \text{for all } s, s' \in \mathcal{S}, \\ p_s \geqslant p_{s'} \text{ if and only if } s \leqslant s' \end{array} \right\}.$$

Clearly, there is a bijection between the set $\mathcal{V}_L$ and the set $\Pi_{[L]}^{\downarrow}$, and so elements of $\Pi_{[L]}^{\downarrow}$ will also be called distribution histograms. Moreover, any element of $\boldsymbol{\pi}$ of $\Pi_{\mathcal{S}}$ can be mapped to an element $\Pi_{\mathcal{S}}^{\downarrow}$ via the mapping $\boldsymbol{\pi} \mapsto \boldsymbol{\pi}^{\downarrow}$, where $\boldsymbol{\pi}^{\downarrow}$ contains the same components as $\boldsymbol{\pi}$, but sorted in non-increasing order.

Example 2 *We continue Example 2. Note that the histogram function v is in $\mathcal{V}_6$ and that it is associated with $\boldsymbol{p} \in \Pi_6^{\downarrow}$ with entries $p_1 = 0.35$, $p_2 = 0.20$, $p_3 = 0.20$, $p_4 = 0.15$, $p_5 = 0.10$, $p_6 = 0$.* □

B. The Main Setup

We are now ready to describe the setup under consideration in this paper (see also Fig. 1):

- Fix some positive integer k.
- Let $\mathcal{X}$ be a set of cardinality k. When not mentioned otherwise, we choose $\mathcal{X} \triangleq [k]$.
- Consider a memoryless source with alphabet $\mathcal{X}$ and distribution $\boldsymbol{\pi}$. The source's distribution histogram is called $\boldsymbol{p}$, where $\boldsymbol{p} \triangleq \boldsymbol{\pi}^{\downarrow}$.
- The source produces a string $\boldsymbol{x} \triangleq x_1, x_2, \ldots, x_n$, where $x_\ell \in \mathcal{X}$.
- The main task is to estimate $\boldsymbol{p}$ based on the knowledge of k and the observation of $\boldsymbol{x}$. (Of course, $\boldsymbol{\pi}$ is not known.)

Note that in contrast to most papers on the PML estimate, here we assume that k is known. We leave it as an exercise for the reader to extend the results of this paper to the case where k is unknown a priori.

C. The Sequence Maximum Likelihood Estimate

A simple approach to estimate $\boldsymbol{p}$ based on $\boldsymbol{x}$ is specified in the next definition: in a first step it computes the ML estimate of $\boldsymbol{\pi}$ based on $\boldsymbol{x}$ and in a second step it sorts the resulting vector in non-increasing order towards obtaining a vector in $\Pi_{[k]}^{\downarrow}$.

Definition 3 *The vector $\hat{\boldsymbol{p}}^{\mathrm{SML}} \in \Pi_{[k]}^{\downarrow}$, called the sequence maximum likelihood (SML) estimate of $\boldsymbol{p}$, is defined to be the vector*

$$\hat{\boldsymbol{p}}^{\mathrm{SML}} \triangleq \left(\arg\max_{\boldsymbol{\pi}} \; P(\boldsymbol{x} \,|\, \boldsymbol{\pi}) \right)^{\downarrow},$$

where $\boldsymbol{\pi}$ varies over $\Pi_{[k]}$. □

Lemma 4 *It holds that*

$$\hat{\boldsymbol{p}}^{\mathrm{SML}} = \hat{\boldsymbol{\pi}}^{\downarrow}, \quad \textit{where} \quad \hat{\pi}_i = \frac{\big|\{\ell \,|\, x_\ell = i\}\big|}{n},\ i \in [k].$$

Proof: Straightforward. ■

Although the estimate $\hat{\boldsymbol{p}}^{\mathrm{SML}}$ is efficiently computable from $\boldsymbol{x}$, and although it is asymptotically consistent (in the limit $n \to \infty$ for fixed k), it has several drawbacks when n is on the order k, as mentioned earlier in this introduction.

D. The Pattern Maximum Likelihood Estimate

Orlitsky *et al.* [8], [9] have proposed a different approach for obtaining an estimate of $\boldsymbol{p}$. First, they noted that because the vector $\boldsymbol{p} = \boldsymbol{\pi}^{\downarrow}$ associated with the distribution $\boldsymbol{\pi}$ does not care about the labels in $\mathcal{X}$, we can base, w.l.o.g., the estimation of $\boldsymbol{p}$ not on the string $\boldsymbol{x}$, but on the string of integers that is

obtained by replacing every symbol in the sequence $\boldsymbol{x}$ by its order or appearance; this new string is called the pattern $\boldsymbol{\psi}$ of $\boldsymbol{x}$. Second, they proposed that the estimate of $\boldsymbol{p}$ should be given by the $\boldsymbol{p}$ that maximizes the probability that the pattern $\boldsymbol{\psi}$ is observed; the resulting estimate $\hat{\boldsymbol{p}}^{\mathrm{PML}}$ of $\boldsymbol{p}$ is called the pattern maximum likelihood (PML) estimate. The following definitions formalize this approach.

Definition 5 *The* pattern $\boldsymbol{\psi} \triangleq \boldsymbol{\psi}(\boldsymbol{x}) \triangleq \psi_1, \psi_2, \ldots, \psi_n$ *of a sequence* $\boldsymbol{x}$ *is defined to be the integer sequence derived by replacing each symbol in* $\boldsymbol{x}$ *by its order of appearance.* □

Example 6 *We continue Examples 1 and 2. Assume that the memoryless source produces the sequence* $\boldsymbol{x} = \mathrm{c,c,a,b,c,b,b,e,a}$ *of length* $n = 9$*. (Note that "*f*" does not appear in* $\boldsymbol{x}$ *despite the fact that* $\pi_{\mathrm{f}} > 0$*.) This sequence has* $m = 4$ *distinct symbols, and has the pattern* $\boldsymbol{\psi} = 1,1,2,3,1,3,3,4,2$. □

Let the number of distinct symbols in $\boldsymbol{x}$ by m. Clearly, the length of $\boldsymbol{\psi}$ is n, *i.e.*, the same as the length of $\boldsymbol{x}$, and the number of distinct integers in $\boldsymbol{\psi}$ equals m, *i.e.*, the same as the number of distinct symbols in $\boldsymbol{x}$.

Definition 7 *Let the* multiplicity $\mu_\psi(\boldsymbol{\psi})$ *of the integer* ψ *in the pattern* $\boldsymbol{\psi}$ *be the number of times that* ψ *appears in* $\boldsymbol{\psi}$*,* i.e.,

$$\mu_\psi(\boldsymbol{\psi}) \triangleq \{\ell \mid \psi_\ell = \psi\}, \quad \psi \in [k].$$

(Note that $\mu_\psi(\boldsymbol{\psi}) = 0$*,* $m < \psi \leqslant k$*, where* m *is the number of distinct symbols in* $\boldsymbol{x}$*, and therefore also the number of distinct symbols in* $\boldsymbol{\psi}$*.) All multiplicities are collected in the multiplicity vector* $\boldsymbol{\mu}(\boldsymbol{\psi})$*. Obviously,* $\sum_{\psi \in [k]} \mu_\psi(\boldsymbol{\psi}) = n$*.* [1] □

Example 8 *We continue Examples 1, 2, and 6. The multiplicities of the integers in the pattern* $\boldsymbol{\psi} = 1,1,2,3,1,3,3,4,2$ *are* $\mu_1 = 3$*,* $\mu_2 = 2$*,* $\mu_3 = 3$*,* $\mu_4 = 1$*,* $\mu_5 = 0$*, and* $\mu_6 = 0$*.* □

We are now ready for the main definition of this paper.

Definition 9 (PML estimate) *For a given positive integer* k *and a pattern* $\boldsymbol{\psi}$*, the pattern maximum likelihood (PML) estimate is defined to be*

$$\hat{\boldsymbol{p}}^{\mathrm{PML}} \triangleq \hat{\boldsymbol{p}}^{\mathrm{PML}}(\boldsymbol{\psi}) \triangleq \arg\max_{\boldsymbol{p}} P(\boldsymbol{\psi} \,|\, \boldsymbol{p}),$$

where $\boldsymbol{p}$ *varies over* $\Pi^{\downarrow}_{[k]}$*.* □

The PML estimate has many interesting properties and uses; for an in-depth discussion we refer the interested reader to the relevant literature (in particular [8], but also [3], [9]–[17]).

[1] In this definition we follow [10] that defines the multiplicity vector based on the pattern $\boldsymbol{\psi}$. This is in contrast to other papers by Orlitsky *et al.* that define the multiplicity vector based on $\boldsymbol{x}$. Clearly, both definitions yield the same vector up to reordering.

The pattern probability $P(\boldsymbol{\psi} \,|\, \boldsymbol{p})$ of a pattern $\boldsymbol{\psi}$ with m distinct integers and multiplicity vector $\boldsymbol{\mu} \triangleq \boldsymbol{\mu}(\boldsymbol{\psi})$ can be written as follows

$$\begin{aligned} P(\boldsymbol{\psi} \,|\, \boldsymbol{p}) &= \sum_{y_1 \in [k]} \sum_{\substack{y_2 \in [k], \\ y_2 \neq y_1}} \cdots \sum_{\substack{y_m \in [k] \\ y_m \neq y_1, y_2, \ldots}} p_{y_1}^{\mu_1} p_{y_2}^{\mu_2} \cdots p_{y_m}^{\mu_m} \\ &= \frac{1}{(k-m)!} \sum_{y_1 \in [k]} \sum_{\substack{y_2 \in [k], \\ y_2 \neq y_1}} \cdots \sum_{\substack{y_k \in [k] \\ y_k \neq y_1, y_2, \ldots}} p_{y_1}^{\mu_1} p_{y_2}^{\mu_2} \cdots p_{y_k}^{\mu_k}. \end{aligned}$$

As was shown in [18], and discussed next, one can rewrite the right-hand side of this expression in terms of the permanent of a suitably defined square matrix. In order to do that, we recall the definition of the permanent of a square matrix.

Definition 10 (see, *e.g.*, [19]) *Let* $\boldsymbol{\theta} = (\theta_{i,j})_{i,j}$ *be a real matrix of size* $L \times L$*. The permanent of* $\boldsymbol{\theta}$ *is defined to be the scalar*

$$\mathrm{perm}(\boldsymbol{\theta}) = \sum_{\sigma} \prod_{i \in [L]} \theta_{i,\sigma(i)}, \tag{5}$$

where the summation is over all $L!$ *permutations of the set* $[L]$*.* [2] □

Lemma 11 *Let* k *be a positive integer, let* $\boldsymbol{p} \in \Pi^{\downarrow}_{[k]}$*, let* $\boldsymbol{\psi}$ *be a pattern of length* n *with* m *distinct integers. Then*

$$P(\boldsymbol{\psi} \,|\, \boldsymbol{p}) = \frac{1}{(k-m)!} \cdot \mathrm{perm}\Big(\boldsymbol{\theta}\big(\boldsymbol{p}, \boldsymbol{\mu}(\boldsymbol{\psi})\big)\Big),$$

where $\boldsymbol{\theta}(\boldsymbol{p}, \boldsymbol{\mu})$ *is the* $k \times k$ *non-negative matrix*

$$\boldsymbol{\theta}(\boldsymbol{p}, \boldsymbol{\mu}) \triangleq \begin{pmatrix} p_1^{\mu_1} & p_1^{\mu_2} & \cdots & p_1^{\mu_k} \\ p_2^{\mu_1} & p_2^{\mu_2} & \cdots & p_2^{\mu_k} \\ \vdots & \vdots & & \vdots \\ p_k^{\mu_1} & p_k^{\mu_2} & \cdots & p_k^{\mu_k} \end{pmatrix}, \tag{6}$$

i.e., *the components of* $\boldsymbol{\theta} \triangleq \boldsymbol{\theta}(\boldsymbol{p}, \boldsymbol{\mu})$ *are*

$$\theta_{i,j} = p_i^{\mu_j}, \quad (i,j) \in [k]^2.$$

Note that for $m < j \leqslant k$ *we have* $\mu_j = 0$*, and so* $\theta_{i,j} = 1$*.*

Proof: Follows from the above expression for $P(\boldsymbol{\psi} \,|\, \boldsymbol{p})$ and Definition 10. ■

Definition 12 (Restatement of the PML estimate) *For a given positive integer* k *and a pattern* $\boldsymbol{\psi}$*, the PML estimate is defined to be*

$$\hat{\boldsymbol{p}}^{\mathrm{PML}} \triangleq \hat{\boldsymbol{p}}^{\mathrm{PML}}(\boldsymbol{\psi}) \triangleq \arg\max_{\boldsymbol{p}} \mathrm{perm}\Big(\boldsymbol{\theta}\big(\boldsymbol{p}, \boldsymbol{\mu}(\boldsymbol{\psi})\big)\Big),$$

where $\boldsymbol{p}$ *varies over* $\Pi^{\downarrow}_{[k]}$ *and where* $\boldsymbol{\theta}\big(\boldsymbol{p}, \boldsymbol{\mu}(\boldsymbol{\psi})\big)$ *is defined in* (6)*.* □

[2] Contrast this definition with the definition of the *determinant* of $\boldsymbol{\theta}$, *i.e.*, $\det(\boldsymbol{\theta}) = \sum_\sigma \mathrm{sgn}(\sigma) \prod_{i \in [L]} \theta_{i,\sigma(i)}$, where $\mathrm{sgn}(\sigma)$ equals $+1$ if σ is an even permutation and equals -1 if σ is an odd permutation.

E. Approximations of the PML Estimate

Because finding the $\hat{\boldsymbol{p}}^{\mathrm{PML}}$ appears to be intractable for practically relevant problem sizes, it is desirable to devise efficient algorithms that find approximations to $\hat{\boldsymbol{p}}^{\mathrm{PML}}$. In the following, we list a variety of approaches.

- **(PML estimate via approximately solving a fixed-point equation)** As will be discussed in the extended version of this paper [20], one can formulate an algorithm for which $\hat{\boldsymbol{p}}^{\mathrm{PML}}$ is a fixed point. Because the algorithm requires some quantities that appear to be intractable, one can use Markov chain Monte Carlo (MCMC) sampling to approximate them. One can show that this approach is equivalent to an algorithm that was proposed by Orlitsky *et al*. (an early version of that algorithm was mentioned in [21]; for a detailed discussion see [9]). Although this fixed-point based algorithm is relatively simple to derive, it is not immediately obvious how to prove its convergence. This problem is remedied by the algorithm described in the next item.
- **(PML estimate via approximately minimizing the Gibbs free energy function)** As was discussed in Section II of [22], one can express $\hat{\boldsymbol{p}}^{\mathrm{PML}}$ as the solution of a certain Gibbs free energy function minimization problem. (We refer to [22] for details.) One can then approximately find a minimum (or at least a stationary point) of this Gibbs free energy function by applying an alternating minimization algorithm where the steps are based on quantities that are obtained by Markov chain Monte Carlo (MCMC) sampling. One can show that this algorithm is equivalent to the algorithm discussed in the previous item. The advantage of the Gibbs free energy function framework is that it allows a principled approach to prove convergence of the algorithm.
- **(Bethe PML estimate)** In this approach, which was discussed in Section III of [22], we approximated the above Gibbs free energy function by the more tractable Bethe free energy function and applied an alternating minimization algorithm to this function; the resulting estimate was called the Bethe PML (BPML) estimate.
  The Bethe free energy function approach leveraged results from [23]–[25] on how to approximate the permanent of a non-negative matrix with the help of the sum-product algorithm (SPA) (see also [26]–[29]). Note that there are some high-level similarities between the problem setup in [23], [24] and the problem setup here: namely, in both setups one tries to learn some model parameters (flow parameters in [23], [24] *vs.* distribution histogram here) that maximize a model fitness function that is expressed in terms of some permanent. The conclusions in both setups were very similar: for the same level of accuracy, the Bethe free energy based approach is much faster than the Gibbs free energy / MCMC based approach. There are two notable differences between the setups though. First, our model has typically many more parameters than the setup in [23], [24]. Second, the structure of the matrix $\boldsymbol{\theta}(\boldsymbol{p}, \boldsymbol{\mu})$ can be used towards significantly speeding up the SPA computations.
- **(Sinkhorn PML estimate)** In this approach, discussed in Section II of the present paper, we replace the above Gibbs free energy function by a tractable approximation that we call the Sinkhorn free energy function and apply an alternating minimization algorithm to this function; the resulting estimate is called the Sinkhorn PML (SPML) estimate. The corresponding algorithm is slightly simpler than for the Bethe PML estimate, but the Sinkhorn PML estimate is usually slightly further away from the PML estimate than the Bethe PML estimate from the PML estimate.
- **(Dictionary Bethe and Sinkhorn PML estimates)** As will be discussed in the extended version of this paper [20], one can use the special structure of the Bethe and the Sinkhorn free energies to formulate linear programs that can be solved very efficiently and whose solutions give good approximations of the BPML and SPML estimate, respectively. In other words, one obtains approximations of the approximations of the PML estimate; the resulting estimates are called Dictionary Bethe PML (DBPML) estimate and Dictionary Sinkhorn PML (DSPML) estimate, respectively.

The above free energy framework also highlights the central role played by a certain doubly stochastic matrix (called $\boldsymbol{\gamma}$) for the PML estimate and its approximations. Leveraging results for doubly stochastic matrices allows us then to simplify some earlier proofs of properties of the PML estimate and to derive some new properties of the PML estimate and its approximations. In particular, as will be discussed in [20], the Bethe and the Sinkhorn PML estimates behave in many ways similar to the PML estimate.

Recently, an alternative way of estimating the distribution histogram has been presented by Valiant and Valiant [4]. In Section III of the present paper, with the help of the above-mentioned doubly stochastic matrix $\boldsymbol{\gamma}$, we discuss connections between their estimate and the Sinkhorn PML estimate. In fact, these connections inspired the definition of the Dictionary Bethe PML estimate and the Dictionary Sinkhorn PML estimate.

Experimental results for some of the above-mentioned estimates were shown in [22]. More experimental results will be presented in [20].

F. Basic Notations and Definitions

This subsection discusses the most important notations that will be used in this paper. More notational definitions will be given in later sections.

We let $\mathbb{R}$ be the field of real numbers, $\mathbb{R}_{\geqslant 0}$ be the set of non-negative real numbers, $\mathbb{R}_{>0}$ be the set of positive real numbers, $\mathbb{Z}$ be the ring of integers, $\mathbb{Z}_{\geqslant 0}$ be the set of non-negative integers, and $\mathbb{Z}_{>0}$ be the set of positive integers. Scalars are denoted by non-boldface characters, whereas vectors and matrices by boldface characters. The size of a set $\mathcal{S}$ is denoted by $|\mathcal{S}|$.

As usual, the entropy of a probability mass function $\boldsymbol{\pi}$ is defined to be $H(\boldsymbol{\pi}) \triangleq -\sum_{\ell} \pi_{\ell} \log(\pi_{\ell})$, whereas the relative entropy between two probability mass functions $\boldsymbol{\pi}$ and $\boldsymbol{\pi}'$ is

defined to be $D(\boldsymbol{\pi} \,\|\, \boldsymbol{\pi}') \triangleq \sum_\ell \pi_\ell \log(\pi_\ell/\pi'_\ell)$. Unless stated otherwise, all logarithms are natural logarithms. Moreover, we let $0 \cdot \log(0) \triangleq 0$, and so $0^0 = 1$.

The inner product of two vectors $\boldsymbol{a} = (a_i)_i$ and $\boldsymbol{b} = (b_i)_i$ of the same length is defined to be $\langle \boldsymbol{a}, \boldsymbol{b} \rangle \triangleq \sum_i a_i b_i$. Similarly, the inner product of two matrices $\boldsymbol{A} = (A_{i,j})_{i,j}$ and $\boldsymbol{B} = (B_{i,j})_{i,j}$ of the same size is defined to be $\langle \boldsymbol{A}, \boldsymbol{B} \rangle \triangleq \sum_{i,j} A_{i,j} B_{i,j}$.

A matrix is called non-negative (positive) if all entries are non-negative (positive). For a non-negative matrix $\boldsymbol{\theta} = (\theta_{i,j})_{i,j}$ we define $\log^\circ(\boldsymbol{\theta})$ to be the matrix where the logarithm function is applied component-wise to $\boldsymbol{\theta}$, *i.e.*, $\log^\circ(\boldsymbol{\theta}) \triangleq \big(\log(\theta_{i,j})\big)_{i,j}$.

For any positive integer L, we let $\Gamma_{L\times L}$ be the set of doubly stochastic matrices of size $L \times L$, *i.e.*,

$$\Gamma_{L\times L} \triangleq \left\{ \boldsymbol{\gamma} = \big(\gamma_{i,j}\big) \,\middle|\, \begin{array}{l} \gamma_{i,j} \geqslant 0 \text{ for all } (i,j) \in [L] \times [L] \\ \sum_{j\in[L]} \gamma_{i,j} = 1 \text{ for all } i \in [L] \\ \sum_{i\in[L]} \gamma_{i,j} = 1 \text{ for all } j \in [L] \end{array} \right\}.$$

Finally, throughout the paper we assume that the source alphabet is $\mathcal{X} = \{1, 2, \ldots, k\}$, where k is some positive integer, and that the sequence $\boldsymbol{x}$ has length n and contains m distinct symbols. Similarly, the pattern $\boldsymbol{\psi} \triangleq \boldsymbol{\psi}(\boldsymbol{x})$ has length n and contains m distinct integers. Moreover, if not stated otherwise, $\sum_i$, $\sum_{i'}$, $\sum_j$, $\sum_{j'}$ will stand for, respectively, $\sum_{i\in[k]}$, $\sum_{i'\in[k]}$, $\sum_{j\in[k]}$, $\sum_{j'\in[k]}$.

II. The Sinkhorn Approximation to the Pattern Maximum Likelihood Estimate

In this section we discuss the Sinkhorn pattern maximum likelihood (SPML) estimate, which is based on a permanent approximation that we call the Sinkhorn permanent. The name for this approximation stems from the algorithm, namely the Sinkhorn balancing algorithm, that can be used to compute the Sinkhorn permanent (at least in the case of positive matrices).

Definition 13 *Fix some temperature $T \geqslant 0$ and let $\boldsymbol{\theta}$ be a non-negative matrix of size $k \times k$. Define the functions*

$$\begin{aligned} F_{\mathrm{S}}(\,\cdot\,;\boldsymbol{\theta}) &: \Gamma_{k\times k} \to \mathbb{R}, \quad \boldsymbol{\gamma} \mapsto U_{\mathrm{S}}(\boldsymbol{\gamma};\boldsymbol{\theta}) - T \cdot H_{\mathrm{S}}(\boldsymbol{\gamma}), \\ U_{\mathrm{S}}(\,\cdot\,;\boldsymbol{\theta}) &: \Gamma_{k\times k} \to \mathbb{R}, \quad \boldsymbol{\gamma} \mapsto -\big\langle \boldsymbol{\gamma}, \log^\circ(\boldsymbol{\theta}) \big\rangle \\ H_{\mathrm{S}}(\,\cdot\,) &: \Gamma_{k\times k} \to \mathbb{R}, \quad \boldsymbol{\gamma} \mapsto -\big\langle \boldsymbol{\gamma}, \log^\circ(\boldsymbol{\gamma}) \big\rangle. \end{aligned}$$

Here, F_{S}, U_{S}, and H_{S} are called the Sinkhorn free energy function, the Sinkhorn average energy function, and the Sinkhorn entropy function, respectively. The Sinkhorn permanent of a non-negative matrix $\boldsymbol{\theta}$ of size $k \times k$ is then defined to be

$$\mathrm{perm}_{\mathrm{S}}(\boldsymbol{\theta}) \triangleq \max_{\boldsymbol{\gamma}} \exp\big(-F_{\mathrm{S}}(\boldsymbol{\gamma};\boldsymbol{\theta})\big),$$

where the maximization is over all $\boldsymbol{\gamma} \in \Gamma_{k\times k}$. □

Definition 14 *For a given pattern $\boldsymbol{\psi}$, the Sinkhorn PML (SPML) estimate $\hat{\boldsymbol{p}}^{\mathrm{SPML}}$ is defined to be*

$$\begin{aligned} \hat{\boldsymbol{p}}^{\mathrm{SPML}} \triangleq \hat{\boldsymbol{p}}^{\mathrm{SPML}}(\boldsymbol{\psi}) &\triangleq \arg\max_{\boldsymbol{p}} \mathrm{perm}_{\mathrm{S}}\Big(\boldsymbol{\theta}\big(\boldsymbol{p}, \boldsymbol{\mu}(\boldsymbol{\psi})\big)\Big) \\ &= \arg\min_{\boldsymbol{p}} \min_{\boldsymbol{\gamma}} F_{\mathrm{S}}\Big(\boldsymbol{\gamma}; \boldsymbol{\theta}\big(\boldsymbol{p}, \boldsymbol{\mu}(\boldsymbol{\psi})\big)\Big), \end{aligned}$$

where $\boldsymbol{p}$ varies over $\Pi^{\downarrow}_{\mathcal{S}}$ and where $\boldsymbol{\gamma}$ varies over $\Gamma_{k\times k}$. □

Note that the function $(\boldsymbol{p}, \boldsymbol{\gamma}) \mapsto F_{\mathrm{S}}\big(\boldsymbol{\gamma}; \boldsymbol{\theta}\big(\boldsymbol{p}, \boldsymbol{\mu}(\boldsymbol{\psi})\big)\big)$ is convex in $\boldsymbol{p}$ for fixed $\boldsymbol{\gamma}$, and convex in $\boldsymbol{\gamma}$ for fixed $\boldsymbol{p}$, but in general it is not convex in $(\boldsymbol{p}, \boldsymbol{\gamma})$.

In the following, we will consider the problem of finding the SPML estimate for temperature $T > 0$. Actually, since one can show (see, e.g. [20]) that the problem of finding the SPML estimate for temperature $T > 0$ can be reduced to the problem of finding the SPML estimate for temperature $T = 1$, we will only consider the case $T = 1$. Finding the SPML estimate for temperature $T = 0$ will be discussed in [20].

A. The Case $T = 1$

In this subsection we assume that $T = 1$. Let us start by commenting on Definition 13.

- Clearly, U_{S} is linear in $\boldsymbol{\gamma}$, H_{S} is concave in $\boldsymbol{\gamma}$, and so F_{G} is convex is $\boldsymbol{\gamma}$. The last fact follows also from rewriting $F_{\mathrm{S}}(\boldsymbol{\gamma};\boldsymbol{\theta})$ in terms of a relative-entropy-like expression[3] as follows
$$F_{\mathrm{S}}(\boldsymbol{\gamma};\boldsymbol{\theta}) = \sum_{i,j} \gamma_{i,j} \cdot \log\left(\frac{\gamma_{i,j}}{\theta_{i,j}}\right).$$
- If $\boldsymbol{\theta}$ is a positive matrix of size $k \times k$, then Sinkhorn's theorem [30]) (see also [31] and [25, Section 3.3]) states that there is a unique doubly stochastic matrix $\boldsymbol{\gamma}_{\boldsymbol{\theta}}$ such that
$$\boldsymbol{\theta} = \boldsymbol{D} \cdot \boldsymbol{\gamma}_{\boldsymbol{\theta}} \cdot \boldsymbol{E}, \tag{7}$$
where $\boldsymbol{D}$ and $\boldsymbol{E}$ are diagonal matrices with positive diagonals. (The matrices $\boldsymbol{D}$ and $\boldsymbol{E}$ are themselves unique, up to a scalar factor.)
- The matrices $\boldsymbol{D}$, $\boldsymbol{\gamma}_{\boldsymbol{\theta}}$, and $\boldsymbol{E}$ in the previous item can be found by the following algorithm (called Sinkhorn's balancing algorithm):
  - Let $\boldsymbol{\gamma}^{(0)} \triangleq \boldsymbol{\theta}$, and let $\boldsymbol{D}^{(0)}$ and $\boldsymbol{E}^{(0)}$ be the identity matrices of size $k \times k$.
  - For $t = 1, 2, \ldots$ alternatingly apply the following two steps until a suitable convergence criterion is met.
    - Let $\boldsymbol{S}$ be the $k \times k$ diagonal matrix that has s_i, $i \in [k]$, in the diagonal, where $s_i \triangleq \sum_j \gamma^{(t-1)}_{i,j}$. Let
    $$\begin{aligned} \boldsymbol{\gamma}^{(t-0.5)} &\triangleq \boldsymbol{S}^{-1} \cdot \boldsymbol{\gamma}^{(t-1)}, \\ \boldsymbol{D}^{(t-0.5)} &\triangleq \boldsymbol{D}^{(t-1)} \cdot \boldsymbol{S}, \\ \boldsymbol{E}^{(t-0.5)} &\triangleq \boldsymbol{E}^{(t-1)}. \end{aligned}$$
    - Let $\boldsymbol{S}$ be the $k \times k$ diagonal matrix that has s_j, $j \in [k]$, in the diagonal, where $s_j \triangleq \sum_i \gamma^{(t-0.5)}_{i,j}$. Let
    $$\begin{aligned} \boldsymbol{\gamma}^{(t)} &\triangleq \boldsymbol{\gamma}^{(t-0.5)} \cdot \boldsymbol{S}^{-1}, \\ \boldsymbol{D}^{(t)} &\triangleq \boldsymbol{D}^{(t-0.5)}, \\ \boldsymbol{E}^{(t)} &\triangleq \boldsymbol{S} \cdot \boldsymbol{E}^{(t-0.5)}. \end{aligned}$$

[3] We write "relative-entropy-like expression" because $\boldsymbol{\gamma}$ is a doubly stochastic matrix and not a distribution. Moreover, $\boldsymbol{\theta}$ is neither a distribution nor (in general) a doubly stochastic matrix.

- Fix some positive matrix $\boldsymbol{\theta}$ of size $k \times k$ and decompose it as in (7), *i.e.*, $\boldsymbol{\theta} = \boldsymbol{D} \cdot \boldsymbol{\gamma}_{\boldsymbol{\theta}} \cdot \boldsymbol{E}$ with $\boldsymbol{D} = \operatorname{diag}\big((d_i)_{i\in[k]}\big)$ and $\boldsymbol{E} = \operatorname{diag}\big((e_j)_{j\in[k]}\big)$. Then one can show that $\boldsymbol{\gamma} = \boldsymbol{\gamma}_{\boldsymbol{\theta}}$ minimizes $F_{\mathrm{S}}(\boldsymbol{\gamma};\boldsymbol{\theta})$ and that

$$\operatorname{perm}_{\mathrm{S}}(\boldsymbol{\theta}) = \left(\prod_{i\in[k]} d_i\right) \cdot \left(\prod_{j\in[k]} e_j\right). \tag{8}$$

- Fix some matrix $\boldsymbol{\theta} \triangleq \boldsymbol{\theta}\big(\boldsymbol{p}, \boldsymbol{\mu}(\boldsymbol{\psi})\big)$. If all entries of $\boldsymbol{p}$ are positive, then the matrix $\boldsymbol{\theta}$ is a positive matrix. However, if there is an $(i,j) \in [k]^2$ with $p_i = 0$ and $\mu_j > 0$, then at least one entry of $\boldsymbol{\theta}$ equals zero. A priori, the Sinkhorn balancing algorithm cannot be used in this case because there might be no matrix decomposition as in (7). However, one can show, by symmetry considerations, that the $\boldsymbol{\gamma}^*$ minimizing $F_{\mathrm{S}}(\boldsymbol{\gamma};\boldsymbol{\theta})$ satisfies

$$\gamma^*_{i,j} = \begin{cases} 0 & ((i,j) \in [k]^2 \text{ such that } p_i = 0,\ \mu_j > 0) \\ \frac{1}{k-m} & ((i,j) \in [k]^2 \text{ such that } p_i = 0,\ \mu_j = 0) \end{cases},$$

  where m is the number of distinct symbols in the pattern, *i.e.*, $k - m$ equals the number of $j \in [k]$ with $\mu_j = 0$. This partial solution of $\boldsymbol{\gamma}^*$ can be used towards finding the remaining entries of $\boldsymbol{\gamma}^*$.

The upcoming Algorithm 16 is motivated by the following theorem.

Theorem 15 *Fix a pattern $\boldsymbol{\psi}$ and let $\boldsymbol{\mu} \triangleq \boldsymbol{\mu}(\boldsymbol{\psi})$ be its multiplicity vector.*

- *For a fixed $\boldsymbol{p} \in \Pi_{[k]}$ define*

$$\boldsymbol{\gamma}^* \triangleq \arg\min_{\boldsymbol{\gamma}}\ F_{\mathrm{S}}(\boldsymbol{\gamma};\boldsymbol{\theta}), \tag{9}$$

  where $\boldsymbol{\theta} \triangleq \boldsymbol{\theta}(\boldsymbol{p},\boldsymbol{\mu})$. This matrix satisfies

$$\gamma^*_{i,j} \triangleq \theta^*_{i,j} \cdot \exp(-\lambda_i - \lambda'_j), \quad (i,j) \in [k]^2, \tag{10}$$

  where $(\lambda_i)_{i\in[k]}, (\lambda'_j)_{j\in[k]} \in \mathbb{R}^k$ are suitable Lagrange multipliers. In the case where all entries of $\boldsymbol{p}$ are positive, the matrix $\boldsymbol{\gamma}^*$ can for example be found with the help of the Sinkhorn balancing algorithm applied to the matrix $\boldsymbol{\theta}$.
- *For a fixed $\boldsymbol{\gamma} \in \Gamma_{k\times k}$ define*

$$\boldsymbol{p}^* \triangleq \arg\min_{\boldsymbol{p}}\ F_{\mathrm{S}}\big(\boldsymbol{\gamma};\boldsymbol{\theta}(\boldsymbol{p},\boldsymbol{\mu})\big). \tag{11}$$

  Then

$$\boldsymbol{p}^* = \boldsymbol{\gamma} \cdot \frac{\boldsymbol{\mu}}{n}. \tag{12}$$

Proof: See Appendix A. ■

Algorithm 16 *Fix a pattern $\boldsymbol{\psi}$ and let $\boldsymbol{\mu} \triangleq \boldsymbol{\mu}(\boldsymbol{\psi})$ be its multiplicity vector.*

- *Arbitrarily initialize the vector $\boldsymbol{p}^{(0)} \in \Pi_{[k]}$ with positive components.*
- *For $t = 1, 2, \ldots$ alternatingly apply the following two steps until a suitable convergence criterion is met.*
  - *Compute $\boldsymbol{\gamma}^{(t)}$ according to*

$$\boldsymbol{\gamma}^{(t)} \triangleq \arg\min_{\boldsymbol{\gamma}}\ F_{\mathrm{S}}\big(\boldsymbol{\gamma};\boldsymbol{\theta}(\boldsymbol{p}^{(t-1)},\boldsymbol{\mu})\big). \tag{13}$$

    In the case where all entries of $\boldsymbol{p}$ are positive, the matrix $\boldsymbol{\gamma}^*$ can for example be found with the help of the Sinkhorn balancing algorithm applied to the matrix $\boldsymbol{\theta}$.
  - *Compute $\boldsymbol{p}^{(t)}$ according to*

$$\begin{aligned}\boldsymbol{p}^{(t)} &\triangleq \arg\min_{\boldsymbol{p}}\ F_{\mathrm{S}}\big(\boldsymbol{\gamma}^{(t)};\boldsymbol{\theta}(\boldsymbol{p},\boldsymbol{\mu})\big) \\ &= \boldsymbol{\gamma}^{(t)} \cdot \frac{\boldsymbol{\mu}}{n}.\end{aligned} \tag{14}$$

□

Let us comment on this algorithm.

- For showing convergence of $(\boldsymbol{p}^{(t)}, \boldsymbol{\gamma}^{(t)})$ in Algorithm 16 to a stationary point of of $(\boldsymbol{p},\boldsymbol{\gamma}) \mapsto F_{\mathrm{S}}\big(\boldsymbol{\gamma};\boldsymbol{\theta}(\boldsymbol{p},\boldsymbol{\mu})\big)$, one can for example use standard results for alternating minimization algorithms (see, *e.g.*, [32, Proposition 2.7.1]).
- A naive implementation of the Sinkhorn balancing algorithm to solve (13) in round t for a given vector $\boldsymbol{p}^{(t-1)}$ has a time and space complexity proportional to k^2 per iteration. However, as will be discussed in [20], the structure of the matrix $\boldsymbol{\theta}\big(\boldsymbol{p}^{(t-1)}, \boldsymbol{\mu}(\boldsymbol{\psi})\big)$ can be used towards significantly reducing the time and space complexity. Moreover, typically only few iterations are necessary because instead of initializing the Sinkhorn balancing algorithm with $\boldsymbol{\gamma}^{(0)} = \boldsymbol{\theta}\big(\boldsymbol{p}^{(t-1)}, \boldsymbol{\mu}\big)$, one can initialize it with $\boldsymbol{\gamma}^{(0)} = \operatorname{diag}\big((\exp(-\lambda_i)_{i\in[k]})\big) \cdot \boldsymbol{\theta}\big(\boldsymbol{p}^{(t-1)}, \boldsymbol{\mu}\big) \cdot \operatorname{diag}\big((\exp(-\lambda_j)_{j\in[k]})\big)$, where $(\lambda_i)_{i\in[k]}$ and $(\lambda'_j)_{j\in[k]}$ are the Lagrange multipliers (10) that were found in the previous round.

III. Connections to the Valiant–Valiant Estimate

Recently, a very interesting approach for estimating the distribution histogram of a memoryless source based on the observed pattern was presented by Valiant and Valiant [4]. In the following, we will call their estimate the Valiant–Valiant (VV) estimate. The aim of this section is to give a rough outline of the VV estimate, along with connecting it to the other estimates in this paper, in particular the Sinkhorn PML (SPML) estimate.[4] For the exact details of the VV estimate, we refer the interested reader to the paper [4], [33].

The rest of this section is structured as follows. In Section III-A we introduce a concept called prevalence [3], [8], which will turn out to be helpful for expressing the upcoming results. In Section III-B we present what we call a "simplified version" of the VV estimate. The simplification comes from the fact that we consider only multiplicity vectors $\boldsymbol{\mu}$ that satisfy a certain constraint. In Section III-C we discuss the (general) VV estimate.

[4] In [20] we will connect the VV estimate also to the Dictionary Sinkhorn PML (DSPML) estimate.

A. Prevalence

Definition 17 *Let $\boldsymbol{x}$ be a string with pattern $\boldsymbol{\psi}$ and multiplicity vector $\boldsymbol{\mu} \triangleq \boldsymbol{\mu}(\boldsymbol{\psi})$. The prevalence $\varphi_\mu \triangleq \varphi_\mu(\boldsymbol{\psi})$ of a multiplicity μ is defined to be the number of symbols appearing μ times in $\boldsymbol{\psi}$,* i.e.,

$$\varphi_\mu(\boldsymbol{\psi}) \triangleq \big\{\psi \in [k] \;\big|\; \mu_\psi = \mu\big\}.$$

Equivalently, φ_μ is also the number of symbols appearing μ times in $\boldsymbol{x}$. □

Based on this definition, Orlitsky *et al.* [3], [8] have also defined the profile $\boldsymbol{\varphi}$ of a pattern $\boldsymbol{\psi}$ to be the formal product $\prod_\mu \mu^{\varphi_\mu}$. One can show that this is a sufficient statistic to compute the PML estimate. Therefore, "PML estimate" sometimes also stands for "profile maximum likelihood estimate" instead of "pattern maximum likelihood estimate." (See also Figure 1.)

Example 18 *We continue Examples 1, 2, 6, and 8. Recall that the pattern is $\boldsymbol{\psi} = 1,1,2,3,1,3,3,4,2$ and that its multiplicities are $\mu_1 = 3$, $\mu_2 = 2$, $\mu_3 = 3$, $\mu_4 = 1$, $\mu_5 = 0$, and $\mu_6 = 0$. With this, the prevalences of $\boldsymbol{\mu}$ are $\varphi_3 = 2$, $\varphi_2 = 1$, $\varphi_1 = 1$, and $\varphi_0 = 2$. Therefore, the profile of $\boldsymbol{\mu}$ is $\boldsymbol{\varphi} = 3^2 2^1 1^1 0^2$.* □

B. Simplified Valiant–Valiant Estimate

Here we assume that the pattern has no "often appearing" symbols, *i.e.*, there are no ψ where μ_ψ is "large." (We refer to the paper [4] for details, in particular, what "large" exactly means.)

The main idea of the VV estimate is to look for a distribution histogram $v : [0,1] \to \mathbb{Z}_{\geqslant 0}$ that satisfies certain constraints, *cf.* Definition 19 for the simplified VV estimate and Definition 21 for the (general) VV estimate. In particular, the support of v support should be in $\mathcal{Q}$, where $\mathcal{Q}$ is a carefully selected finite subset of $[0,1]$, which implies that v can be represented by a vector $\boldsymbol{v} \triangleq (v(q))_{q\in\mathcal{Q}}$ with non-negative integral entries. Actually, one first looks for a vector $\boldsymbol{v}' \triangleq (v'(q))_{q\in\mathcal{Q}}$ with non-negative real entries, whereupon $\boldsymbol{v}$ is obtained form $\boldsymbol{v}'$ by the application of a relatively simple rounding and adjustment procedure that makes sure that $\boldsymbol{v}$ is "close" to $\boldsymbol{v}'$, yet is a proper distribution histogram.

Definition 19 (Simplified VV Estimate) *Fix a pattern $\boldsymbol{\psi}$, let $\boldsymbol{\mu} \triangleq \boldsymbol{\mu}(\boldsymbol{\psi})$ be its multiplicity vector, and let $\boldsymbol{\varphi} \triangleq \boldsymbol{\varphi}(\boldsymbol{\mu})$ be its profile. Moreover, for some positive integer L, let $\mathcal{Q} \triangleq \{q_1, \ldots, q_L\}$ be a certain subset of $[0,1]$ of size L. Let $\boldsymbol{v}' = (v'_1, \ldots, v'_L)$ be a vector that satisfies the following constraints*

$$v'_\ell \geqslant 0 \quad \ell \in [L], \tag{15}$$

$$\sum_\ell v'_\ell \cdot q_\ell = 1, \tag{16}$$

$$\sum_\ell v'_\ell \cdot \mathrm{e}^{-nq_\ell} \frac{(nq_\ell)^\mu}{\mu!} \in \big[\varphi_\mu - f(n), \varphi_\mu + f(n)\big], \;\; \mu \in \mathbb{Z}_{\geqslant 0}, \tag{17}$$

where $\sum_\ell$ is shorthand for $\sum_{\ell\in[L]}$ and where $f(n)$ is a certain function with sublinear growth. The (potentially improper) simplified VV distribution histogram estimate $v^{\mathrm{VV}'} : [0,1] \to \mathbb{R}_{\geqslant 0}$ is then given by

$$v^{\mathrm{VV}'}(q) \triangleq \begin{cases} v_\ell & \text{(if } q = q_\ell \text{ for some } \ell \in [L]\text{)} \\ 0 & \text{(otherwise)} \end{cases}. \tag{18}$$

A proper distribution histogram estimate $v^{\mathrm{VV}} : [0,1] \to \mathbb{Z}_{\geqslant 0}$ is obtained from $v^{\mathrm{VV}'}$ with the help of a suitable rounding and adjustment procedure. (We refer to [4] for details.) □

In the above definition, the constraints in (15) and (16) ensure that $v^{\mathrm{VV}'}$ is a (potentially improper) distribution histogram, whereas the constraints in (17) are based on computations what the statistics of the profile $\boldsymbol{\varphi} \triangleq \boldsymbol{\varphi}(\boldsymbol{\mu})$ of the multiplicity vector $\boldsymbol{\mu} \triangleq \boldsymbol{\mu}(\boldsymbol{\psi})$ of a pattern $\boldsymbol{\psi}$ generated by a source with distribution histogram v' looks like. The size and the elements of the set $\mathcal{Q}$ are chosen such that a feasible vector $\boldsymbol{v}'$ can be found efficiently, yet so that one can guarantee nice properties of the solution. As pointed out in the paper [4], the larger $\mathcal{Q}$ is, the easier it is to establish the proofs of the properties of $\boldsymbol{v}$ and $\boldsymbol{v}'$, but the more computationally demanding the search of a feasible vector $\boldsymbol{v}'$ will be. Note though that there is a non-zero probability that the constraints in Definition 19 are infeasible; however, this issue can be alleviated by replacing the problem of finding a vector that satisfies the constraints in Definition 19 by the problem of finding a vector that is the solution of a suitable optimization problem (see Chapter 8 of [33] for details).

In the following, we want to connect the simplified VV estimate to the Sinkhorn PML estimate.

Remark 20 *Consider a vector $\boldsymbol{v}'$ that satisfies the constraints in Definition 19. For reasons of simplicity, we assume that $\boldsymbol{v}'$ has integral entries.*

- *Define $k \triangleq \sum_\ell v'_\ell$,* i.e., *$k$ is the positive integer such that $v^{\mathrm{VV}'} \in \mathcal{V}_k$.*
- *If necessary, stretch the $\boldsymbol{\mu}$ vector by including zero components so that its length is k.*
- *Let the vector $\boldsymbol{p}^* \in \Pi^{\downarrow}_{[k]}$ be obtained from $\boldsymbol{v}' \in \mathcal{V}_k$ by using the bijection between the set $\Pi^{\downarrow}_{[k]}$ and $\mathcal{V}_k$ (see Section I-A).*
- *Let the matrix $\boldsymbol{\gamma}^*$ of size $k \times k$ have components*

$$\gamma^*_{i,j} \triangleq \mathrm{e}^{-np^*_i} \cdot \frac{(np^*_i)^{\mu_j}}{\mu_j! \cdot \varphi_{\mu_j}}, \quad (i,j) \in [k]^2.$$

One can make the following statements about $\boldsymbol{p}^*$ and $\boldsymbol{\gamma}^*$.

- *The matrix $\boldsymbol{\gamma}^*$ is approximately doubly stochastic. By this we mean that all entries are non-negative and that the row and column sums are approximately* 1.
- *The vector-matrix pair $(\boldsymbol{p}^*, \boldsymbol{\gamma}^*)$ is close to being a fixed point of Algorithm 16.* [5]

Proof: See Appendix B. ■

[5] Note that "close to being a fixed point of Algorithm 16" does not imply that $(\boldsymbol{p}^*, \boldsymbol{\gamma}^*)$ is close to an actual fixed point of Algorithm 16.

C. Valiant–Valiant Estimate

In contrast to the simplified VV estimate, here we allow patterns that have "often appearing" symbols, *i.e.*, we allow j where μ_j is "large." The approach behind the VV estimate is to consider large μ_j values to be reliable enough to conclude that the distribution histogram estimate v^{VV} should be such that $v^{\mathrm{VV}}(\mu_j/n) = \varphi_{\mu_j}$.

More specifically, the VV estimate defines the tentative support of v to be $\mathcal{Q} \triangleq \mathcal{Q}_{\mathrm{low}} \cup \mathcal{Q}_{\mathrm{high}}$, where $\mathcal{Q}_{\mathrm{low}}$ and $\mathcal{Q}_{\mathrm{high}}$ are two disjoint sets such that all elements of $\mathcal{Q}_{\mathrm{low}}$ are smaller than the elements of $\mathcal{Q}_{\mathrm{high}}$. According to the above comment concerning "large" μ_j values, $\mathcal{Q}_{\mathrm{high}}$ is chosen to be

$$\mathcal{Q}_{\mathrm{high}} \triangleq \left\{ \frac{\mu_j}{n} \;\middle|\; \mu_j \text{ is "large"} \right\}.$$

On the other hand, $\mathcal{Q}_{\mathrm{low}}$ is a carefully selected finite subset of $[0,1]$. Moreover, let $\mathcal{Q}'_{\mathrm{low}}$ be a certain subset of the largest values in $\mathcal{Q}_{\mathrm{low}}$. (We refer to the paper [4] for details.)

Definition 21 (VV Estimate) *Fix a pattern $\boldsymbol{\psi}$, let $\boldsymbol{\mu} \triangleq \boldsymbol{\mu}(\boldsymbol{\psi})$ be its multiplicity, and let $\boldsymbol{\varphi} \triangleq \boldsymbol{\varphi}(\boldsymbol{\mu})$ be its profile. Moreover, let $\mathcal{Q}$, $\mathcal{Q}_{\mathrm{low}}$, $\mathcal{Q}_{\mathrm{high}}$, and $\mathcal{Q}'_{\mathrm{low}}$ be defined as just mentioned. Assume that $\mathcal{Q}$ has size L and that $\mathcal{Q} = \{q_1, \ldots, q_L\}$. Let $\boldsymbol{v}' = (v'_1, \ldots, v'_L)$ be a vector that satisfies the following constraints*

$$v'_\ell \geqslant 0, \quad \ell \in [L], \tag{19}$$

$$\sum_\ell v'_\ell \cdot q_\ell = 1, \tag{20}$$

$$v'_\ell = \varphi_{nq_\ell}, \quad \ell : q_\ell \in \mathcal{Q}_{\mathrm{high}}, \tag{21}$$

$$\sum_{\ell:\, q_\ell \in \mathcal{Q}'_{\mathrm{low}}} v'_\ell \cdot q_\ell \leqslant f_1(n), \tag{22}$$

$$\sum_{\ell:\, q_\ell \in \mathcal{Q}_{\mathrm{low}}} v'_\ell \cdot \mathrm{e}^{-nq_\ell} \frac{(nq_\ell)^\mu}{\mu!} \in \big[\varphi_\mu - f_2(n), \varphi_\mu + f_2(n)\big], \quad \mu \in \mathbb{Z}_{\geqslant 0}, \mu \leqslant nq', \tag{23}$$

where $f_1(n)$ and $f_2(n)$ are certain functions with sublinear growth and where $q' \triangleq \frac{1}{2}(\min(\mathcal{Q}'_{\mathrm{low}}) + \max(\mathcal{Q}'_{\mathrm{low}}))$. The (potentially improper) VV distribution histogram estimate $v^{\mathrm{VV}'} : [0,1] \to \mathbb{R}_{\geqslant 0}$ is then given by

$$v^{\mathrm{VV}'}(q) \triangleq \begin{cases} v'_\ell & \text{(if } q = q_\ell \text{ for some } \ell \in [L]\text{)} \\ 0 & \text{(otherwise)} \end{cases}. \tag{24}$$

A proper distribution histogram estimate $v^{\mathrm{VV}} : [0,1] \to \mathbb{Z}_{\geqslant 0}$ is obtained from $v^{\mathrm{VV}'}$ with the help of a suitable rounding and adjustment procedure. (We refer to [4] for details.) □

Here, the constraints in (19) and (20) ensure that $v^{\mathrm{VV}'}$ is a (potentially improper) distribution histogram; the constraints in (21) follow from our comments on "large" μ_j values; the constraint in (22) is included to avoid certain issues that might otherwise arise for v_ℓ with $q_\ell \in \mathcal{Q}_{\mathrm{low}}$ close to the lowest value of $\mathcal{Q}_{\mathrm{high}}$; finally, the constraints in (23) are similar to the constraints in (17).

Based on Definition 21, we can state a remark whose content is similar to the content of Remark 20 and where $\boldsymbol{\gamma}^*$ is chosen as follows

$$\gamma^*_{i,j} \triangleq \begin{cases} \frac{1}{\varphi_{\mu_j}} & (i,j) \in [k]^2 \text{ s.t. } p_i \in \mathcal{Q}_{\mathrm{high}},\ p_i = \frac{\mu_j}{n} \\ 0 & (i,j) \in [k]^2 \text{ s.t. } p_i \in \mathcal{Q}_{\mathrm{high}},\ p_i \neq \frac{\mu_j}{n} \\ \mathrm{e}^{-np_i} \cdot \frac{(np_i)^{\mu_j}}{\mu_j! \cdot \varphi_{\mu_j}} & (i,j) \in [k]^2 \text{ s.t. } p_i \in \mathcal{Q}_{\mathrm{low}} \end{cases}$$

(Alternatively, one can also set $\gamma^*_{i,j} \triangleq 0$ for $(i,j) \in [k]^2$ s.t. $p_i \in \mathcal{Q}_{\mathrm{low}}$, $\mu_j \geqslant n \cdot \min(\mathcal{Q}_{\mathrm{high}})$.) We omit further details.

IV. Conclusion

This paper discussed various approximations of the PML estimate and showed connections between them. These connections were established by associating an approximately doubly stochastic matrix $\boldsymbol{\gamma}^*$ with the Valiant–Valiant estimate and comparing $\boldsymbol{\gamma}^*$ with the doubly stochastic matrices $\boldsymbol{\gamma}$ that appear in the description of the PML estimate and its approximations, in particular the doubly stochastic matrix $\boldsymbol{\gamma}$ that appears in the description of the SPML estimate. Further connections will be discussed in [20].

Appendix A
Proof of Theorem 15

The proof of the first part of the theorem follows from a standard application of Lagrange multipliers (note that there is no need for Lagrange multipliers that enforce the non-negativity constraints on the entries of $\boldsymbol{\gamma}^*$ since the solution automatically satisfies these constraints) and from the properties of the Sinkhorn balancing algorithm [30], [31].

Towards proving the second part of the theorem, fix a matrix $\boldsymbol{\gamma} \in \Gamma_{k \times k}$ and define

$$\boldsymbol{p}^* \triangleq \arg\min_{\boldsymbol{p}} \; F_{\mathrm{S}}\big(\boldsymbol{\gamma}; \boldsymbol{\theta}(\boldsymbol{p}, \boldsymbol{\mu})\big).$$

Introducing the Lagrange multiplier λ'' for the constraint $\sum_{i \in [k]} p_i = 1$, we see that $\boldsymbol{p}^*$ must satisfy

$$0 \stackrel{!}{=} \frac{\partial}{\partial p_i} \left(F_{\mathrm{S}}\big(\boldsymbol{\gamma}, \boldsymbol{\theta}(\boldsymbol{p}, \boldsymbol{\mu})\big) + \lambda'' \cdot \left(\sum_{i'} p_{i'} - 1 \right) \right) \Bigg|_{\boldsymbol{p} = \boldsymbol{p}^*}$$

$$\stackrel{\text{(a)}}{=} - \sum_j \frac{\gamma_{i,j} \cdot \mu_j}{p_i} + \lambda'' \Bigg|_{\boldsymbol{p} = \boldsymbol{p}^*}.$$

Here, step (a) used Definition 13, the fact that

$$U'_{\mathrm{S}}\Big(\boldsymbol{\gamma}, \boldsymbol{\theta}\big(\boldsymbol{p}, \boldsymbol{\mu}(\boldsymbol{\psi})\big)\Big) = -\sum_{i,j} \gamma_{i,j} \log\big(p_i^{\mu_j}\big) = -\sum_{i,j} \gamma_{i,j} \mu_j \log(p_i),$$

and the fact that $H_{\mathrm{S}}(\boldsymbol{\gamma})$ does not depend on $\boldsymbol{p}$. Solving for $\boldsymbol{p}^*$, we obtain

$$p^*_i = \frac{\sum_j \gamma_{i,j} \cdot \mu_j}{\sum_{i'} \sum_j \gamma_{i',j} \cdot \mu_j}, \quad i \in [k].$$

Observing that $\sum_{i'} \sum_j \gamma_{i',j} \cdot \mu_j = \sum_j \mu_j \cdot \sum_{i'} \gamma_{i',j} = \sum_j \mu_j = n$, we can rewrite this as

$$p^*_i = \sum_j \gamma_{i,j} \cdot \frac{\mu_j}{n}, \quad i \in [k],$$

which is the expression that appears in the theorem statement.

Appendix B
Proof of Remark 20

In the first part of this appendix we show that the matrix $\boldsymbol{\gamma}^*$ is approximately doubly stochastic. Clearly, all the entries of $\boldsymbol{\gamma}^*$ are non-negative. Therefore, let us proceed to the investigation of the row and column sums.

Let $\mathcal{M}$ be the set of integers μ for which there is at least one $j \in [k]$ such that $\mu_j = \mu$. For $i \in [k]$, the i-th row sum of $\boldsymbol{\gamma}^*$ is then

$$\begin{aligned}\sum_j \gamma^*_{i,j} &= \sum_j \mathrm{e}^{-np^*_i} \cdot \frac{(np^*_i)^{\mu_j}}{\mu_j! \cdot \varphi_{\mu_j}} \\ &= \mathrm{e}^{-np^*_i} \cdot \sum_j \frac{(np^*_i)^{\mu_j}}{\mu_j! \cdot \varphi_{\mu_j}} \\ &= \mathrm{e}^{-np^*_i} \cdot \sum_{\mu \in \mathcal{M}} \sum_{j\colon \mu_j = \mu} \frac{(np^*_i)^{\mu}}{\mu! \cdot \varphi_{\mu}} \\ &= \mathrm{e}^{-np^*_i} \cdot \sum_{\mu \in \mathcal{M}} \frac{(np^*_i)^{\mu}}{\mu!} \\ &\overset{\text{(a)}}{\approx} \mathrm{e}^{-np^*_i} \cdot e^{np^*_i} \\ &= 1,\end{aligned}$$

where at step (a) we have assumed that $\mathcal{M}$ contains the relevant integers μ so that the approximation is allowed. (Note that $\mu \mapsto e^{-np^*_i} \frac{(np^*_i)^{\mu}}{\mu!}$ is the probability mass function of a Poisson random variable with mean np^*_i.) Similarly, for $j \in [k]$, the j-the column sum of $\boldsymbol{\gamma}^*$ is

$$\begin{aligned}\sum_i \gamma^*_{i,j} &= \sum_i \mathrm{e}^{-np^*_i} \cdot \frac{(np^*_i)^{\mu_j}}{\mu_j! \cdot \varphi_{\mu_j}} \\ &= \sum_{q \in \mathcal{Q}} v(q) \cdot \mathrm{e}^{-nq} \frac{(nq)^{\mu_j}}{\mu_j! \cdot \varphi_{\mu_j}} \\ &= \frac{1}{\varphi_{\mu_j}} \cdot \sum_{q \in \mathcal{Q}} v(q) \cdot \mathrm{e}^{-nq} \frac{(nq)^{\mu_j}}{\mu_j!} \\ &\overset{\text{(a)}}{\approx} \frac{1}{\varphi_{\mu_j}} \cdot \varphi_{\mu_j} \\ &= 1,\end{aligned}$$

where step (a) follows from (17).

We come now to the second part of this appendix. According to Theorem 15, a fixed point $(\boldsymbol{p}^*, \boldsymbol{\gamma}^*)$ of Algorithm 16 satisfies

$$\boldsymbol{\gamma}^* = \arg\min_{\boldsymbol{\gamma}} \; F_{\mathrm{S}}\big(\boldsymbol{\gamma}; \boldsymbol{\theta}(\boldsymbol{p}^*, \boldsymbol{\mu})\big), \tag{25}$$

$$\boldsymbol{p}^* = \arg\min_{\boldsymbol{p}} \; F_{\mathrm{S}}\big(\boldsymbol{\gamma}^*; \boldsymbol{\theta}(\boldsymbol{p}, \mu)\big) = \boldsymbol{\gamma}^* \cdot \frac{\boldsymbol{\mu}}{n}. \tag{26}$$

In the following, we want to show that the vector-matrix pair $(\boldsymbol{p}^*, \boldsymbol{\gamma}^*)$ given in the remark statement approximately satisfies (25) and (26).

The following argument shows that (26) is approximately satisfied. Namely, for any $i \in [k]$ we have

$$\begin{aligned}\sum_j \gamma^*_{i,j} \cdot \frac{\mu_j}{n} &= \sum_j \mathrm{e}^{-np^*_i} \cdot \frac{(np^*_i)^{\mu_j}}{\mu_j! \cdot \varphi_{\mu_j}} \cdot \frac{\mu_j}{n} \\ &= p^*_i \cdot \mathrm{e}^{-np^*_i} \cdot \sum_{j\colon \mu_j > 0} \frac{(np^*_i)^{\mu_j - 1}}{(\mu_j - 1)! \cdot \varphi_{\mu_j}} \\ &= p^*_i \cdot \mathrm{e}^{-np^*_i} \cdot \sum_{\mu \in \mathcal{M}\colon \mu > 0} \sum_{j\colon \mu_j = \mu} \frac{(np^*_i)^{\mu - 1}}{(\mu - 1)! \cdot \varphi_{\mu}} \\ &= p^*_i \cdot \mathrm{e}^{-np^*_i} \cdot \sum_{\mu \in \mathcal{M}\colon \mu > 0} \frac{(np^*_i)^{\mu - 1}}{(\mu - 1)!} \\ &\overset{\text{(a)}}{=} p^*_i \cdot \mathrm{e}^{-np^*_i} \cdot \sum_{\mu \in \mathcal{M}'} \frac{(np^*_i)^{\mu}}{\mu!} \\ &\overset{\text{(b)}}{\approx} p^*_i \cdot \mathrm{e}^{-np^*_i} \cdot e^{np^*_i} \\ &= p^*_i,\end{aligned}$$

where $\mathcal{M}' \triangleq \{\mu \in \mathbb{Z}_{\geqslant 0} \mid \mu + 1 \in \mathcal{M}\}$ was introduced at step (a) and where the reasoning at step (b) is similar to the reasoning that was used while computing the row sums of $\boldsymbol{\gamma}^*$.

Let $\boldsymbol{\theta}^* \triangleq \boldsymbol{\theta}(\boldsymbol{p}^*, \boldsymbol{\mu})$ with components $\boldsymbol{\theta}^*_{i,j} = (p^*_i)^{\mu_j}$, $(i,j) \in [k]^2$. From Theorem 15 it follows that condition (25) is equivalent to the condition that $\boldsymbol{\gamma}^*$ is a doubly stochastic matrix and the existence of Lagrange multipliers $(\lambda_i)_{i \in [k]}$, $(\lambda'_j)_{j \in [k]}$ such that

$$\gamma^*_{i,j} \triangleq \theta^*_{i,j} \cdot \exp(-\lambda_i - \lambda'_j) \quad (i,j) \in [k]^2. \tag{27}$$

As shown earlier, $\boldsymbol{\gamma}^*$ is approximately doubly stochastic and the choice

$$\begin{aligned}\lambda_i &\triangleq np^*_i, && i \in [k], \\ \lambda'_j &\triangleq -\mu_j \log(n) + \log(\mu_j! \cdot \varphi_{\mu_j}) && j \in [k],\end{aligned}$$

shows that (27) can be satisfied with equality.

Although not required for the proof of the remark statement, it is informative to compute $F_{\mathrm{S}}\big(\boldsymbol{\gamma}^*; \boldsymbol{\theta}(\boldsymbol{p}^*, \boldsymbol{\mu})\big)$. Namely,

$$\begin{aligned}&F_{\mathrm{S}}\big(\boldsymbol{\gamma}^*; \boldsymbol{\theta}(\boldsymbol{p}^*, \boldsymbol{\mu})\big) \\ &\quad = -\sum_{i,j} \gamma_{i,j} \log\big((p^*_i)^{\mu_j}\big) + \sum_{i,j} \gamma^*_{i,j} \log(\gamma^*_{i,j}) \\ &\quad = \sum_{i,j} \gamma^*_{i,j} \log\left(\mathrm{e}^{-np^*_i} \cdot \frac{n^{\mu_j}}{\mu_j! \cdot \varphi_{\mu_j}}\right) \\ &\quad \overset{\text{(a)}}{\approx} -n + n\log(n) - \sum_j \log\big(\mu_j! \cdot \varphi_{\mu_j}\big)\end{aligned}$$

where at step (a) we have used the approximate doubly stochasticity of $\boldsymbol{\gamma}^*$, along with $\sum_j \mu_j = n$. Note that the last expression is a function only of $\boldsymbol{\mu}$ and $\boldsymbol{\varphi} \triangleq \boldsymbol{\varphi}(\boldsymbol{\mu})$, which corroborates the above findings about $\boldsymbol{p}^*$ and $\boldsymbol{\gamma}^*$. In particular, $F_{\mathrm{S}}\big(\boldsymbol{\gamma}^*; \boldsymbol{\theta}(\boldsymbol{p}^*, \boldsymbol{\mu})\big)$ is approximately constant for all vector-matrix pairs $(\boldsymbol{p}^*, \boldsymbol{\gamma}^*)$ derived from distribution histograms satisfying the constraints in Definition 19.

Acknowledgment

I am indebted to Krishna Viswanathan for introducing me to the topic of the PML estimate and for answering all my subsequent questions.

References

[1] R. A. Fisher, A. Corbet, and C. B. Williams, "The relation between the number of species and the number of individuals in a random sample of an animal population," *J. British Ecological Society*, vol. 12, no. 1, pp. 42–58, May 1943.

[2] I. J. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, no. 16, pp. 237–264, Dec. 1953.

[3] A. Orlitsky, N. P. Santhanam, K. Viswanathan, and J. Zhang, "On estimating the probability multiset," *draft*, Jun. 2011, available online at `http://www-ee.eng.hawaii.edu/~prasadsn/skelnew59.pdf`.

[4] G. Valiant and P. Valiant, "Estimating the unseen: a sublinear-sample canonical estimator of distributions," in *Proc. 43rd Annual ACM Symp. Theory of Computing*, San Jose, CA, USA, June 6–8 2011, extended version available at `http://eccc.hpi-web.de/report/2010/179/` and `http://eccc.hpi-web.de/report/2010/180/`.

[5] J. Bunge and K. Barger, "Estimating the number of classes in a population," archive bibliography, `http://www.stat.cornell.edu/~bunge/bibliography.html`, Cornell University.

[6] R. Rubinfeld, "Taming big probability distributions," *XRDS: Crossroads, The ACM Magazine for Students*, vol. 19, no. 1, pp. 24–28, Fall 2012.

[7] G. Valiant and P. Valiant, "The power of linear estimators," in *Proc. 52nd Ann. Symp. on Foundations of Computer Science*, Palm Springs, CA, USA, Oct. 22–25 2011.

[8] A. Orlitsky, N. P. Santhanam, K. Viswanathan, and J. Zhang, "On modeling profiles instead of values," in *Proc. Uncertainty in Artificial Intelligence*, Banff, Canada, Jul. 7–11 2004.

[9] A. Orlitsky, S. Pan, Sajama, N. P. Santhanam, K. Viswanathan, and J. Zhang, "Approximation of the pattern maximum likelihood distribution," *draft*, Oct. 2011.

[10] A. Orlitsky and S. Pan, "The maximum likelihood probability of skewed patterns," in *Proc. IEEE Int. Symp. Inf. Theory*, Seoul, Korea, June 28–July 3 2009.

[11] A. Orlitsky, N. P. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *IEEE Trans. Inf. Theory*, vol. 50, no. 7, pp. 1469–1481, Jul. 2004.

[12] G. I. Shamir, "Universal lossless compression with unknown alphabets — the average case," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 4915–4944, Nov. 2006.

[13] A. B. Wagner, P. Viswanath, and S. R. Kulkarni, "A better Good-Turing estimator for sequence probabilities," in *Proc. IEEE Int. Symp. Inf. Theory*, Nice, France, June 24–29 2007, pp. 2356–2360.

[14] J. Acharya, A. Orlitsky, and S. Pan, "Recent results on pattern maximum likelihood," in *Proc. IEEE Inf. Theory Workshop*, Volos, Greece, Jun. 10–12 2009.

[15] A. B. Wagner, P. Viswanath, and S. R. Kulkarni, "Probability estimation in the rare-events regime," *IEEE Trans. Inf. Theory*, vol. 57, no. 6, pp. 3207–3229, Jun. 2011.

[16] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and S. Pan, "Estimating multiple concurrent processes," in *Proc. IEEE Int. Symp. Inf. Theory*, Cambridge, MA, USA, Jul. 1–6 2012, pp. 1628–1632.

[17] W. Fernandes and N. Kashyap, "A phase transition for the uniform distribution in the pattern maximum likelihood problem," in *Proc. IEEE Inf. Theory Workshop*, Seville, Spain, Sep. 9–13 2013, pp. 1–5.

[18] K. Viswanathan, "Pattern maximum-likelihood," Talk at Workshop on "Permanents and modeling probability distributions," American Institute of Mathematics, Palo Alto, CA, USA, Sep. 1 2009.

[19] H. Minc, *Permanents*. Reading, MA: Addison-Wesley, 1978.

[20] P. O. Vontobel, "Fast approximations of the pattern maximum likelihood estimate," *in preparation, to be submitted to IEEE Trans. Inform. Theory*, 2014.

[21] A. Orlitsky, Sajama, N. P. Santhanam, K. Viswanathan, and J. Zhang, "Algorithms for modeling distributions over large alphabets," in *Proc. IEEE Int. Symp. Inf. Theory*, Chicago, IL, USA, June 27–July 2 2004, p. 304.

[22] P. O. Vontobel, "The Bethe approximation of the pattern maximum likelihood distribution," in *Proc. IEEE Int. Symp. Inf. Theory*, Cambridge, MA, USA, Jul. 1–6 2012, pp. 2012–2016.

[23] M. Chertkov, L. Kroc, and M. Vergassola, "Belief propagation and beyond for particle tracking," *CoRR, available online under* `http://arxiv.org/abs/0806.1199`, Jun. 2008.

[24] M. Chertkov, L. Kroc, F. Krzakala, M. Vergassola, and L. Zdeborová, "Inference in particle tracking experiments by passing messsages between images," *Proc. Natl. Acad. Sci.*, vol. 107, no. 17, pp. 7663–7668, Apr. 2010.

[25] B. Huang and T. Jebara, "Approximating the permanent with belief propagation," *CoRR, available online under* `http://arxiv.org/abs/0908.1769`, Aug. 2009.

[26] Y. Watanabe and M. Chertkov, "Belief propagation and loop calculus for the permanent of a non-negative matrix," *Journal of Physics A: Mathematical and Theoretical*, vol. 43, p. 242002, 2010.

[27] P. O. Vontobel, "The Bethe permanent of a non-negative matrix," *IEEE Trans. Inf. Theory*, vol. 59, no. 3, pp. 1866–1901, Mar. 2013.

[28] L. Gurvits, "Unleashing the power of Schrijver's permanental inequality with the help of the Bethe approximation," *Elec. Coll. Comp. Compl.*, Dec. 2011.

[29] M. Chertkov and A. B. Yedidia, "Approximating the permanent with fractional belief propagation," *J. Mach. Learn. Res.*, vol. 14, pp. 2029–2066, Jul. 2013.

[30] R. Sinkhorn, "A relationship between arbitrary positive matrices and doubly stochastic matrices," *Ann. Math. Statist.*, vol. 35, no. 2, pp. 876–879, Jun. 1964.

[31] N. Linial, A. Samorodnitsky, and A. Wigderson, "A deterministic strongly polynomial algorithm for matrix scaling and approximate permanents," *Combinatorica*, vol. 20, no. 4, pp. 545–568, 2000.

[32] D. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA: Athena Scientific, 1999.

[33] G. J. Valiant, "Algorithmic approaches to statistical questions," Ph.D. dissertation, Dept. of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA, USA, 2012.