2nd Symposium on Advances in Approximate Bayesian Inference, 2019 1–9

Sinkhorn Permutation Variational Marginal Inference

Gonzalo Mena GOMENA@FAS.HARVARD.EDU
Harvard University

Erdem Varol, Amin Nejatbakhsh, Eviatar Yemini, Liam Paninski
Columbia University

Abstract

We address the problem of marginal inference for an exponential family defined over the set of permutation matrices. This problem is known to quickly become intractable as the size of the permutation increases, since its involves the computation of the permanent of a matrix, a #P-hard problem. We introduce Sinkhorn variational marginal inference as a scalable alternative, a method whose validity is ultimately justified by the so-called Sinkhorn approximation of the permanent. We demonstrate the effectiveness of our method in the problem of probabilistic identification of neurons in the worm *C.elegans*.

1. Introduction

Let $P \in \mathbb{R}^{n\times n}$ be a binary matrix representing a permutation of n elements (i.e. each row and column of P contains a unique **1**). We consider the distribution over P defined as

$$p(P|L) = \frac{1}{Z_L} \exp\left(\langle \log L, P\rangle_F\right), \tag{1.1}$$

where $\langle A, B\rangle_F$ is the Frobenius matrix inner product, $\log L$ is a parameter matrix and Z_L is the normalizing constant. Here we address the problem of marginal inference, i.e. computing the matrix of expectations $\rho := E(P)$. This problem is known to be intractable since it requires access to Z_L, also known as the permanent of L, and whose computation is known to be a #P-hard problem Valiant (1979)

To overcome this difficulty we introduce Sinkhorn variational marginal inference, which can be computed efficiently and is straightforward to implement. Specifically, we approximate ρ as $S(L)$, the Sinkhorn operator applied to L (Sinkhorn, 1964). $S(L)$ is defined as the (infinite) successive row and column normalization of L (Adams and Zemel, 2011; Linderman et al., 2018), a limit that is known to result in a doubly stochastic matrix (Altschuler et al., 2017). In section 2 we argue the Sinkhorn approximation is sensible, and in section 3 we describe the problem of probabilistic inference of neural identity in *C.elegans* and demonstrate the Sinkhorn approximation produces the best results.

2. Sinkhorn permutation variational marginal inference

Our argument bases on the well-known relation between marginal inference and the normalizing constant (Wainwright and Jordan, 2008), valid for exponential families. Specifically, (1.1) defines

an exponential family with sufficient statistic P and parameter $\log L$. By virtue of Theorem 3.4 in Wainwright and Jordan (2008):

$$\log Z_L = \sup_{\mu \in \mathcal{M}} \langle \log L, \mu \rangle - A^*(\mu), \tag{2.1}$$

where $\mathcal{M}$ is the marginal polytope (here, the Birkhoff polytope, the set of doubly stochastic matrices) and $A^*(\mu)$ is the dual function $\log Z_L$, i.e.

$$A^*(\mu) = \sup_{L} \langle \log L, \mu \rangle - \log Z_L. \tag{2.2}$$

Moreover, for a given L, $\mu(L)$ achieving the supremum in (2.1) is exactly the matrix of marginals, $\mu(L) = \rho^L$ and the dual function $A^*(\mu(L))$ coincides with the negative entropy of (1.1). Then, marginal inference of ρ^L and computation of the permanent $Z_L = perm(L)$ are linked by the optimization problem in (2.2).

As in any generic variational inference scheme Wainwright and Jordan (2008), we obtain an approximate ρ by replacing the variational representation of Z_L in (2.1), by a different, more tractable optimization problem. Typically, the quality of the approximated ρ depends on how tight is the approximation to Z_L. Our approximation is based on replacing the intractable dual function $A^*(\mu)$ by a component-wise entropy, and whose solution is exactly $S(L)$. In detail, the following variational representation holds (Mena et al., 2018; Helmbold and Warmuth, 2009):

$$S(L) = \arg \sup_{\mu \in \mathcal{M}} \langle \log L, \mu \rangle - \sum_{i,j} \mu_{i,j} \log \mu_{i,j}. \tag{2.3}$$

By using the component-wise entropy in (2.1) we obtain an approximation of the normalizing constant, that we call as the Sinkhorn permanent (Linial et al., 2000), $perm_S(L)$. In the following proposition we provide bounds for this approximation.

Proposition 1 *The following bounds hold*

$$perm(L) \leq perm_S(L) \leq e^n perm(L). \tag{2.4}$$

We note the Sinkhorn approximation has recently been proposed independently (Powell and Smith, 2019). However, there the approximation is proposed rather heuristically, without any appeal to a theoretical framework.

2.1. Related work

Additionally, the so-called Bethe variational inference method (Wainwright and Jordan, 2008) is a rather general rationale for obtaining variational approximations in graphical models, where the dual function $A^*(\mu)$ is approximated by the value it would take if the underlying Markov random field had a tree structure (Yedidia et al., 2001). This approximation has successfully been applied to permutations (Huang and Jebara, 2007; Chertkov et al., 2010; Vontobel, 2014; Tang et al., 2015), where the corresponding approximate marginal $B(L)$ is computed through belief propagation (Huang and Jebara, 2007; Vontobel, 2013), enjoying also better theoretical guarantees

than the Sinkhorn approximation. Indeed, for the Bethe approximation of the permanent, $perm_B(\cdot)$ the following bounds are known (Gurvits and Samorodnitsky, 2014; Anari and Rezaei, 2018)

$$\sqrt{2}^{-n} perm(L) \leq perm_B(L) \leq perm(L). \tag{2.5}$$

However, there are also important computational differences. A single iteration of the Sinkhorn algorithms corresponds to a row and column normalization, but the message computations in the belief propagation-like routine for the Bethe approximation are more complex. Explicit formulae of such Sinkhorn and Bethe iterations are available in Appendix C.

Fig 1(b) shows that in practice the Bethe approximation also produces better permanent approximations, confirming theoretical predictions. We considered the simple case where $n = 8$ and the permanent and marginal can be computed by enumeration, so comparisons with ground truth are possible. However, and quite interestingly, in many cases (see Figs 1(a) and A.1(a) in the Appendix) the Sinkhorn approximation produced qualitatively better marginals, putting more mass on more non-zero entries than the Bethe approximation, regardless of possibly worse permanents.

Additionally, we observed that for moderate n the Sinkhorn approximation scaled better. For example, if $n = 710$, each Bethe iteration took on average 0.035 seconds, while each Sinkhorn iteration took only 0.0027 seconds (see Fig A.2 in the Appendix for details).

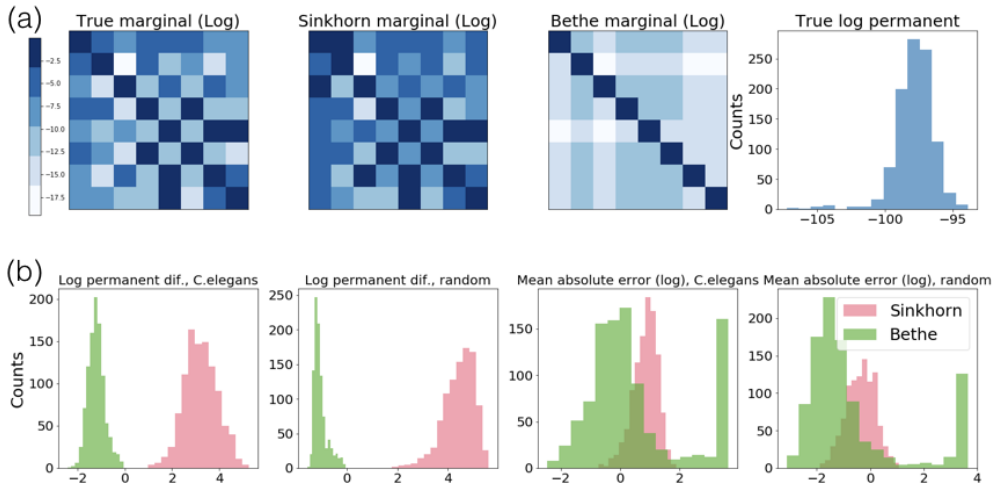

Figure 1: Comparison of Bethe and Sinkhorn approximations. 1,000 submatrices of size $n = 8$ were randomly sampled from the *C.elegans* dataset described in section 3. (a) Examples of a (log) true marginal matrix ρ along with Sinkhorn and Bette approximation. The rightmost plot is a histogram of the log permanent across the samples. (b) Differences between approximate and true log permanent (left) and mean absolute errors of log marginals (right) for our two approximations. We considered additional 1,000 'random' submatrices made by uniformly sampling entries between the minimum and maximum values of each *C.elegans* submatrix

Finally, we note that sampling-based methods may be also used for marginal inference. Indeed, quite sophisticated samplers have been proposed to show polynomial approximability of the permanent (Jerrum and Sinclair, 1989); however, their practical appeal is limited. In section 3 we show that an elementary MCMC sampler failed to produce sensible marginal inferences at reasonable time.

3. Probabilistic inference of neurons in *C.elegans*

The worm *C.elegans* is a unique species since their nervous system is stereotypical; i.e., the number of neurons (roughly, 300) and the connections between those neurons remain unchanged from animal to animal. Recent advances in neurotechnology have enabled whole brain imaging so that the long-standing fundamental question about how the activity in the worm brain relates to its behavior in the world can be now studied and settled. However, before that, a technical problem has to be solved: given volumetric images of the worm neurons have to be identified; that is, canonical labels (names) must be assigned to each.

We applied our methodology for such *probabilistic* neural identification in the context of NeuroPAL (Yemini et al., 2019), a multicolor *C.elegans* transgene where neuron colors were designed to facilitate neural identification (see Fig 1 for an example). Specifically, given n observed neurons represented as vectors in $\mathbb{R}^6$ (position and color), we aim to estimate the matrix of marginal ρ such that $\rho_{k,i}$ is the probability that observed neuron k is identified with the canonical identity i. These probabilities are relevant as they provide *uncertainty* estimates for model predictions, giving a much more complete picture than point estimates (e.g. a permutation found via maximum likelihood).

We consider a gaussian model for each canonical neuron, whose parameters (μ_k, Σ_k) are inferred beforehand from previously annotated worms (see Yemini et al. (2019) for details). Let π denote the permutation so that $\pi(k)$ is the canonical index of the $k-th$ observed neuron. Then, the likelihood of observing data $Y = (y_k)$ writes as:

$$p(Y|P, \mu, \Sigma) = \prod_{k=1}^{n} \mathcal{N}\left(y_k; \mu_{\pi(k)}, \Sigma_{\pi(k)}\right). \tag{3.1}$$

Suppose a flat prior is assumed over P. Then, it is plain to verify that equation (3.1) induces a posterior over P that has the form of (1.1), with L defined as

$$\log L_{k,i} = -\frac{1}{2}(y_k - \mu_i)^T \Sigma_i^{-1}(y_k - \mu_i). \tag{3.2}$$

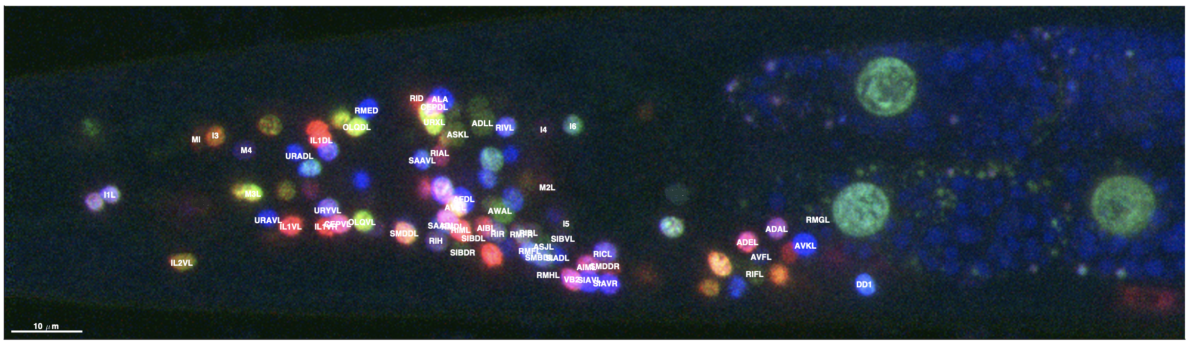

Figure 2: A worm's head displaying the deterministic coloring scheme identical across all NeuroPAL worms, with neuron names (determined by a human) over each neuron.

3.1. Results

In the context of NeuroPAL we consider a downstream task involving the computation of the approximate probabilistic neural identifies ρ. Specifically, in this task a human is asked to manually

label the neurons for which the model estimates are the most uncertain; i.e., the rows of ρ that are closest to the uniform distribution. As the human progressively annotates neurons this uncertainty resolves and the corresponding model update lead to an increases in identification accuracy for the remaining neurons. Ideally the human will only require a few annotations to reach a high accuracy, and therefore, as a proxy for approximation quality we measure how much faster accuracy increases in comparison to simple baselines; e.g., where at each time a neuron is randomly chosen.

Results are shown in Fig 3, and further details are described in the Appendix. We considered several alternatives: i) Sinkhorn approximation, ii) Bethe approximation, iii) MCMC, iv) the random baseline described above, v) a naive baseline where uncertainty estimates are made by scaling only the rows of the likelihood matrix, i.e., without imposing any one-to-one assignment structure, and vi) a 'ground truth', the protocol where the labels that are chosen are the ones where the model makes a wrong prediction (this oracle cannot be realized in practice). Results of Sinkhorn and Bethe approximations are similar but the former slightly better, presumably a consequence of more accurate estimates of low probability marginals (see Figs 1(a) and A.1(a)). They both are substantially better than any baseline other than the oracle. Contrarily, we see MCMC does not provide better results than the naive baseline, suggesting lack of convergence for chain lengths leading to computational times comparable to the ones of approximated methods.

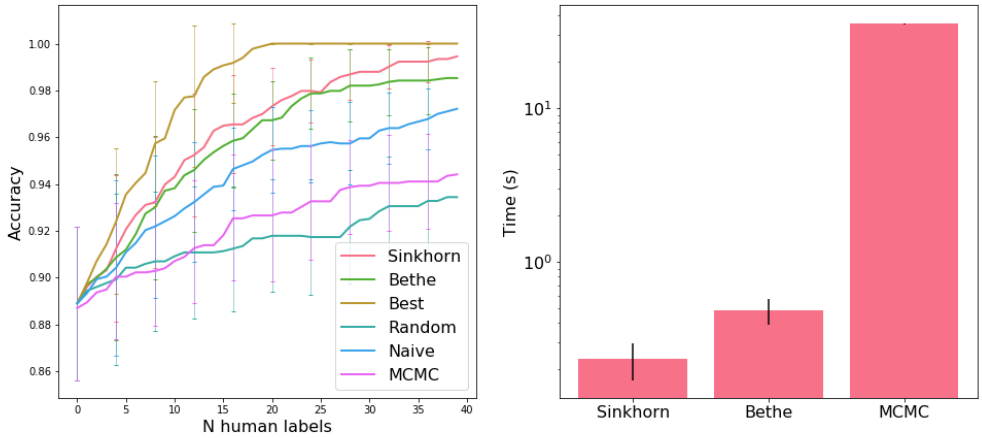

Figure 3: Results in the neural identification downstream task. Here, $n \approx 180$. Left: mean accuracy (standard deviation) as a number of human labels for the different ways for suggesting uncertain neurons. Right: average times (seconds) for computing the ρ matrix.

4. Conclusion

We have introduced the Sinkhorn approximation for marginal inference, and our it is a sensible alternative to sampling, and it may provide faster, simpler and more accurate approximate marginals than the Bethe approximation, despite typically leading to worse permanent approximations. We leave for future work a thorough analysis of the relation between quality of permanent approximation and corresponding marginals.

References

Ryan Prescott Adams and Richard S Zemel. Ranking via sinkhorn propagation. *arXiv preprint arXiv:1106.1925*, 2011.

Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pages 1964–1974, 2017.

Nima Anari and Alireza Rezaei. A tight analysis of bethe approximation for permanent. *arXiv preprint arXiv:1811.02933*, 2018.

Michael Chertkov, Lukas Kroc, F Krzakala, M Vergassola, and L Zdeborová. Inference in particle tracking experiments by passing messages between images. *Proceedings of the National Academy of Sciences*, 107(17):7663–7668, 2010.

Persi Diaconis. The markov chain monte carlo revolution. *Bulletin of the American Mathematical Society*, 46(2):179–205, 2009.

Leonid Gurvits and Alex Samorodnitsky. Bounds on the permanent and some applications. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 90–99. IEEE, 2014.

David P Helmbold and Manfred K Warmuth. Learning permutations with exponential weights. *Journal of Machine Learning Research*, 10(Jul):1705–1736, 2009.

Bert Huang and Tony Jebara. Loopy belief propagation for bipartite maximum weight b-matching. In *Artificial Intelligence and Statistics*, pages 195–202, 2007.

Mark Jerrum and Alistair Sinclair. Approximating the permanent. *SIAM journal on computing*, 18 (6):1149–1178, 1989.

Scott Linderman, Gonzalo Mena, Hal Cooper, Liam Paninski, and John Cunningham. Reparameterizing the birkhoff polytope for variational permutation inference. In *International Conference on Artificial Intelligence and Statistics*, pages 1618–1627, 2018.

Nathan Linial, Alex Samorodnitsky, and Avi Wigderson. A deterministic strongly polynomial algorithm for matrix scaling and approximate permanents. *Combinatorica*, 20(4):545–568, 2000.

Gonzalo Mena, David Belanger, Scott Linderman, and Jasper Snoek. Learning latent permutations with gumbel-sinkhorn networks. In *International Conference on Learning Representations*, 2018.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Pascal Pontobel. Personal Communication, December 2019.

Ben Powell and Paul A Smith. Computing expectations and marginal likelihoods for permutations. *Computational Statistics*, pages 1–21, 2019.

Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.

Kui Tang, Nicholas Ruozzi, David Belanger, and Tony Jebara. Bethe learning of conditional random fields via map decoding. *arXiv preprint arXiv:1503.01228*, 2015.

Leslie G Valiant. The complexity of computing the permanent. *Theoretical computer science*, 8(2): 189–201, 1979.

Pascal O Vontobel. The bethe permanent of a nonnegative matrix. *IEEE Transactions on Information Theory*, 59(3):1866–1901, 2013.

Pascal O Vontobel. The bethe and sinkhorn approximations of the pattern maximum likelihood estimate and their connections to the valiant-valiant estimate. In *ITA*, pages 1–10, 2014.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Jonathan S Yedidia, William T Freeman, and Yair Weiss. Bethe free energy, kikuchi approximations, and belief propagation algorithms. *Advances in neural information processing systems*, 13, 2001.

Eviatar Yemini, Albert Lin, Amin Nejatbakhsh, Erdem Varol, Ruoxi Sun, Gonzalo E Mena, Aravinthan DT Samuel, Liam Paninski, Vivek Venkatachalam, and Oliver Hobert. Neuropal: A neuronal polychromatic atlas of landmarks for whole-brain imaging in c. elegans. *bioRxiv*, page 676312, 2019.

Appendix A. Proof of proposition 1

Proof We essentially condense the arguments in Linial et al. (2000). First, we use the fact that the permanent of a doubly stochastic matrix B of size n satisfies (Linial et al., 2000):

$$e^{-n} \leq perm(B) \leq 1.$$

Also, it can be verified that $S(L) = diag(x)Ldiag(y)$, where $diag(x), diag(y)$ are some positive vectors x, y turned into diagonal matrices (Peyré et al., 2019). Then,

$$perm(S(L)) = \left(\prod_{i=1}^{n} x_i\right)\left(\prod_{i=1}^{n} y_i\right) perm(L).$$

Additionally, we obtain the (log) Sinkhorn approximation of the permanent of L, $perm_S(L)$, by evaluating $S(L)$ in the problem it solves, (2.3). By simple algebra and using the fact that $S(L)$ is a doubly stochastic matrix we see that

$$\log perm_S(L) = -\sum_{i=1}^{n} \log(x_i) - \sum_{j=1}^{n} \log(y_i).$$

By combining the last three displays we obtain

$$e^{-n} \leq perm(L)/perm_S(L) \leq 1,$$

from which the result follows. ■

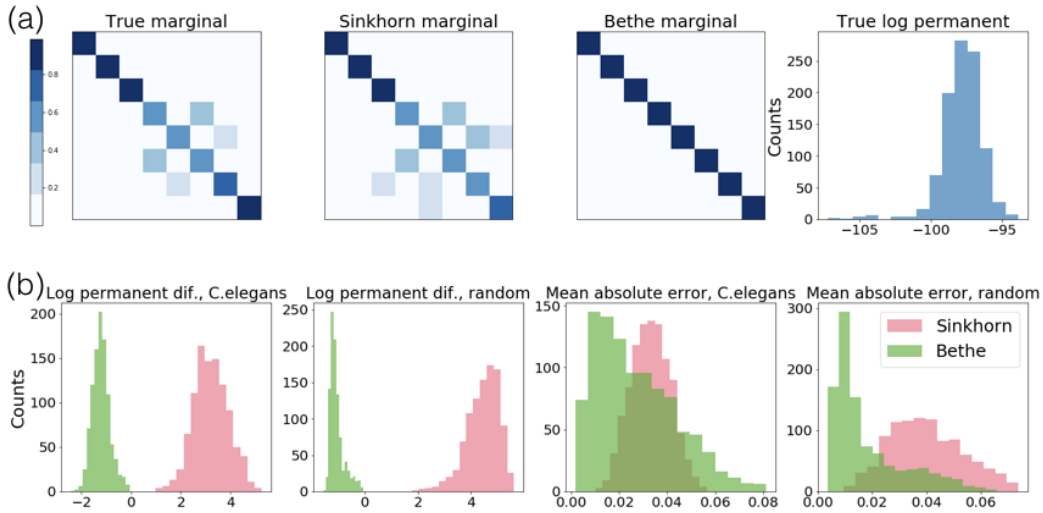

Figure A.1: Same as Fig 1, but now true marginals are plotted in (a), and mean absolute errors in (b).

Appendix B. Experimental details

We used the dataset described in Yemini et al. (2019). This consists on ten NeuroPAL worm heads with available human labels, and with number of neurons n ranging from 180 to 195. Each of these worms is summarized through a $n \times n$ log-likelihood matrix L computed with the methods described in (Yemini et al., 2019, Supplemental Information).

For both the Sinkhorn and Bethe approximation we used 200 iterations. These values led to the computation times described in Fig 1, and preliminary results showed were sufficient to ensure convergence (that is, none of the results would change dramatically for a larger number of iterations). For the MCMC sampler we used the method described in Diaconis (2009). We used 100 chains of length 1000, and for each of them considered we took as samples of the multiples of 10 starting from iteration 500 on.

All results were obtained on a desktop computer with an Intel Xeon W-2125 processor.

Appendix C. Code

Here we provide Python implementations of Sinkhorn and Bethe marginal approximations. These are defined for an arbitrary number of iterations, which in practice may be determined by a convergence criteria.

Sinkhorn approximation The following is a log-space implementation of Sinkhorn approximation as described in Mena et al. (2018).

```
def sinkhorn_logspace(logP, niters):
    for _ in range(niters):
        logP = logP - logsumexp(logP, axis=0, keepdims=True)
        logP = logP - logsumexp(logP, axis=1, keepdims=True)
```

```
    return np.exp(logP)
```

Bethe approximation The following is an efficient log-space implementation of the message passing algorithm described in (Vontobel, 2013, Lemma 29), which was subsequently simplified by Pontobel (2019). The parameter *eps* is introduced for numerical estability.

```
def belief_propagation_log2(M, nIters= 1, eps=1e-20):
    N = M.shape[0]
    logV1 = np.log((1 / N) * np.ones((N, N)))
    logV2 = M - logsumexp(M, axis=1, keepdims=1)
    for _ in range(niters):
        logexpV2 = np.log(-np.expm1(logV2)+eps)
        HelpMat = logV2 + logexpV2
        HelpMat = HelpMat - np.log(-np.expm1(logV2)+eps)
        logV1   = HelpMat - logsumexp(HelpMat,0, keepdims=True)

        HelpMat = logV1 + logexpV2
        HelpMat = HelpMat - np.log(-np.expm1(logV1)+eps)
        logV2   = HelpMat  - logsumexp(HelpMat,1, keepdims=True)
    return np.exp(logV1)
```

in logarithmic space. For the Bethe algorithm we present a more efficient formulation than the one originally presented in To our understading, this is the most efficient implementation.

Appendix D. Supplemental Figure

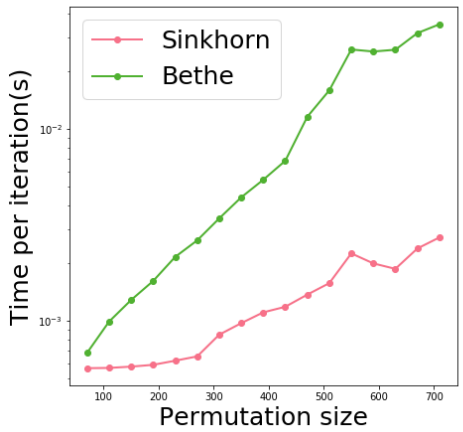

Figure A.2: Computation time per iteration for Sinkhorn and Bethe approximations, as a function of size of the matrix. For each value of $n = 70, 110, 150, \ldots, 710$, a number of 1,000 submatrices of size n were randomly drawn from the ten available log likelihood *C.elegans* matrices (see text on Appendix B, indexes were drawn with replacement). Error bars are omitted because they were too small to be noticed.