Modeling Large-scale Joint Distributions and Inference by Randomized Assignment

Bastian Boll[1 [0000−0002−3490−3350]], Jonathan Schwarz[1 [0000−0003−1825−3826]], Daniel Gonzalez-Alvarado[1 [0000−0002−4636−3697]], Dmitrij Sitenko[1 [0000−0002−0022−3891]], Stefania Petra[2 [0000−0002−7189−2275]], and Christoph Schnörr[1 [0000−0002−8999−2338]]

[1] Image and Pattern Analysis Group, Heidelberg University, Germany
[2] Mathematical Imaging Group, Heidelberg University, Germany
bastian.boll@iwr.uni-heidelberg.de

Abstract. We propose a novel way of approximating energy-based models by randomizing the parameters of assignment flows, a class of smooth dynamical data labeling systems. Our approach builds on averaging flow limit points within the combinatorially large simplex of joint distributions. In an initial learning stage, the distribution of flow parameters is selected to match a given energy-based model. This entails the difficult problem of estimating model entropy which we address by differentiable approximation of a bias-corrected estimator. The model subsequently allows to perform probabilistic inference by computationally efficient draws of structured integer samples which are approximately governed by the energy-based target Gibbs measure in the low-temperature regime. We conduct a rigorous quantitative assessment by approximating a small two-dimensional Ising model and find close approximation of the combinatorial solution in terms of relative entropy which outperforms a mean-field approximation baseline.

Keywords: Probabilistic Inference · Assignment Flows · Energy-based Models · Structured Prediction

1 Introduction

Probabilistic models for context-sensitive decision making and structured prediction have been a focal point of research during the last decades, devoted to data modeling and analysis, imaging science and machine learning. Major paradigms for representing mathematically complex probability distributions include Gibbs distributions, probabilistic graphical models [32,17] and measure transport using push-forward maps parameterized by neural networks [26,16]. A range of variational approximations [3] have been developed for the – typically intractable – problems of inference and parameter learning.

The most basic one, the so-called ('naive') mean-field approximation [32, Section 5], minimizes the Kullback-Leibler distance of a fully factorized distribution and the intractable target distribution. More advanced structured mean-field

approaches include the well-known Bethe- and Kikuchi approximations and related algorithms for approximating marginals of the target distribution by belief propagation [20], [32, Section 4], convexified Bethe approximations [31,11] and related methods in statistical physics, like the cavity method [25].

This paper presents a preliminary step for approaching the problem from quite a different angle. Specifically, we consider a *randomized* dynamical system, the assignment flow approach, proposed by [1] for data and image labeling. Using learned *deterministic* parameters, this approach provides a continuous-time model for deep networks whose layers emerge by geometric integration. Key differences to established image labeling methods based on minimizing energies over discrete variables, like Maximum A-Posterior (MAP) inference using Markov Random Fields [13], include (i) inherent smoothness, (ii) efficient inference by geometric integration and (iii) amenability to learning parameters directly from data. Our goal is to achieve *probabilistic* inference, beyond *deterministic* MAP *point* estimates, by efficient evaluations of *randomized* assignment flows.

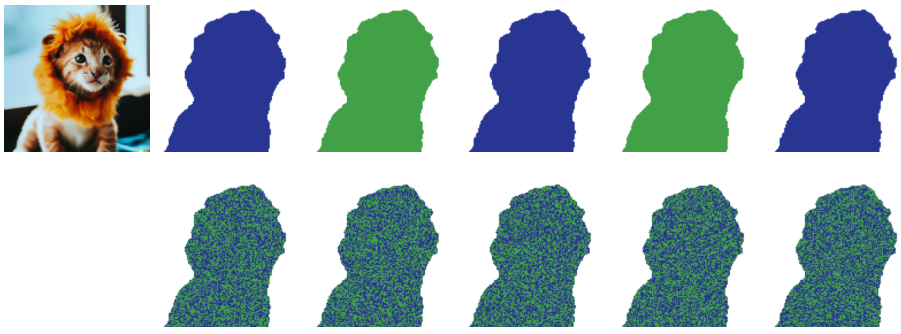

Fig. 1. Image segmentation of an ambiguous subject[4](cat or lion with equal probability). Our model of the joint distribution (samples in first row) captures the structure of the data by coupling subject pixels. In contrast, the marginal distribution (samples in second row) makes a pixelwise independence assumption and therefore fails to represent spatial context.

The rationale behind our approach appears natural when considering the embedding of assignment flows into a meta-simplex, as recently proposed by [4]. *Any* probability distribution over discrete variables can be represented as a point in a combinatorially large probability simplex, each vertex of which represents a *single* labeling corresponding to a single *joint* configuration of *all* involved discrete assignment variables. The embedding of assignment flows then ranges

[4] This image was created by DALL-E 2 [24].

over a submanifold in this simplex corresponding to *factorized* distributions, akin to the basic mean-field approach mentioned above.

In this situation, we utilize (i) that assignment flows converge to labelings under mild conditions, i.e. they approach a vertex in the meta-simplex and, consequently, (ii) that *randomized* assignment flows define *implicitly* via (i) a *probability distribution* on the set of all vertices of the meta-simplex. This distribution, in turn, defines a *barycenter* in the meta-simplex which generally lies *outside* the submanifold corresponding to embedded assignment flows. In other words, using *randomized* assignment flows, we achieve approximate probabilistic representations and inference that are *more* expressive than *any* (naive) mean-field model. Figure 1 provides an example which illustrates this difference.

We point out that our approach to probabilistic modeling and inference, by convex combination of extreme points of compact convex sets of probability distributions, is not at all new in mathematics, but in fact extends far beyond the scenarios with discrete random variables considered here [8]. Our approach to constructing these representations, using randomized assignment flows, is novel however. Randomized assignment flows were also used in [5], yet within the different context of PAC-Bayes risk certification.

Organization. Section 2 briefly presents a specific parameterization of assignment flows introduced in [27] and the embedding approach of [4]. Randomized assignment flows are introduced in Section 3. The approximation of energy-based probability target distributions by our approach using the Gibbs variational principle is detailed in Section 4. Experiments which validate quantitatively our claims are presented in Section 5, using problems which are small enough such that the results of *exact* probabilistic inference can be computed as unequivocal baseline.

Basic notation. We set $[n] = \{1, 2, \ldots, n\}$ for $n \in \mathbb{N}$ and $\mathbb{1}_n = (1, 1, \ldots, 1)^\top \in \mathbb{R}^n$. $\langle \cdot, \cdot \rangle$ denotes the Euclidean vector inner product or the Frobenius matrix inner product. The canonical basis of $\mathbb{R}^n$ is denoted by $(e_1, \ldots, e_n) = I_n$. $\mathbb{R}^n_{++}$ denotes the set of vectors in $\mathbb{R}^n$ with strictly positive entries. $\otimes$ denotes the Kronecker product.

2 (S-)Assignment Flows

2.1 Definition

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \omega)$ denote an undirected weighted graph with $n = |\mathcal{V}|$ vertices and a nonnegative weight function $\omega \colon \mathcal{E} \to [0, \infty)$ on graph edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. Let $S \colon \mathcal{V} \to \mathcal{S}_c$ denote a function that takes values $S_i \in \mathcal{S}_c$, $i \in \mathcal{V}$ in the relative interior $\mathcal{S}_c = \{s \in \mathbb{R}^c_{++} \colon \langle \mathbb{1}_c, s \rangle = 1\}$ of the probability simplex. $(\mathcal{S}_c, g)$ is a Riemannian manifold with the trivial tangent bundle $T_p\mathcal{S}_c \cong \mathcal{S}_c \times T_0\mathcal{S}_c$, with tangent space $T_0\mathcal{S}_c = \{v \in \mathbb{R}^c \colon \langle \mathbb{1}_c, v \rangle = 0\}$ and with the Fisher-Rao metric

$$g \colon T_p\mathcal{S}_c \times T_p\mathcal{S}_c \to \mathbb{R}, \qquad (v_1, v_2) \mapsto \langle v_1, v_2 \rangle_g = \left\langle \frac{v_1}{p}, v_2 \right\rangle . \tag{1}$$

This metric can be seen as an infinitessimal version of relative entropy on $\mathcal{S}_c$. The assignment manifold $(\mathcal{W}, g)$ is the product manifold $\mathcal{W} = \mathcal{S}_c \times \cdots \times \mathcal{S}_c$ with $n = |\mathcal{V}|$ factors and the natural corresponding extension of the Fisher-Rao metric g. It defines the set of feasible assignment matrices $S \in \mathcal{W} \subset \mathbb{R}^{n \times c}_{++}$.

We consider a version of the assignment flow approach [1,28] introduced in [27],

$$\dot{S} = R_S(\Omega S),\ S(0) = S_0, \qquad \Omega_{ij} = \omega_{ij}, \quad ij \in \mathcal{E}, \tag{2a}$$

$$\big(R_S(\Omega S)\big)_i = R_{S_i}(\Omega S)_i, \qquad R_{S_i} = \mathrm{Diag}(S_i) - S_i S_i^\top,\ i \in \mathcal{V}. \tag{2b}$$

Determining $S(t)$ by geometric numerical integration [34] converges for $t \to \infty$ under mild conditions towards unit vectors $S_i^* = e_{j(i)},\ i \in \mathcal{V}$ that assign the label j to the data point at vertex i represented by the initial point $S_{0;i}$ [35]. Using the row-stacking operator $s(t) = \mathrm{vec}_r(S(t))$ and extending Ω from $\mathrm{vec}_r(\Omega S) = (\Omega \otimes I_c)\,\mathrm{vec}_r(S)$ to $\Omega^{\mathfrak{v}}\,\mathrm{vec}_r(S)$, as done in [5] in order to take both spatial and label interaction into account, vectorization of (2) yields

$$\dot{s}(t) = R^{\mathfrak{v}}_{s(t)} \Omega^{\mathfrak{v}} s(t), \qquad s(0) = s_0 \tag{3}$$

with $R^{\mathfrak{v}}_s \Omega^{\mathfrak{v}} s = \mathrm{Diag}(R_{S_1}, \ldots, R_{S_n}) \Omega^{\mathfrak{v}}\,\mathrm{vec}_r(S)$. We further define the *lifting map* $\exp_S(V) = \mathrm{softmax}(V + \log S)$ where softmax is applied along the second dimension (c) and log applies componentwise. To simplify notation, we will in the following re-use the symbol Ω to mean the extended operator $\Omega^{\mathfrak{v}} \in \mathbb{R}^{nc \times nc}$.

2.2 Embedding

In [4], a formal reduction of assignment flows to replicator dynamics has been proposed. To this end, the authors regard the joint state of n nodes each carrying a distribution over c classes as point on a single meta-simplex $\mathcal{S}_N$ with $N = c^n$ vertices (extreme points). As in this work, we will use multi-index notation $\alpha \in [c]^n$ instead of single indices $i \in [N]$ to refer to entries of $\mathcal{S}_N$ and its tangent space. We will use the embedding result [4, Theorem 1] as well as the following associated definitions.

Definition 1 (Embedding Maps). *The maps*

$$T \colon \mathcal{W} \to \mathcal{S}_N, \qquad T(S)_\alpha := \prod_{i \in [n]} S_{i,\alpha_i}, \quad N := c^n, \tag{4a}$$

$$Q \colon T_0\mathcal{W} \to T_0\mathcal{S}_N, \qquad Q(V)_\alpha := \sum_{l \in [n]} V_{l,\alpha_l}\ . \tag{4b}$$

are diffeomorphisms between their domain and a subset of their range. In the case of (4a)*, the range* $\mathrm{img}\, T$ *is the set of joint distributions in* $\mathcal{S}_N$ *which factorize into marginals* S_i*.*

With abuse of notation, we will use the same symbol Q to denote the linear map $\mathbb{R}^{n \times c} \to \mathbb{R}^N$ defined analogously by (4b). From this perspective, the adjoint linear

map $Q^\top$ was shown in [4, Lemma 2] to perform marginalization of distributions in $\mathcal{S}_N$ which is the inverse map of T on its range. In addition, we will analogously apply the maps T and Q to vectorized arguments $s = \mathrm{vec}(S)$ and $v = \mathrm{vec}(V)$. With these definitions, the central result of [4] is that for $s(t)$ with dynamics (3), the quantity $p(t) = T(s(t)) \in \mathcal{S}_N$ has the dynamics

$$\dot{p}(t) = R_{p(t)}[Q\Omega Q^\top p(t)], \qquad p(0) = T(s_0)\ . \tag{5}$$

3 Randomized Assignment Flow

We regard the interaction Ω as a random variable with distribution μ. This in turn makes $p(t) = p(t, \Omega, S_0)$ defined by the dynamics (5) a random variable whose distribution $\nu(t)$ in $\mathcal{S}_N$ varies over time. We will use the first moment

$$P(t) = \mathbb{E}_{p\sim\nu(t)}[p] = \mathbb{E}_{\Omega\sim\mu}[p(t, \Omega, S_0)] \tag{6}$$

to represent a joint distribution of random variables on the graph $\mathcal{G}$. In this section, we examine properties of $P(t)$ and its limit P^∞ over time.

First note that $P(t)$ lies in $\mathcal{S}_N$ because every $p(t, \Omega, S_0)$ lies in $\mathcal{S}_N$ and $\mathcal{S}_N$ is a convex set. Thus, the limit P^∞ lies in the closure $\overline{\mathcal{S}}_N$ which contains all joint distributions of random variables on $\mathcal{G}$ with possibly non-full support. Suppose that (5) converges to a unit vector $e_{\gamma(\Omega)} \in \overline{\mathcal{S}}_N$ for (almost) every Ω drawn from μ. Then

$$P^\infty_\alpha = \lim_{t\to\infty} P(t)_\alpha \to \mathbb{E}_{\Omega\sim\mu}[e_{\gamma(\Omega)}]_\alpha = \mathbb{E}_{\Omega\sim\mu}[\![\alpha = \gamma(\Omega)]\!] = \mathbb{P}_{\Omega\sim\mu}(\alpha = \gamma(\Omega)). \tag{7}$$

Thus, we can draw samples of P^∞ *efficiently* by numerical integration of (3) and these samples *will be integer distributions*, i.e. hard node-label assignments. A distribution μ governing Ω, which meets these requirements is specified next.

Theorem 1 ([35, Thm. 2]). *Let $\Omega = \max(Z + Z^\top, 0) + \epsilon\mathbb{1}_n,\ Z \in \mathbb{R}^{n\times n},\ \epsilon > 0$, where the entries of Z follow a multivariate normal distribution and maximization is componentwise. Then the embedded S-flow* (5) *converges to a unit vector for every draw of Ω and almost every $S_0 \in \mathcal{W}$.*

Proof. For the given shape of Ω, the assumptions of [35, Thm. 2] are met, which guarantees that the solution $S(t)$ of (2) converges to an integral solution for almost every initialization $S_0 \in \mathcal{W}$. Because T bijectively maps the corners of $\mathcal{W}$ to the corners of $\mathcal{S}_N$, the assertion is immediate from the embedding theorem [4, Theorem 1].

Crucially, even though every $p(t, \Omega, S_0)$ lies in the image of T and thus factorizes into node marginals, the expected value $P(t)$ typically *does not factorize*. This is due to the fact that $\mathrm{img}\, T$, the set of rank-1 tensors, is a relatively low-dimensional, curved, non-convex subset of $\overline{\mathcal{S}}_N$. To see this, consider the following Lemma.

Lemma 1 (Lifting Map Lemma [4]). *For any $S \in \mathcal{W}$ and $V \in \mathbb{R}^{n\times c}$ it holds*

$$T(\exp_S(V)) = \exp_{T(S)}(QV) \tag{8}$$

By using Lemma 1, we can show the following properties of $\operatorname{img} T$.

Lemma 2 (Properties of $\operatorname{img} T$). *$\operatorname{img} T$ is a curved, non-convex submanifold of $\mathcal{S}_N$ with dimension at most $n(c-1)$.*

Proof. By choosing $S \in \mathcal{W}$ as the uniform distribution on every node, the statement of Lemma 1 becomes

$$T(\operatorname{softmax}(V)) = \operatorname{softmax}(QV) \tag{9}$$

Since $\operatorname{softmax}\colon T_0\mathcal{W} \to \mathcal{W}$ is surjective onto $\mathcal{W}$, (9) characterizes $\operatorname{img} T$ as the image of the linear subspace $\operatorname{img} Q \subseteq T_0\mathcal{S}_N$ under $\operatorname{softmax}\colon T_0\mathcal{S}_N \to \mathcal{S}_N$. Thus, $\operatorname{img} T$ is flat in e-coordinates on $\mathcal{S}_N$, making it curved in m-coordinates. The subspace $\operatorname{img} Q$ has dimension at most $n(c-1)$ because Q is linear and $T_0\mathcal{W}$ has dimension $n(c-1)$. To see that $\operatorname{img} T$ is not convex, note that the extreme points of $\overline{\mathcal{W}}$ are bijectively mapped to the extreme points of $\overline{\mathcal{S}}_N$ by T. Suppose $\operatorname{img} T$ was convex. Then $\operatorname{img} T$ contains the convex hull of every subset of $\operatorname{img} T$. But the convex hull of the extreme points of $\mathcal{S}_N$ is just all of $\mathcal{S}_N$, contradicting the fact that $\operatorname{img} T$ has lower dimension than $\mathcal{S}_N$. □

4 Approximation of Energy-based Models

Here we consider the approximation of energy-based models, i.e. models in which the probability of configuration α is given by

$$P^*_\alpha = \frac{1}{Z^*}\exp(-E_\alpha), \qquad Z^* = \sum_{\alpha\in[c]^n} \exp(-E_\alpha) \tag{10}$$

where energy E_α of each individual configuration is tractable but the *partition function* Z^* is intractable, because it contains a combinatorially large number of summands. We enumerate the energies of all configurations and collect them in the single vector $E \in \mathbb{R}^N$. As an instance of (10), consider the class of pairwise graphical models with energy

$$E_\alpha = \sum_{i\in[n]} \langle \theta^i, e_{\alpha_i}\rangle + \sum_{ij\in\mathcal{E}} \langle e_{\alpha_j}, \theta^{ij} e_{\alpha_i}\rangle \qquad \theta^i \in \mathbb{R}^c, \quad \theta^{ij} \in \mathbb{R}^{c\times c} \tag{11}$$

Tying back to the notation of earlier sections, we transform the pairwise term in (11) to

$$\sum_{ij\in\mathcal{E}} \langle e_{\alpha_j}, \theta^{ij} e_{\alpha_i}\rangle = \sum_{ij\in\mathcal{E}} \langle (Q^\top e_\alpha)_j, \theta^{ij}(Q^\top e_\alpha)_i\rangle = \langle Q^\top e_\alpha, \theta^{(p)} Q^\top e_\alpha\rangle \tag{12}$$

with matrix $\theta^{(p)} \in \mathbb{R}^{nc \times nc}$ built from blocks $\theta^{ij} \in \mathbb{R}^{c \times c}$, $i, j \in [n]$. Similarly, combining unary parameters θ^i into a single vector $\theta^{(u)} \in \mathbb{R}^{nc}$ yields the vectorized form of (11)

$$E = Q\theta^{(u)} + \operatorname{diag}(Q\theta^{(p)}Q^\top) \ . \tag{13}$$

Suppose one approximates P^* by a tractable $P \in \mathcal{S}_N$. This entails minimization of

$$\begin{aligned} \mathrm{KL}[P \colon P^*] &= \left\langle P, \log \frac{P}{P^*} \right\rangle = \langle P, \log P \rangle - \langle P, \log P^* \rangle && \text{(14a)} \\ &= -H(P) - \langle P, E \rangle + \underbrace{\log Z^*}_{\text{const}} \underbrace{\langle P, \mathbb{1}_N \rangle}_{=1} && \text{(14b)} \end{aligned}$$

which mirrors the well-known conjugacy relation [7, Lemma 1.1.3]

$$\log \left\langle \frac{1}{N}\mathbb{1}_N, \exp(-E) \right\rangle = \sup_P \ -\langle E, P \rangle - \mathrm{KL}[P \colon \frac{1}{N}\mathbb{1}_N] \ . \tag{15}$$

Thus, in order to learn P efficiently, we need to be able to compute its entropy and expected energy as well as their respective gradients. Since the energy of each individual configuration is tractable, the expected energy of a tractable model is typically easy to estimate. However, estimating entropy from samples is generally a difficult problem, which makes tractable entropy a key design criterium for surrogate models P. Along this line of reasoning, the basic *mean-field approach* is to approximate P^* by a factorizing distribution $T(M)$. The model entropy in (14) then simplifies to

$$-H(T(M)) = \langle T(M), \log T(M) \rangle = \langle T(M), Q \log M \rangle = \langle M, \log M \rangle \tag{16}$$

by [4, Lemma 2]. Because both the uniform distribution in $\mathcal{S}_N$ and every extreme point of $\mathcal{S}_N$ is a factorizing distribution, the mean-field approximation generally works best if either (a) all configurations have close to the same probability (high temperature regime) or (b) P^* is close to an integer distribution (the system is essentially deterministic). It is the challenging medium or low temperature regime in which a more sophisticated model is typically required – entailing the problem of entropy estimation.

Here we propose to approximate P^* by P as defined in (6). This goes beyond mean field approaches because, as discussed in Section 3, P typically lies outside $\operatorname{img} T$ i.e. does not factorize. Expected model energy reads

$$\langle P, E \rangle = \langle \mathbb{E}_\Omega T(S(\Omega)), E \rangle = \mathbb{E}_\Omega \langle T(S(\Omega)), E \rangle \tag{17}$$

which amounts to an expected value of mean field energies. Thus, if mean field energy is tractable, the empirical energy over samples of Ω is an unbiased estimator of model energy.

We turn to the more challenging problem of entropy estimation. Typically, estimating model entropy $H(P)$ from samples is difficult because the support $|\operatorname{supp} P| = s$ of P is large compared to the number m of available samples.

The support of P^* can be arbitrarily large in principle. In fact, as a prerequisite for the Hammersley-Clifford theorem [6, Thm. 9.1.10], full support has formal merit in Markov random fields. On the other hand, many situations of practical interest do not benefit from a model with very large support. For instance, in image segmentation, most configurations of classes on the pixels of an image will have very little semantic content. In statistical mechanics, full support is beneficial to model high temperature systems. However, the behavior of these systems is dominated by randomness and they are well-described by a mean field approximation. More sophisticated models are beneficial in the challenging medium or low temperature regime and in this case small support can suffice.

Suppose the support size s is small compared to the number m of available samples $\{\alpha(i)\}_{i\in[m]} \subseteq [c]^n$ drawn from P^∞. Denote by

$$\widehat{p} = \frac{1}{m} \sum_{i\in[m]} e_{\alpha(s)} \in \mathcal{S}_N \tag{18}$$

the empirical distribution of samples. A classical analysis by [18] shows that the *plugin estimator*

$$H(P) \approx H(\widehat{p}) = -\sum_{\alpha\in\operatorname{supp}(\widehat{p})} \widehat{p}_\alpha \log \widehat{p}_\alpha \tag{19}$$

has bias

$$\mathbb{E}[H(\widehat{p})] - H(P) = -\frac{s-1}{2m} + \mathcal{O}\Big(\frac{1}{m^2}\Big) \tag{20}$$

which leads to the Miller-Maddows bias correction for known support s. It was shown that this only achieves a consistent estimator if $m \gg s$ [21] which is far from the optimal rate of $m \gg s/\log s$ [12] and thus motivates the use of more advanced approaches such as [29,12,33,30].

The support of P as defined in (6) is typically not known. However, in our experiments (Section 5) we observe that the empirical distribution is only supported on relatively few configurations. For this reason, we judge (19) with bias correction (20) to be sufficient for the case at hand. Note that an unbiased estimator of entropy from samples exists [19] but is not practical for our use case, because it entails drawing a potentially large number of samples which is not known a priori.

As a key issue it remains to find a differentiable approximation of $-H(\widehat{p})$. Under the assumption of integer samples, we find

$$-H(\widehat{p}) = \Big\langle \frac{1}{m} \sum_{k\in[m]} T(S^k), \log \frac{1}{m} \sum_{k\in[m]} T(S^k) \Big\rangle \tag{21a}$$

$$= \Big\langle \frac{1}{m} \sum_{k\in[m]} e_{\alpha(k)}, \log \frac{1}{m} \sum_{k\in[m]} e_{\alpha(k)} \Big\rangle = \frac{1}{m} \sum_{k\in[m]} \log \Big(\frac{1}{m} \sum_{l\in[m]} e_{\alpha(l)}\Big)_{\alpha(k)} \tag{21b}$$

$$= -\log m + \frac{1}{m} \sum_{k\in[m]} \log \Big(\sum_{l\in[m]} e_{\alpha(l)}\Big)_{\alpha(k)} . \tag{21c}$$

This motivates the approximation

$$-H(\widehat{p}) \approx -\log m + \frac{1}{m} \sum_{k\in[m]} \log\Big(\sum_{l\in[m]} T(S^l)\Big)_{\alpha(k)} \tag{22}$$

for non-integer samples S^l. The fact that assignment flows converge to integer labelings is crucial to this construction, because quantities in $\mathcal{S}_N$ can not be explicitly represented in numerical computations and integrality of S allows to reduce the sparse sum in (21b) to a tractable quantity. Note that $T(S^l)_{\alpha(k)}$ above is a product of n numbers in $(0,1)$. We thus rewrite the summands in (22) as

$$\log\Big(\sum_{l\in[m]} T(S^l)\Big)_{\alpha(k)} \overset{(4a)}{=} \log\Big(\sum_{l\in[m]} \prod_{i\in[n]} S^l_{i,\alpha(k)_i}\Big)_{\alpha(k)} \tag{23a}$$

$$= \log \sum_{l\in[m]} \exp\Big(\sum_{i\in[n]} \log S^l_{i,\alpha(k)_i}\Big) \tag{23b}$$

to avoid numerical underflow problems by leveraging a stabilized implementation of $\mathrm{logexp}_{\epsilon=1}$. Note that the right-hand side is differentiable.

Once a suitable approximation of P^* is found by minimizing (14) with respect to parameters governing μ (Section 5), the model P can be used for probabilistic inference. Marginal distributions are easily estimated via

$$Q^\top P = \mathbb{E}_{\Omega\sim\mu}[Q^\top T(S(\Omega))] = \mathbb{E}_{\Omega\sim\mu}[S(\Omega)] \tag{24}$$

and more generally, any quantity $Q\phi$ with $\phi \in \mathbb{R}^{nc}$ can be inferred by

$$\mathbb{E}_P[Q\phi] = \langle \mathbb{E}_{\Omega\sim\mu}[T(S(\Omega))], Q\phi\rangle = \mathbb{E}_{\Omega\sim\mu}[\langle S(\Omega), \phi\rangle] \ . \tag{25}$$

5 Experiments

The introductory example in Fig. 1 was produced by approximating a large Potts model on the grid graph of image pixels[5]. This was achieved by randomizing the generalized S-flow (3), giving $\Omega \in \mathbb{R}^{nc\times nc}$ the structure of multi-channel convolution with weights following a multivariate normal distribution. Suitable moments for this normal distribution together with a suitable flow initialization s_0 are the result of a *training procedure* which minimizes (14). To this end, a reparameterization trick [15] is applied in conjunction with the approximation (22) and bias correction (20) where the unknown support s is replaced by the empirical support $\widehat{s} = |\operatorname{supp}\widehat{p}|$ smoothed by the mean entropy of nodewise assignment. Numerical integration of (3) via the simple geometric Euler scheme [34] (step size 0.1, end time 1.0) is unwound and automatically differentiated by PyTorch [22] which allows to find a local optimum of parameters by employing the Adam optimizer [14] with step length 0.01.

[5] All experiments were run on a single NVIDIA RTX 2080ti graphics card

In this section, we further demonstrate the approximation of energy-based models on a small *two-dimensional Ising model*, i.e. a system of binary random variables with nearest-neighbor interaction on a grid graph, governed by a Gibbs distribution of the form (10) with a corresponding energy function E_α.

These systems are classical ones in statistical mechanics [23]. They prototypically represent a combinatorially large configuration space and *long range* correlation at low temperatures. As a consequence, in the presence of an 'external field' [2], i.e. data defining unary potentials in E_α, minimizing E_α and probabilistic inference become NP-hard even for moderate problem sizes. Such models initiated research on image segmentation and Bayesian inference [9,10] and have been stimulating research on variational approximations for many years [32,17]. As a consequence, they define an ideal testbed for evaluating our approach and validating also experimentally our claims.

$\mathcal{G}$ is chosen as a 3×8 grid graph such that the combinatorial partition function and true marginals can still be computed by brute force. This allows to give numerical values for the distance to the combinatorial model in terms of relative entropy via (14). The number of classes is $c = 2$. Unary energy is chosen as -3.0 for the 0-configuration of nodes on the left boundary and as 3.0 on the right boundary. All other unary energies are zero. Pairwise energy is set to $\theta^{ij} = \frac{7}{10} \cdot (\mathbb{1}_c \mathbb{1}_c^\top - \mathbb{I}_c)$ for each edge.

We approximate this model by the same training procedure as above with reduced learning rate $5 \cdot 10^{-3}$ over 5k iterations. This takes around 21 minutes on a single desktop graphics card. To guarantee S-flow convergence via Theorem 1, we omit label interaction as afforded by the generalization (3) and instead use (2) with symmetric matrix $\Omega \in \mathbb{R}^{n\times n}$ parameterized as $\Omega = \max(Z + Z^\top, 0) + 10^{-3}\mathbb{I}_n$ with entries of $Z \in \mathbb{R}^{n\times n}$ following a multivariate normal distribution. We initialize the distribution of Z centered at $\frac{1}{20}\mathbb{1}_n\mathbb{1}_n^\top$ and with componentwise variance 10^{-1}. In the early stages of optimization, samples are not integral due to the finite time horizon, but we observe that the sample entropy gradually decreases over the course of optimization, making the approximation (22) already close to exact for finite time. Once a model is learnt through convergence to a local minimum, integrality of samples is guaranteed by Theorem 1 for $t \to \infty$. In fact, it was shown in [35] that the same integer limit is also found by rounding after sufficiently large but finite time t which is relevant for numerical implementation.

As a baseline, we compute a mean field approximation $M \in \mathcal{W}$ by parameterizing $M = \mathrm{softmax}(V)$ and using the Adam optimizer to learn V by minimizing the tractable form of (14). This procedure is repeated for 1k initializations drawn randomly from a standard normal distribution of $V \in \mathbb{R}^{n\times c}$ and a model with minimal KL distance is selected from resulting local optima. The true distribution has multiple modes, of which mean field approximation can only represent just one. In contrast, our model is able to capture the multimodality as is apparent from samples (Fig. 2), close approximation of marginals (Fig. 3) and low relative entropy (Tab. 1).

Fig. 2. Samples of Ising model from the mean-field baseline (first row) and from our model (second row). This demonstrates that, unlike the mean-field approximation, our approach can explore multiple modes in the low-temperature regime.

Fig. 3. Marginals of the true distribution (left), our approximation via randomized assignment (middle) and the baseline mean-field approximation (right).

6 Discussion and Conclusion

In the low temperature regime (E large), the mass of P^* concentrates around its modes. For this reason, the proposed model – for which small support is computationally beneficial – actually becomes more effective at lower temperature. This unusual performance characteristic makes our approach promising in challenging structured prediction scenarios where mean-field approximation fails to capture the structure of interest. A natural direction of future work is to construct differentiable approximations such as (22) for more advanced entropy estimators.

Acknowledgements This work is funded by the Deutsche Forschungsgemeinschaft (DFG), grant SCHN 457/17-1, within the priority programme SPP 2298: "Theoretical Foundations of Deep Learning". This work is funded by the Deutsche Forschungsgemeinschaft (DFG) under Germany's Excellence Strategy EXC-2181/1 - 390900948 (the Heidelberg STRUCTURES Excellence Cluster).

Table 1. Summary of Ising model approximation. Relative entropy to the true distribution is computed by brute-force evaluation of the combinatorial partition function. Entropy of our model is closely approximated by (19) with bias correction (20) using $m = 1\text{M}$ integer samples.

Model	KL	Energy	Entropy	Marginal Difference
AF (ours)	**0.599**	-1.98	2.56	**0.090**
Mean Field	1.974	-1.57	1.60	0.198

References

1. Åström, F., Petra, S., Schmitzer, B., Schnörr, C.: Image Labeling by Assignment. Journal of Mathematical Imaging and Vision **58**(2), 211–238 (2017)
2. Baxter, R.: Exactly Solved Models in Statistical Mechanics. Academic Press (1982)
3. Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational Inference: A Review for Statisticians. Journal of the American Statistical Association **112**(518), 859–877 (04 2017)
4. Boll, B., Schwarz, J., Schnörr, C.: On the Correspondence between Replicator Dynamics and Assignment Flows. In: SSVM 2021: Scale Space and Variational Methods in Computer Vision. LNCS, vol. 12679, pp. 373–384. Springer (2021)
5. Boll, B., Zeilmann, A., Petra, S., Schnörr, C.: Self-Certifying Classification by Linearized Deep Assignment. preprint arXiv:2201.11162 (2022)
6. Brèmaud, P.: Discrete Probability Models and Methods. Springer (2017)
7. Catoni, O.: PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning. Institute of Mathematical Statistics (2007)
8. Dynkin, E.B.: Sufficient Statistics and Extreme Points. Ann. Probability **6**(5), 705–730 (1978)
9. Geman, S., Geman, D.: Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Trans. Patt. Anal. Mach. Intell. **6**(6), 721–741 (1984)
10. Gidas, B.: A Renormalization Group Approach to Image Processing Problems. IEEE Trans. Patt. Anal. Mach. Intell. **11**(11), 164–180 (1989)
11. Heskes, T.: Convexity Arguments for Efficient Minimization of the Bethe and Kikuchi Free Energies. J. Artif. Intell. Res. **26**, 153–190 (2006)
12. Jiao, J., Venkat, K., Han, Y., Weissman, T.: Minimax Estimation of Functionals of Discrete Distributions. IEEE Transactions on Information Theory **61**(5), 2835–2885 (2015)
13. Kappes, J., Andres, B., Hamprecht, F., Schnörr, C., Nowozin, S., Batra, D., Kim, S., Kausler, B., Kröger, T., Lellmann, J., Komodakis, N., Savchynskyy, B., Rother, C.: A Comparative Study of Modern Inference Techniques for Structured Discrete Energy Minimization Problems. Int. J. Computer Vision **115**(2), 155–184 (2015)
14. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. preprint arXiv:1412.6980 (2014)
15. Kingma, D.P., Welling, M.: Auto-encoding Variational Bayes. preprint arXiv:1312.6114 (2013)
16. Kobyzev, I., Prince, S.D., Brubaker, M.A.: Normalizing Flows: An Introduction and Review of Current Methods. IEEE Trans. Pattern Anal. Mach. Intell. **43**(11), 3964–3979 (2021)
17. Mézard, M., Montanari, A.: Information, Physics, and Computation. Oxford Univ. Press (2009)
18. Miller, G.: Note on the Bias of Information Estimates. Information Theory in Psychology: Problems and Methods (1955)
19. Montgomery-Smith, S., Schürmann, T.: Unbiased estimators for entropy and class number. arXiv preprint arXiv:1410.5002 (2014)
20. Pakzad, P., Anantharam, V.: Estimation and Marginalization using Kikuchi Approximation Methods. Neural Computation **17**(8), 1836–1873 (2005)
21. Paninski, L.: Estimation of Entropy and Mutual Information. Neural computation **15**(6), 1191–1253 (2003)

22. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An Imperative Style, High-performance Deep Learning Library. NIPS (2019)
23. Pathria, R.K., Beale, P.D.: Statistical Mechanics. Academic Press, 3rd edn. (2011)
24. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical Text-Conditional Image Generation with CLIP Latents (2022)
25. Rizzo, T., Wemmenhove, B., Kappen, H.J.: Cavity Approximation for Graphical Models. Phys. Rev. E **76**(1) (2007)
26. Ruthotto, L., Haber, E.: An Introduction to Deep Generative Modeling. GAMM Mitt. **44**(2), 24 pages (2021)
27. Savarino, F., Schnörr, C.: Continuous-Domain Assignment Flows. Europ. J. Appl. Math. **32**(3), 570–597 (2021)
28. Schnörr, C.: Assignment Flows. In: Grohs, P., Holler, M., Weinmann, A. (eds.) Variational Methods for Nonlinear Geometric Data and Applications, pp. 235–260. Springer (2020)
29. Valiant, G., Valiant, P.: Estimating the Unseen: an n/log (n)-Sample Estimator for Entropy and Support Size, shown optimal via new CLTs. In: Proc. 43th ACM Symposium on Theory of Computing. pp. 685–694 (2011)
30. Valiant, G., Valiant, P.: Estimating the Unseen: Improved Estimators for Entropy and other Properties. Journal of the ACM **64**(6), 1–41 (2017)
31. Wainwright, M.J., Jaakola, T.S., Willsky, A.S.: Tree-Based Reparameterization Framework for Analysis of Sum-Product and Related Algorithms. IEEE Trans. Inf. Theory **49**(5), 1120–1146 (2003)
32. Wainwright, M., Jordan, M.: Graphical Models, Exponential Families, and Variational Inference. Found. Trends Mach. Learn. **1**(1-2), 1–305 (2008)
33. Wu, Y., Yang, P.: Minimax rates of entropy estimation on large alphabets via best polynomial approximation. IEEE Transactions on Information Theory **62**(6), 3702–3720 (2016)
34. Zeilmann, A., Savarino, F., Petra, S., Schnörr, C.: Geometric Numerical Integration of the Assignment Flow. Inverse Problems **36**(3), 034004 (33pp) (2020)
35. Zern, A., Zeilmann, A., Schnörr, C.: Assignment Flows for Data Labeling on Graphs: Convergence and Stability. Information Geometry **5**, 355–404 (2022)