Variational Inference for Sparse and Undirected Models

John Ingraham [1] **Debora Marks** [1]

Abstract

Undirected graphical models are applied in genomics, protein structure prediction, and neuroscience to identify sparse interactions that underlie discrete data. Although Bayesian methods for inference would be favorable in these contexts, they are rarely used because they require doubly intractable Monte Carlo sampling. Here, we develop a framework for scalable Bayesian inference of discrete undirected models based on two new methods. The first is Persistent VI, an algorithm for variational inference of discrete undirected models that avoids doubly intractable MCMC and approximations of the partition function. The second is Fadeout, a reparameterization approach for variational inference under sparsity-inducing priors that captures *a posteriori* correlations between parameters and hyperparameters with noncentered parameterizations. We find that, together, these methods for variational inference substantially improve learning of sparse undirected graphical models in simulated and real problems from physics and biology.

1. Introduction

Hierarchical priors that favor sparsity have been a central development in modern statistics and machine learning, and find widespread use for variable selection in biology, engineering, and economics. Among the most widely used and successful approaches for inference of sparse models has been L_1 regularization, which, after introduction in the context of linear models with the LASSO (Tibshirani, 1996), has become the standard tool for both directed and undirected models alike (Murphy, 2012).

Despite its success, however, L_1 is a pragmatic compromise. As the closest convex approximation of the idealized L_0 norm, L_1 regularization cannot model the hypothesis of sparsity as well as some Bayesian alternatives (Tipping, 2001). Two Bayesian approaches stand out as more accurate models of sparsity than L_1. The first, the spike and slab (Mitchell & Beauchamp, 1988), introduces discrete latent variables that directly model the presence or absence of each parameter. This discrete approach is the most direct and accurate representation of a sparsity hypothesis (Mohamed et al., 2012), but the discrete latent space that it imposes is often computationally intractable for models where Bayesian inference is difficult.

The second approach to Bayesian sparsity uses the scale mixtures of normals (Andrews & Mallows, 1974), a family of distributions that arise from integrating a zero mean-Gaussian over an unknown variance as

$$p(\theta) = \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\theta^2}{2\sigma^2}\right\} p(\sigma) d\sigma. \qquad (1)$$

Scale-mixtures of normals can approximate the discrete spike and slab prior by mixing both large and small values of the variance σ^2. The implicit prior of L_1 regularization, the Laplacian, is a member of the scale mixture family that results from an exponentially distributed variance σ^2. Thus, mixing densities $p(\sigma^2)$ with subexponential tails and more mass near the origin more accurately model sparsity than L_1 and are the basis for approaches often referred to as "Sparse Bayesian Learning" (Tipping, 2001). Both the Student-t of Automatic Relevance Determination (ARD) (MacKay et al., 1994) and the Horseshoe prior (Carvalho et al., 2010) incorporate these properties.

Applying these favorable, Bayesian approaches to sparsity has been particularly challenging for discrete, undirected models like Boltzmann Machines. Undirected models possess a representational advantage of capturing 'collective phenomena' with no directions of causality, but their likelihoods require an intractable normalizing constant (Murray & Ghahramani, 2004). For a fully observed Boltzmann Machine with $\mathbf{x} \in \{0,1\}^D$ the distribution[1] is

$$p(\mathbf{x}|\mathbf{J}) = \frac{1}{Z(\mathbf{J})} \exp\left\{\sum_{i<j} J_{ij} x_i x_j\right\}, \qquad (2)$$

[1]Harvard Medical School, Boston, Massachusetts. Correspondence to: John Ingraham <ingraham@fas.harvard.edu>, Debora Marks <debbie@hms.harvard.edu>.

Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

[1]We exclude biases for simplicity.

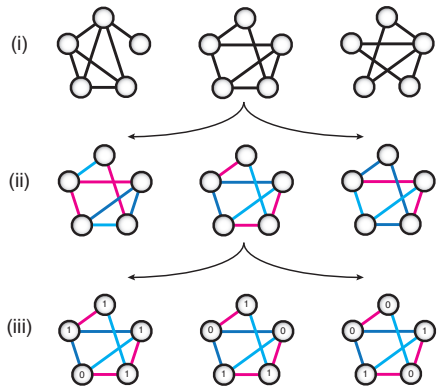

Figure 1. Bayesian inference for discrete undirected graphical models with sparse priors is triply intractable, as the space of possible models spans: (i) all possible sparsity patterns, each of which possesses its own (ii) parameter space, for which every distinct set of parameters has its own (iii) intractable normalizing constant.

where the partition function $Z(\mathbf{J})$ depends on the couplings. Whenever a new set of couplings $\mathbf{J}$ are considered during inference, the partition function $Z(\mathbf{J})$ and corresponding density $p(\mathbf{x}|\mathbf{J})$ must be reevaluated. This requirement for an an intractable calculation embedded within already-intractable nonconjugate inference has led some to term Bayesian learning of undirected graphical models "doubly intractable" (Murray et al., 2006). When all $2^{\binom{D}{2}}$ patterns of discrete spike and slab sparsity are added on top of this, we might call this problem "triply intractable" (Figure 1). Triple-intractability does not mean that this problem is impossible, but it will typically require expensive approaches based on MCMC-within-MCMC (Chen & Welling, 2012).

Here we present an alternative to MCMC-based approaches for learning undirected models with sparse priors based on stochastic variational inference (Hoffman et al., 2013). We combine three ideas: (i) stochastic gradient variational Bayes (Kingma & Welling, 2014; Rezende et al., 2014; Titsias & Lázaro-Gredilla, 2014)[2], (ii) persistent Markov chains (Younes, 1989), and (iii) a noncentered parameterization of scale-mixture priors, to inherit the benefits of hierarchical Bayesian sparsity in an efficient variational framework. We make the following contributions:

- We extend stochastic variational inference to undirected models with intractable normalizing constants by developing a learning algorithm based on *persistent* Markov chains, which we call Persistent Variational Inference (PVI) (Section 2).
- We introduce a reparameterization approach for variational inference under sparsity-inducing scale-mixture priors (e.g. the Laplacian, ARD, and the Horseshoe) that significantly improves approximation quality by capturing *scale uncertainty* (Section 3). When combined with Gaussian stochastic variational inference, we call this Fadeout.
- We demonstrate how a Bayesian approach for learning sparse undirected graphical models with PVI and Fadeout yields significantly improved inferences of both synthetic and real applications in physics and biology (Section 4).

[2]This is also a type of noncentered parameterization, but of the variational distribution rather than the posterior.

2. Persistent Variational Inference

Background: Learning in undirected models Undirected graphical models, also known as Markov Random Fields, can be written in log-linear form as

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left\{\sum_{i=1}^{k} \theta_i f_i(\mathbf{x})\right\}, \tag{3}$$

where i indexes a set of k features $\{f_i(\mathbf{x})\}_{i=1}^k$ and the partition function $Z(\boldsymbol{\theta}) = \sum_{\mathbf{x}} \exp\{\sum_i \theta_i f_i(\mathbf{x})\}$ normalizes the distribution (Koller & Friedman, 2009). Maximum Likelihood inference selects parameters $\boldsymbol{\theta}$ that maximize the probability of data $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$ by ascending the gradient of the (averaged) log likelihood

$$\frac{\partial}{\partial \theta_i} \frac{1}{N} \log p(\mathcal{D}|\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{D}}\left[f_i(\mathbf{x})\right] - \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})}\left[f_i(\mathbf{x})\right]. \tag{4}$$

The first term in the gradient is a data-dependent average of feature $f_i(\mathbf{x})$ over $\mathcal{D}$, while the second term is a data-independent average of feature $f_i(\mathbf{x})$ over the model distribution that often requires sampling (Murphy, 2012)[3].

Bayesian learning for undirected models is confounded by the partition function $Z(\boldsymbol{\theta})$. Given the data $\mathcal{D}$, a prior $p(\boldsymbol{\theta})$, and the log potentials $H[\mathbf{x}|\boldsymbol{\theta}] = -\sum_i \theta_i f_i(\mathbf{x})$, the posterior distribution of the parameters is

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\boldsymbol{\theta}) \prod_i e^{-H[\mathbf{x}^{(i)}|\boldsymbol{\theta}]}/Z(\boldsymbol{\theta})}{\int p(\boldsymbol{\theta}') \prod_i e^{-H[\mathbf{x}^{(i)}|\boldsymbol{\theta}']}/Z(\boldsymbol{\theta}') d\boldsymbol{\theta}'}, \tag{5}$$

which contains an intractable partition function $Z(\boldsymbol{\theta})$ *within* the already-intractable evidence term. As a result, most algorithms for Bayesian learning of undirected models require either doubly-intractable MCMC and/or approximations of the likelihood $p(\mathbf{x}|\boldsymbol{\theta})$.

[3]Depending on the details of the MCMC and the community these approaches are known as Boltzmann Learning, Stochastic Maximum Likelihood, or Persistent Contrastive Divergence (Tieleman, 2008).

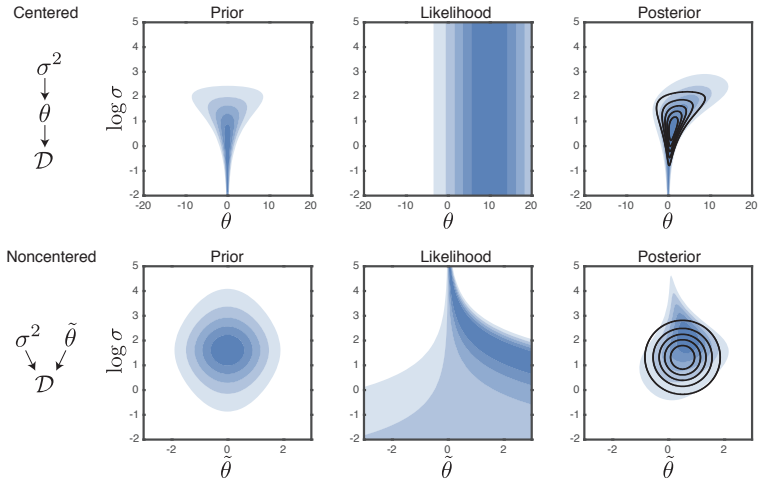

Figure 2. Variational inference for sparse priors with noncentered reparameterizations. Several sparsity-inducing priors such as the Laplacian, Student-t, and Horseshoe (shown here) can be derived as scale-mixture priors in which each model parameter θ is drawn from a zero-mean Gaussian with random variance σ^2 (top row). The dependency of θ on σ^2 gives rise to a strongly curved "funnel" distribution (blue, top left and right) that is poorly modeled by a factorized variational distribution (not shown). A noncentered reparameterization with $\tilde{\theta} = \theta/\sigma$ trades independence of θ and σ^2 in the likelihood (blue, top center) for independence in the prior (blue, bottom left), allowing a factorized variational distribution over noncentered parameters (black contours, bottom right) to implicitly capture the *a priori* correlations between θ and σ^2 (black contours, top right). As a result, the variational distribution can more accurately model the bottom of the "funnel", which corresponds to sparse estimates.

A tractable estimator for ∇ELBO of undirected models Here we consider how to approximate the intractable posterior in (5) without approximating the partition function $Z(\boldsymbol{\theta})$ or the likelihood $p(\mathbf{x}|\boldsymbol{\theta})$ by using variational inference. Variational inference recasts inference with $p(\boldsymbol{\theta}|\mathcal{D})$ as an optimization problem of finding a variational distribution $q(\boldsymbol{\theta}|\boldsymbol{\phi})$ that is closest to $p(\boldsymbol{\theta}|\mathcal{D})$ as measured by KL divergence (Jordan et al., 1999). This can be accomplished by maximizing the Evidence Lower BOund

$$\mathcal{L}(\boldsymbol{\phi}) \triangleq \mathbb{E}_q\left[\log p(\mathcal{D},\boldsymbol{\theta}) - \log q(\boldsymbol{\theta}|\boldsymbol{\phi})\right] \leq \log p(\mathcal{D}). \quad (6)$$

For scalability, we would like to optimize the ELBO with methods that can leverage Monte Carlo estimators of the gradient $\nabla_\phi \mathcal{L}(\boldsymbol{\phi})$. One possible strategy for this would be would be to develop an estimator based on the score function (Ranganath et al., 2014) with a Monte-Carlo approximation of

$$\nabla_\phi \mathcal{L} = \mathbb{E}_q\left[\nabla_\phi \log q(\boldsymbol{\theta}|\boldsymbol{\phi}) \log \frac{p(\mathcal{D},\boldsymbol{\theta})}{q(\boldsymbol{\theta}|\boldsymbol{\phi})}\right]. \quad (7)$$

Naively substituting the likelihood (3) in the score function estimator (7) nests the intractable log partition function $\log Z(\boldsymbol{\theta})$ within the average over $q(\boldsymbol{\theta}|\boldsymbol{\phi})$, making this an untenable (and extremely high variance) approach to inference with undirected models.

We can avoid the need for a score-function estimator with the 'reparameterization trick' (Kingma & Welling, 2014; Rezende et al., 2014; Titsias & Lázaro-Gredilla, 2014) that has been incredibly useful for directed models. Consider a variational approximation $q(\boldsymbol{\theta}|\boldsymbol{\phi}) = \prod_i q(\theta_i|\mu_i, s_i)$ that is a fully factorized (mean field) Gaussian with means $\boldsymbol{\mu}$ and log standard deviations $\mathbf{s}$. The ELBO expectations under $q(\boldsymbol{\theta}|\boldsymbol{\phi})$ can be rewritten as expectations wrt an independent noise source $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ where[4] $\boldsymbol{\theta}(\boldsymbol{\epsilon}) = \boldsymbol{\mu} + \exp\{\mathbf{s}\} \odot \boldsymbol{\epsilon}$. Then the gradients are

$$\nabla_\mu \mathcal{L} = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\nabla_\theta \log p(\mathcal{D},\boldsymbol{\theta}(\boldsymbol{\epsilon}))\right], \quad (8)$$

$$\nabla_{\mathbf{s}} \mathcal{L} = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\nabla_\theta \log p(\mathcal{D},\boldsymbol{\theta}(\boldsymbol{\epsilon})) \odot (\boldsymbol{\theta}(\boldsymbol{\epsilon}) - \boldsymbol{\mu})\right] + \mathbf{1}. \quad (9)$$

Because these expectations require only the gradient of the likelihood $\nabla_\theta \log p(\mathcal{D}|\boldsymbol{\theta})$, the gradient for the undirected model (4) can be substituted to form a nested expectation for $\nabla_\phi \mathcal{L}(\boldsymbol{\phi})$. This can then be used as a Monte Carlo gradient estimator by sampling $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I), \mathbf{x} \sim p(\mathbf{x}|\boldsymbol{\theta}(\boldsymbol{\epsilon}))$.

Persistent gradient estimation In Stochastic Maximum Likelihood estimation for undirected models, the intractable gradients of (4) are estimated by sampling $p(\mathbf{x}|\boldsymbol{\theta})$. Although sampling-based approaches are slow, they can be made considerably more efficient by running a set of Markov chains in parallel with state that *persists* between iterations (Younes, 1989). Persistent state maintains the Markov chains near their equilibrium distributions, which means that they can quickly re-equilibrate after perturbations to the parameters $\boldsymbol{\theta}$ during learning.

[4]The $\odot$ operator is an element-wise product.

We propose variational inference in undirected models based on persistent gradient estimation of $\nabla_\theta \log p(\mathcal{D}|\boldsymbol{\theta})$ and refer to this as Persistent Variational Inference (PVI) (Algorithm in Appendix). Following the notation of PCD-n (Tieleman, 2008), PVI-n refers to using n sweeps of Gibbs sampling with persistent Markov chains between iterations. This approach is generally compatible with any estimators of ∇ELBO that are based on the gradient of the log likelihood, several examples of which are explained in (Kingma & Welling, 2014; Rezende et al., 2014; Titsias & Lázaro-Gredilla, 2014).

Behavior of the solution for Gaussian q When the variational approximation is a fully factorized Gaussian $q(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\sigma})$ and the prior is flat $p(\boldsymbol{\theta}) \propto 1$, the solution to $\boldsymbol{\mu}^\star, \boldsymbol{\sigma}^\star = \arg\max_{\boldsymbol{\mu},\boldsymbol{\sigma}} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ will satisfy

$$\mathbb{E}_{\mathcal{D}}\left[f_i(\mathbf{x})\right] = \mathbb{E}_{\tilde{p}}\left[f_i(\mathbf{x})\right], \quad \sigma_i^\star = \frac{1}{N\,\mathbb{E}_{\tilde{p}}\left[\epsilon_i f_i(\mathbf{x})\right]} \tag{10}$$

where $\tilde{p} = p(\mathbf{x}|\theta(\epsilon))p(\epsilon)$ is an extended system of the original undirected model in which the parameters $\theta_i = \mu_i + \epsilon_i\sigma_i$ fluctuate according to the variational distribution. This bridges to the Maximum Likelihood solution as $N \to \infty$ and $\sigma_i^\star \to 0$, while accounting for uncertainty in the parameters at finite sample sizes with the inverse of 'sensitivity' $\mathbb{E}_{\tilde{p}}\left[\epsilon_i f_i(\mathbf{x})\right]$.

3. Fadeout

3.1. Noncentered Parameterizations of Hierarchical Priors

Hierarchical models are powerful because they impose *a priori* correlations between latent variables that reflect problem-specific knowledge. For scale-mixture priors that promote sparsity, these correlations come in the form of *scale uncertainty*. Instead of assuming that the scale of a parameter in a model is known *a priori*, we posit that it is normally distributed with a randomly distributed variance $p(\sigma^2)$. The joint prior $p(\theta|\sigma^2)p(\sigma^2)$ gives rise to a strongly curved 'funnel' shape (Figure 2) that illustrates a simple but profound principle about hierarchical models:

Table 1. Common priors as scale-mixtures of normal distributions

Prior	Hyperprior	$p(\log\sigma)$
Gaussian (L_2)	$\sigma^2 = \frac{1}{2\lambda}$	constant
Laplacian (L_1)	$\sigma^2 \sim$ Exponential	$2\lambda e^{-\lambda\sigma^2}\sigma^2$
Student-t (ARD)	$\sigma^2 \sim$ Inv. Gamma	$\frac{2\beta^\alpha}{\Gamma(\alpha)} e^{-\frac{\beta}{\sigma^2}} \sigma^{-2\alpha}$
Horseshoe	$\sigma \sim$ Half-Cauchy	$\frac{2s}{\pi}\frac{\sigma}{s^2+\sigma^2}$

Algorithm 1 Computing ∇ELBO for Fadeout

Require: Global parameters $\{\mu_\tau, s_\tau\}$
Require: Local parameters $\{\mu_{\tilde{\theta}}, \mu_{\log\boldsymbol{\sigma}}, s_{\tilde{\theta}}, s_{\log\boldsymbol{\sigma}}\}$
Require: Hyperprior gradient $\nabla_{\log\boldsymbol{\sigma},\boldsymbol{\tau}} \log p(\log\boldsymbol{\sigma}, \boldsymbol{\tau})$
Require: Likelihood gradient $\nabla_\theta p(\mathbf{x}|\boldsymbol{\theta})$
// Sample from variational distribution
$\mathbf{z}_1 \sim \mathcal{N}(0, I_{|\boldsymbol{\tau}|})$, $\mathbf{z}_2 \sim \mathcal{N}(0, I_{|\tilde{\boldsymbol{\theta}}|})$, $\mathbf{z}_3 \sim \mathcal{N}(0, I_{|\boldsymbol{\sigma}|})$
$\boldsymbol{\tau} \leftarrow \mu_\tau + \exp\{s_\tau\} \odot \mathbf{z}_1$
$\tilde{\boldsymbol{\theta}} \leftarrow \mu_{\tilde{\theta}} + \exp\{s_{\tilde{\theta}}\} \odot \mathbf{z}_2$
$\boldsymbol{\sigma} \leftarrow \exp\left\{\mu_{\log\sigma} + \exp\left\{s_{\log\sigma}\right\} \odot \mathbf{z}_3\right\}$
$\boldsymbol{\theta} \leftarrow \tilde{\boldsymbol{\theta}} \odot \boldsymbol{\sigma}$
// Centered global parameters
$\nabla_{\mu_{\boldsymbol{\tau}}}\mathcal{L} \leftarrow \nabla_{\boldsymbol{\tau}} \log p(\log\sigma, \boldsymbol{\tau})$
$\nabla_{s_{\boldsymbol{\tau}}}\mathcal{L} \leftarrow \exp\left\{s_{\boldsymbol{\tau}}\right\} \odot \mathbf{z}_1 \odot \nabla_{\mu_{\boldsymbol{\tau}}}\mathcal{L} + 1$
// Noncentered local parameters
$\nabla_{\mu_{\tilde{\theta}}}\mathcal{L} \leftarrow \boldsymbol{\sigma} \odot \nabla_\theta \log p(\boldsymbol{x}|\boldsymbol{\theta}) - \tilde{\boldsymbol{\theta}}$
$\nabla_{\mu_{\log\sigma}}\mathcal{L} \leftarrow \boldsymbol{\theta} \odot \nabla_\theta \log p(\boldsymbol{x}|\boldsymbol{\theta}) + \nabla_{\log\sigma} \log p(\log\sigma, \boldsymbol{\tau})$
$\nabla_{s_{\tilde{\theta}}}\mathcal{L} \leftarrow \exp\left\{s_{\tilde{\theta}}\right\} \odot \mathbf{z}_2 \odot \nabla_{\mu_{\tilde{\theta}}}\mathcal{L} + 1$
$\nabla_{s_{\log\sigma}}\mathcal{L} \leftarrow \exp\left\{s_{\log\sigma}\right\} \odot \mathbf{z}_3 \odot \nabla_{\mu_{\log\sigma}}\mathcal{L} + 1$

as the hyperparameter $\log\sigma$ decreases and the prior accepts a smaller range of values for θ, normalization increases the probability density at the origin, favoring sparsity. This normalization-induced sharpening has been called called a Bayesian Occam's Razor (MacKay, 2003).

While normalization-induced sharpening gives rise to sparsity, these extreme correlations are a disaster for mean-field variational inference. Even if a tremendous amount of probability mass is concentrated at the base of the funnel, an uncorrelated mean-field approximation will yield estimates near the top. The result is a potentially non-sparse estimate from a very-sparse prior.

The strong coupling of hierarchical funnels also plagues exact methods based on MCMC with slow mixing, but the statistics community has found that these geometry pathologies can be effectively managed by transformations. Many models can be rewritten in a noncentered form where the parameters and hyperparmeters are *a priori* independen (Papaspiliopoulos et al., 2007; Betancourt & Girolami, 2013). For the scale-mixtures of normals, this change of variables is

$$\{\theta, \log\sigma\} \to \left\{\frac{\theta}{\sigma}, \log\sigma\right\} \tag{11}$$

Then $\tilde{\theta} \triangleq \frac{\theta}{\sigma} \sim N(0,1)$ while preserving $\tilde{\theta}\sigma \sim N(0,\sigma^2)$. In noncentered form, the joint prior is independent and well approximated by a mean-field Gaussian, while the likelihood will be variably correlated depending on the strength of the data (Figure 2). In this sense, centered parameterizations (CP) and noncentered parameterizations (NCP) are usually framed as favorable in strong and weak data

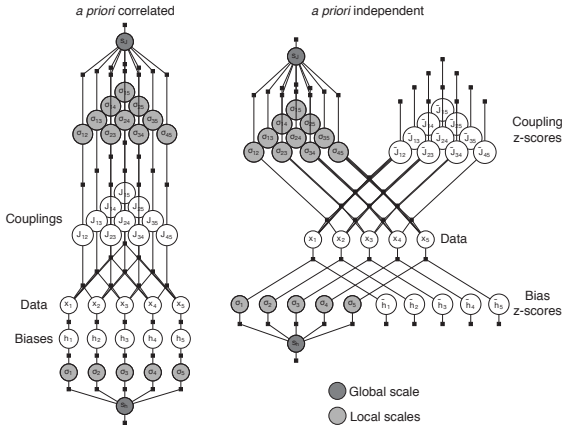

Figure 3. An undirected model with a scale mixture prior (factor graph on left) can be given *a priori* independence of the latent variables by a noncentered parameterization (factor graph on right). This is advantageous for mean-field variational inference that imposes *a posteriori* independence.

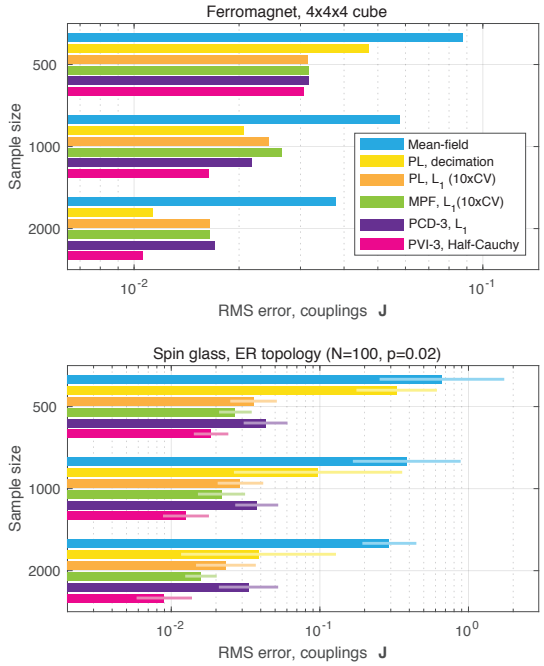

Figure 4. Inverse Ising. Combining Persistent VI with a noncentered Horseshoe prior (Half-Cauchy hyperprior) attains lower error on simulated Ising systems than standard methods for point estimation including: Pseudolikelihood (PL) with L_1 or decimation regularization (Schmidt, 2010; Aurell & Ekeberg, 2012; Decelle & Ricci-Tersenghi, 2014), Minimum Probability Flow (MPF) (Sohl-Dickstein et al., 2011), and Persistent Contrastive Divergence (PCD) (Tieleman, 2008). For the spin glass, error bars are two logarithmic standard deviations across 5 simulated systems.

regimes, respectively.[5]

We propose the use of non-centered parameterizations of scale-mixture priors for mean-field Gaussian variational inference. For convenience, we like to call this Fadeout (see next section). Fadeout can be easily implemented by either (i) using the chain rule to derive the gradient of the Evidence Lower BOund (ELBO) (Algorithm 1) or, for differentiable models, (ii) rewriting models in noncentered form and using automatic differentiation tools such as Stan (Kucukelbir et al., 2017) or `autograd`[6] for ADVI. The only two requirements of the user are the gradient of the likelihood function and a choice of a global hyperprior, several options for which are presented in Table 1.

Estimators for the centered posterior. Fadeout optimizes a mean-field Gaussian variational distribution over the noncentered parameters $q(\tilde{\boldsymbol{\theta}}, \log \boldsymbol{\sigma})$. As an estimator for the centered parameters, we use the mean-field property to compute the centered posterior mean as $\mathbb{E}_q[\theta] = \mathbb{E}_q[\tilde{\boldsymbol{\theta}}] \odot \mathbb{E}_q[\boldsymbol{\sigma}]$, giving [7]

$$\hat{\theta} = \boldsymbol{\mu}_{\tilde{\theta}} \odot \exp\left\{\boldsymbol{\mu}_{\log \boldsymbol{\sigma}} + \frac{1}{2} e^{2 s_{\log \boldsymbol{\sigma}}}\right\} \tag{12}$$

[5] Although "weak data" may seem unrepresentative of typical problems in machine learning, it is important to remember that a sufficiently large and expressive model can make most data weak.

[6] github.com/HIPS/autograd

[7] The term $\frac{1}{2} e^{2 s_{\log \boldsymbol{\sigma}}}$ is optional in the sense that including it corresponds to averaging over the hyperparameters, whereas discarding it corresponds to optimizing the hyperparameters (Empirical Bayes). We included it for all experiments.

3.2. Connection to Dropout

Dropout regularizes neural networks by perturbing hidden units in a directed network with multiplicative Bernoulli or Gaussian noise (Srivastava et al., 2014). Although it was originally framed as a heuristic, Dropout has been subsequently interpreted as variational inference under at least two different schemes (Gal & Ghahramani, 2016; Kingma et al., 2015). Here, we interpret Fadeout the reverse way, where we introduced it as variational inference and now notice that it looks similar to lognormal Dropout.[8] If we take the uncertainty in $\tilde{\boldsymbol{\theta}}$ as low and clamp the other variational parameters, the gradient estimator for Fadeout is:

$$\begin{aligned}
\mathbf{z} &\sim \mathcal{N}(0, I_{|\theta|}) \\
\boldsymbol{\sigma} &\leftarrow \exp\{\mu_{\log \sigma} + \exp\{s_{\log \sigma}\} \odot \mathbf{z}\} \\
\boldsymbol{\theta} &\leftarrow \tilde{\boldsymbol{\theta}} \exp\{\mu_{\log \sigma} + \exp\{s_{\log \sigma}\} \odot \mathbf{z}\} \\
\nabla_{\mu_{\tilde{\theta}}} \mathcal{L} &\leftarrow \boldsymbol{\sigma} \odot \nabla_\theta \log p(\boldsymbol{x}|\theta) - \tilde{\boldsymbol{\theta}}
\end{aligned}$$

[8] Rather than attempting to explain Dropout, the intent is to lend intuition about noncentered scale-mixture VI.

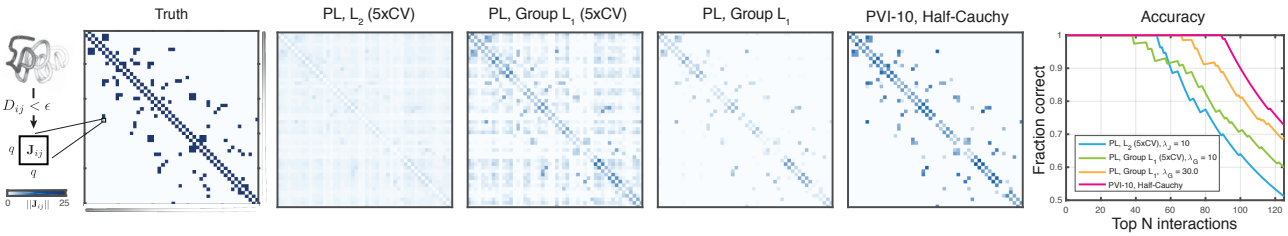

Figure 5. Synthetic protein. For reconstructing interactions in a synthetic 20-letter spin-glass, a hierarchical Bayesian approach based on Persistent VI and a noncentered group Horseshoe prior (Half-Cauchy hyperprior) identifies true interactions with more accuracy and less shrinkage than Group L_1. Each i, j pair is the norm of a 20×20 factor coupling the amino acid at position i to the amino acid at position j.

This is the gradient estimator for a lognormal version of Dropout with an L_2 weight penalty of $\frac{1}{2}$. At each sample from the variational distribution, Fadeout introduces *scale noise* rather than the Bernoulli noise of Dropout. The connection to Dropout would seem to follow naturally from the common interpretation of scale mixtures as continuous relaxations of spike and slab priors (Engelhardt & Adams, 2014) and the idea that Dropout can be related to variational spike and slab inference (Louizos, 2015).

4. Experiments

4.1. Physics: Inferring Spin Models

Ising model The Ising model is a prototypical undirected model for binary systems that includes both pairwise interactions and (potentially) sitewise biases. It can be seen as the fully observed case of the Boltzmann machine, and is typically parameterized with signed spins $\mathbf{x} \in \{-1, 1\}^D$ and a likelihood given by

$$p(\mathbf{x}|\mathbf{h}, \mathbf{J}) = \frac{1}{Z(\mathbf{h}, \mathbf{J})} \exp\left\{\sum_i h_i x_i + \sum_{i<j} J_{ij} x_i x_j\right\}. \quad (13)$$

Originally proposed as a minimal model of how long range order arises in magnets, it continues to find application in physics and biology as a model for phase transitions and quenched disorder in spin glasses (Nishimori, 2001) and collective firing patterns in neural spike trains (Schneidman et al., 2006; Shlens et al., 2006).

Hierarchical sparsity prior One appealing feature of the Ising model is that it allows a sparse set of underlying couplings $\mathbf{J}$ to give rise to long-range, distributed correlations across a system. Since many physical systems are thought to be dominated by a small number of relevant interactions, L_1 regularization has been a favored approach for inferring Ising models. Here, we examine how a more accurate model of sparsity based on the Horseshoe prior (Figure 3) can improve inferences in these systems. Each coupling J_{ij} and bias parameter h_i is given its own scale parameter which are in turn tied under a global Half-Cauchy prior for the scales (Figure 3, Appendix).

Simulated datasets We generated synthetic couplings for two kinds of Ising systems: (i) a slightly sub-critical cubic ferromagnet ($J_{ij} > 0$ for neighboring spins) and (ii) a Sherrington-Kirkpatrick spin glass diluted on an Erdös-Renyi random graph with average degree 2. We sampled synthetic data for each system with the Swendsen-Wang algorithm (Appendix) (Swendsen & Wang, 1987).

Results On both the ferromagnet and the spin glass, we found that Persistent VI with a noncentered Horseshoe prior (Fadeout) gave estimates with systematically lower reconstruction error of the couplings $\mathbf{J}$ (Figure 4) versus a variety of standard methods in the field (Appendix).

4.2. Biology: Reconstructing 3D Contacts in Proteins from Sequence Variation

Potts model The Potts model generalizes the Ising model to non-binary categorical data. The factor graph is the same (Figure 3), except each spin x_i can adopt q different categories with $\mathbf{x} \in \{1, \ldots, q\}^D$ and each $\mathbf{J_{ij}}$ is a $q \times q$ matrix as

$$p(\mathbf{x}|\mathbf{h}, \mathbf{J}) = \frac{1}{Z(\mathbf{h}, \mathbf{J})} \exp\left\{\sum_i h_i(x_i) + \sum_{i<j} J_{ij}(x_i, x_j)\right\}. \quad (14)$$

The Potts model has recently generated considerable excitement in biology, where it has been used to infer 3D contacts in biological molecules solely from patterns of correlated mutations in the sequences that encode them (Marks et al., 2011; Morcos et al., 2011). These contacts are have been sufficient to predict the 3D structures of proteins, protein complexes, and RNAs (Marks et al., 2012).

Group sparsity Each pairwise factor $\mathbf{J}_{ij}$ in a Potts model contains $q \times q$ parameters capturing all possible joint configurations of x_i and x_j. One natural way to enforce spar-

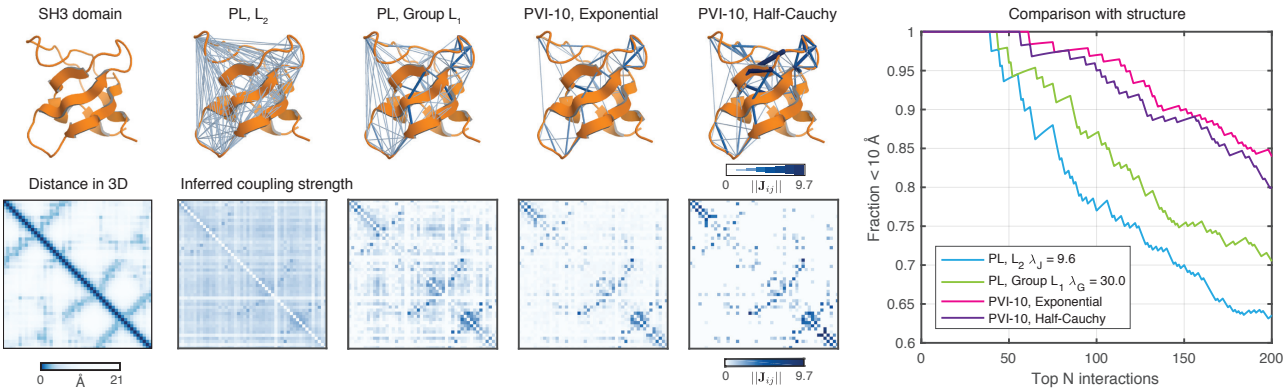

Figure 6. Unsupervised protein contact prediction. When inferring a pairwise undirected model for protein sequences in the SH3 domain family, hierarchical Bayesian approaches based on Persistent VI and noncentered scale mixture priors (Half-Cauchy for Group Horseshoe and Exponential for a Multivariate Laplace) identify local interactions that are close in 3D structure without tuning parameters. When group L_1-regularized maximum Pseudolikelihood estimation is tuned to give the same largest effect size as the Multivariate Laplace, the hierarchical approaches based on Persistent VI are more predictive of 3D proximity (right).

sity in a Potts model is at the level of each $q \times q$ group. This can be accomplished by introducing a single scale parameter σ_{ij} for all $q \times q$ z-scores $\tilde{\mathbf{J}}_{ij}$. We adopt this with the same Half-Cauchy hyperprior as the Ising problem, giving the same factor graph (Figure 3) now corresponding to a Group Horseshoe prior (Hernández-Lobato et al., 2013). In the real protein experiment, we also consider an exponential hyperprior, which corresponds to a Multivariate Laplace distribution (Eltoft et al., 2006) over the groups.

Synthetic protein data We first investigated the performance of Persistent VI with group sparsity on a synthetic protein experiment. We constructed a synthetic Potts spin glass with a topology inspired by biological macromolecules. We generated synthetic parameters based on contacts in a simulated polymer and sampled 2000 sequences with 2×10^6 steps of Gibbs sampling (Appendix).

Results for a synthetic protein We inferred couplings with 400 of the sampled sequences using PVI with group sparsity and two standard methods of the field: L_2 and Group L_1 regularized maximum pseudolikelihood (Appendix). PVI with a noncentered Horseshoe yielded more accurate (Figure 5, right), less shrunk (Figure 5, left) estimates of interactions that were more predictive of the 1600 remaining test sequences (Table 2). The ability to generalize well to new sequences will likely be important to the related problem of predicting mutation effects with unsupervised models of sequence variation (Hopf et al., 2017; Figliuzzi et al., 2015).

Results for natural sequence variation We applied the hierarchical Bayesian model from the protein simulation to model across-species amino acid covariation in the SH3 domain family (Figure 6). Transitioning from simulated to real protein data is particularly challenging for Bayesian methods because available sequence data are highly non-independent due to a shared evolutionary history. We developed a new method for estimating the effective sample size (Appendix) which, when combined standard sequence reweighting techniques, yielded a reweighted effective sample size of 1,012 from 10,209 sequences.

The hierarchical Bayesian approach gave highly localized, sparse estimates of interactions compared to the two predominant methods in the field, L_2 and group L_1 regularized pseudolikelihood (Figure 6). When compared to solved 3D structures for SH3 (Appendix), we found that the inferred interactions were considerably more accurate at predicting amino acids close in structure. Importantly, the hierarchical Bayesian approach accomplished this inference of strong, accurate interactions without a need to prespecify hyperparameters such as λ for L_2 or L_1 regularization. This is particularly important for natural biological sequences because the non-independence of samples limits the utility of cross validation for setting hyperparameters.

5. Related work

5.1. Variational Inference

One strategy for improving variational inference is to introduce correlations in variational distribution by geometric transformations. This can be made particularly powerful

Table 2. Average log-pseudolikelihood for test sequences.

Method	$-\log \mathrm{PL}(\mathbf{x}\|\mathbf{h},\mathbf{J})$	Runtime (s)
PL, L_2 (5xCV)	67.3	375
PL, Group L_1 (5xCV)	59.6	303
PVI-3, Half-Cauchy	54.2	585

by using backpropagation to learn compositions of transformations that capture the geometry of complex posteriors (Rezende & Mohamed, 2015; Tran et al., 2016). Noncentered parameterizations of models may be complementary to these approaches by enabling more efficient representations of correlations between parameters and hyperparameters.

Most related to this work, (Louizos et al., 2017; Ghosh & Doshi-Velez, 2017) show how variational inference with noncentered scale-mixture priors can be useful for Bayesian learning of neural networks, and how group sparsity can act as a form of automatic compression and model selection.

5.2. Maximum Entropy

Much of the work on inference of undirected graphical models has gone under the name of the Maximum Entropy method in physics and neuroscience, which can be equivalently formulated as maximum likelihood in an exponential family (MacKay, 2003). From this maximum likelihood interpretation, L_1 regularized-maximum entropy modeling (MaxEnt) corresponds to the disfavored "integrate-out" approach to inference in hierarchical models[9] (MacKay, 1996) that will introduce significant biases to inferred parameters (Macke et al., 2011). One solution to this bias was foreshadowed by methods for estimating entropy and Mutual Information, which used hierarchical priors to integrate over a large range of possible model complexities (Nemenman et al., 2002; Archer et al., 2013). These hierarchical approaches are favorable because in traditional MAP estimation any top level parameters that are fixed before inference (e.g. a global pseudocount α) introduce strong constraints on allowed model complexity. The improvements from PVI and Fadeout may be seen as extending this hierarchical approach to full systems of discrete variables.

6. Conclusion

We introduced a framework for scalable Bayesian sparsity for undirected graphical models composed of two methods. The first is an extension of stochastic variational inference to work with undirected graphical models that uses persistent gradient estimation to bypass estimating partition functions. The second is a variational approach designed to match the geometry of hierarchical, sparsity-promoting priors. We found that, when combined, these two methods give substantially improved inferences of undirected graphical models on both simulated and real systems from physics and computational biology.

[9] To see this, note that L_1-regularized MAP estimation is equivalent to integrating out a zero-mean Gaussian prior with unknown, exponentially-distributed variance

Acknowledgements

We thank David Duvenaud, Finale Doshi-Velez, Miriam Huntley, Chris Sander, and members of the Marks lab for helpful comments and discussions. JBI was supported by a NSF Graduate Research Fellowship DGE1144152 and DSM by NIH grant 1R01-GM106303. Portions of this work were conducted on the Orchestra HPC Cluster at Harvard Medical School.

References

Andrews, David F and Mallows, Colin L. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 99–102, 1974.

Archer, Evan, Park, Il Memming, and Pillow, Jonathan W. Bayesian and quasi-bayesian estimators for mutual information from discrete data. *Entropy*, 15(5):1738–1755, 2013.

Aurell, Erik and Ekeberg, Magnus. Inverse ising inference using all the data. *Physical review letters*, 108(9): 090201, 2012.

Betancourt, MJ and Girolami, Mark. Hamiltonian monte carlo for hierarchical models. *arXiv preprint arXiv:1312.0906*, 2013.

Carvalho, Carlos M, Polson, Nicholas G, and Scott, James G. The horseshoe estimator for sparse signals. *Biometrika*, pp. asq017, 2010.

Chen, Yutian and Welling, Max. Bayesian structure learning for markov random fields with a spike and slab prior. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 174–184. AUAI Press, 2012.

Decelle, Aurélien and Ricci-Tersenghi, Federico. Pseudolikelihood decimation algorithm improving the inference of the interaction network in a general class of ising models. *Physical review letters*, 112(7):070603, 2014.

Eltoft, Torbjørn, Kim, Taesu, and Lee, Te-Won. On the multivariate laplace distribution. *IEEE Signal Processing Letters*, 13(5):300–303, 2006.

Engelhardt, Barbara E and Adams, Ryan P. Bayesian structured sparsity from gaussian fields. *arXiv preprint arXiv:1407.2235*, 2014.

Figliuzzi, Matteo, Jacquier, Hervé, Schug, Alexander, Tenaillon, Oliver, and Weigt, Martin. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase tem-1. *Molecular biology and evolution*, pp. msv211, 2015.

Gal, Yarin and Ghahramani, Zoubin. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 1050–1059, 2016.

Ghosh, Soumya and Doshi-Velez, Finale. Model selection in bayesian neural networks via horseshoe priors. *arXiv*

preprint arXiv:1705.10388, 2017.

Hernández-Lobato, Daniel, Hernández-Lobato, José Miguel, and Dupont, Pierre. Generalized spike-and-slab priors for bayesian group feature selection using expectation propagation. *Journal of Machine Learning Research*, 14(1):1891–1945, 2013.

Hoffman, Matthew D, Blei, David M, Wang, Chong, and Paisley, John. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Hopf, Thomas A, Ingraham, John B, Poelwijk, Frank J, Schärfe, Charlotta PI, Springer, Michael, Sander, Chris, and Marks, Debora S. Mutation effects predicted from sequence co-variation. *Nature biotechnology*, 35(2): 128–135, 2017.

Jordan, Michael I, Ghahramani, Zoubin, Jaakkola, Tommi S, and Saul, Lawrence K. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Kingma, Diederik P and Welling, Max. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

Kingma, DP, Salimans, T, and Welling, M. Variational dropout and the local reparameterization trick. *Advances in Neural Information Processing Systems*, 28:2575–2583, 2015.

Koller, Daphne and Friedman, Nir. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Kucukelbir, Alp, Tran, Dustin, Ranganath, Rajesh, Gelman, Andrew, and Blei, David M. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(14):1–45, 2017.

Louizos, Christos. Smart regularization of deep architectures. Master's thesis, University of Amsterdam, 2015.

Louizos, Christos, Ullrich, Karen, and Welling, Max. Bayesian compression for deep learning. *arXiv preprint arXiv:1705.08665*, 2017.

MacKay, David JC. Hyperparameters: Optimize, or integrate out? In *Maximum entropy and bayesian methods*, pp. 43–59. Springer, 1996.

MacKay, David JC. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

MacKay, David JC et al. Bayesian nonlinear modeling for the prediction competition. *ASHRAE transactions*, 100 (2):1053–1062, 1994.

Macke, Jakob H, Murray, Iain, and Latham, Peter E. How biased are maximum entropy models? In *Advances in Neural Information Processing Systems*, pp. 2034–2042, 2011.

Marks, Debora S, Colwell, Lucy J, Sheridan, Robert, Hopf, Thomas A, Pagnani, Andrea, Zecchina, Riccardo, and Sander, Chris. Protein 3d structure computed from evolutionary sequence variation. *PloS one*, 6(12):e28766, 2011.

Marks, Debora S, Hopf, Thomas A, and Sander, Chris. Protein structure prediction from sequence variation. *Nature biotechnology*, 30(11):1072–1080, 2012.

Mitchell, Toby J and Beauchamp, John J. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

Mohamed, Shakir, Ghahramani, Zoubin, and Heller, Katherine A. Bayesian and l1 approaches for sparse unsupervised learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pp. 751–758, 2012.

Morcos, Faruck, Pagnani, Andrea, Lunt, Bryan, Bertolino, Arianna, Marks, Debora S, Sander, Chris, Zecchina, Riccardo, Onuchic, José N, Hwa, Terence, and Weigt, Martin. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49): E1293–E1301, 2011.

Murphy, Kevin P. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Murray, Iain and Ghahramani, Zoubin. Bayesian learning in undirected graphical models: approximate mcmc algorithms. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pp. 392–399. AUAI Press, 2004.

Murray, Iain, Ghahramani, Zoubin, and MacKay, David JC. Mcmc for doubly-intractable distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pp. 359–366. AUAI Press, 2006.

Nemenman, Ilya, Shafee, Fariel, and Bialek, William. Entropy and inference, revisited. *Advances in neural information processing systems*, 1:471–478, 2002.

Nishimori, Hidetoshi. *Statistical physics of spin glasses and information processing: an introduction*. Number 111. Clarendon Press, 2001.

Papaspiliopoulos, Omiros, Roberts, Gareth O, and Sköld, Martin. A general framework for the parametrization of hierarchical models. *Statistical Science*, pp. 59–73, 2007.

Ranganath, Rajesh, Gerrish, Sean, and Blei, David. Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pp. 814–822, 2014.

Rezende, Danilo and Mohamed, Shakir. Variational inference with normalizing flows. In *Proceedings of The 32nd International Conference on Machine Learning*, pp. 1530–1538, 2015.

Rezende, Danilo Jimenez, Mohamed, Shakir, and Wierstra, Daan. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of The 31st International Conference on Machine Learning*, pp. 1278–1286, 2014.

Schmidt, Mark. *Graphical model structure learning*

with l1-regularization. PhD thesis, UNIVERSITY OF BRITISH COLUMBIA (Vancouver, 2010.

Schneidman, Elad, Berry, Michael J, Segev, Ronen, and Bialek, William. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012, 2006.

Shlens, Jonathon, Field, Greg D, Gauthier, Jeffrey L, Grivich, Matthew I, Petrusca, Dumitru, Sher, Alexander, Litke, Alan M, and Chichilnisky, EJ. The structure of multi-neuron firing patterns in primate retina. *The Journal of neuroscience*, 26(32):8254–8266, 2006.

Sohl-Dickstein, Jascha, Battaglino, Peter B, and DeWeese, Michael R. New method for parameter estimation in probabilistic models: minimum probability flow. *Physical review letters*, 107(22):220601, 2011.

Srivastava, Nitish, Hinton, Geoffrey, Krizhevsky, Alex, Sutskever, Ilya, and Salakhutdinov, Ruslan. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1): 1929–1958, 2014.

Swendsen, Robert H and Wang, Jian-Sheng. Nonuniversal critical dynamics in monte carlo simulations. *Physical review letters*, 58(2):86, 1987.

Tibshirani, Robert. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

Tieleman, Tijmen. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, pp. 1064–1071. ACM, 2008.

Tipping, Michael E. Sparse bayesian learning and the relevance vector machine. *The journal of machine learning research*, 1:211–244, 2001.

Titsias, Michalis and Lázaro-Gredilla, Miguel. Doubly stochastic variational bayes for non-conjugate inference. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1971–1979, 2014.

Tran, Dustin, Ranganath, Rajesh, and Blei, David M. The variational gaussian process. In *Proceedings of the International Conference on Learning Representations*, 2016.

Younes, Laurent. Parametric inference for imperfectly observed gibbsian fields. *Probability theory and related fields*, 82(4):625–645, 1989.