Abstract
Bayesian approaches for single-variable and group-structured sparsity outperform $L_1$ regularization, but are challenging to apply to large, potentially intractable models. Here we show how noncentered parameterizations, a common trick for improving the efficiency of exact inference in hierarchical models, can similarly improve the accuracy of variational approximations. We develop this with two contributions: First, we introduce Fadeout, an approach for variational inference that uses noncentered parameterizations to capture a posteriori correlations between parameters and hyperparameters. Second, we extend stochastic variational inference to undirected models, enabling efficient hierarchical Bayes without approximations of intractable normalizing constants. We find that this framework substantially improves inferences of undirected graphical models under both sparse and group-sparse priors.

1. Introduction
Hierarchical priors that favor sparsity have been a central development in modern statistics and machine learning. While they have a long history within the confines of statistics, it was perhaps $L_1$-regularization and the LASSO (Tibshirani, 1996) that helped instantiate sparsity-promoting priors as standard tools of applied science and engineering. Nevertheless, $L_1$ is a pragmatic compromise; by being the closest convex approximation of the idealized $L_0$ norm, it cannot model the hypothesis of sparsity as well as some Bayesian alternatives. Bayesian sparsity-promoting priors can yield considerably better results (Tipping, 2001), but only when Bayesian inference is tractable.

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 typically viewed as the ideal prior for sparsity (Mohamed et al., 2011), but the discrete latent space that it imposes is often too significant a barrier for high-dimensional models.

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

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

Scale-mixtures can approximate the discrete spike and slab prior by mixing both large and small values of $\sigma$. $L_1$’s implicit prior, the Laplacian, is a member of this family and results from exponentially distributed variance $\sigma^2$. Thus, mixing densities $p(\sigma^2)$ with subexponential tails and more mass near the origin can better accommodate sparsity than $L_1$ and are the basis for approaches often referred to as “Sparse Bayesian Learning” (Tipping, 2001). The Student-$t$ and Horseshoe (Carvalho et al., 2010) both incorporate these properties, but unfortunately hinder MAP inference with non-convexities and typically require the use of either HMC or augmentation-based approaches (Polson et al., 2013).

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 clear directions of causality, but their likelihoods require an intractable normalizing constant (Murray & Ghahramani, 2004). For a fully observed Boltzmann Machine with $x \in \{0, 1\}^D$ the distribution\footnote{we exclude biases for simplicity} is

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

where the partition function $Z(J)$ depends on the couplings. Whenever a new set of couplings $J$ is considered...
Bayesian Sparsity for Intractable Distributions

Figure 1. Discrete Bayesian sparsity is triply intractable for discrete undirected graphical models, 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.

During inference, the partition function $Z(J)$ and corresponding density $p(x|J)$ must be reevaluated. This requirement for 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., 2012). When all $2^J$ 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 this problem is impossible, merely expensive (Chen & Welling, 2014).

Here we explore an alternative to MCMC-based approaches for Bayesian sparsity with stochastic variational inference (Hoffman et al., 2013). We combine three ideas: (i) a noncentered parameterization of scale-mixture priors, (ii) stochastic gradient variational Bayes (Kingma & Welling, 2013; Rezende et al., 2014; Titsias & Lázaro-Gredilla, 2014)\(^2\), and (iii) persistent Markov chains (Younes, 1989; Neal, 1992; Tieleman, 2008) to inherit the benefits of hierarchical Bayesian sparsity in an efficient variational framework. We make the following contributions:

- We introduce Fadeout, a method for mean-field variational inference under non-centered parameterizations of scale-mixture priors that captures correlations due to scale uncertainty (Section 2).

- We show how stochastic variational inference can be applied to undirected models with intractable normalizing constants by using persistent Markov chains, which we call persistent stochastic variational inference (PSVI) (Section 3).

- We experimentally demonstrate the improvement of Bayesian sparsity with noncentered stochastic variational inference across undirected graphical models with both single-variable and group-structured sparsity (Section 4).

2. Fadeout

2.1. 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: as the hyperparameter $\log \sigma$ decreases and the prior accepts a smaller range of values for $\theta$, 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; Rasmussen & Ghahramani, 2001).

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 hyperparameters are conditionally independent given the data (Papaspiliopoulos et al., 2007; Betancourt & Girolami, 2013). For the scale-mixtures of normals, this change of variables is

$$\{\theta, \log \sigma\} \rightarrow \left\{\frac{\theta}{\sigma}, \log \sigma\right\} \quad (3)$$

Then $\tilde{\theta} \equiv \frac{\theta}{\sigma} \sim N(0, 1)$ while preserving $\tilde{\theta}\sigma \sim N(0, \sigma^2)$. In noncentered form, the joint prior is uncorrelated and well approximated by a mean-field Gaussian, while the likeli-
Bayesian Sparsity for Intractable Distributions

Figure 2. Noncentered variational inference by Fadeout. Hierarchical priors like the Horsehoe (shown here) are extremely correlated and poorly approximated by a mean-field Gaussian. Noncentered reparameterizations trade independence in the likelihood for independence in the prior, allowing a mean-field Gaussian over the noncentered parameters (black contours, bottom right) to capture the funnel geometry of the centered parameters (black contours, top right). By the invariance of the KL divergence, these variational problems are equivalent.

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

| Prior                  | Mixing distribution | $p(\log \sigma)$ |
|------------------------|---------------------|-------------------|
| Gaussian ($L_2$)       | $\sigma^2 = \frac{1}{2\lambda}$ | $\delta (\sigma^2 - \frac{1}{2\lambda})$ |
| Laplacian ($L_1$)      | $\sigma^2 \sim \text{Exp}(\lambda)$ | $2\lambda e^{-\lambda \sigma^2}$ |
| Student-t (ARD)        | $\sigma^2 \sim \mathcal{IG}(\alpha, \beta)$ | $\frac{2\beta}{\Gamma(\alpha)} e^{-\frac{2}{\beta} \sigma^{-2\alpha}}$ |
| Horseshoe              | $\sigma \sim \mathcal{C}^+(0, s)$ | $\frac{2s}{\pi s^2 + \sigma^2}$ |

Noncentered parameterizations may seem antithetical to modern contexts of machine learning, but it is important to remember that a sufficiently large and expressive model can make most data weak. Modeling scenarios where hierarchical priors give interesting results are usually modeling scenarios in which the hierarchical prior has not been overwhelmed by the data. Hence it would seem noncentered parameterizations could be greatly beneficial in maximizing the utility of variational mean-field approximations.

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), as done for single-variable sparsity in Algorithm 1 or (ii) rewriting models in noncentered form and using automatic differentiation tools such as Stan (Kucukelbir et al., 2015) or autograd for black-box SVI. 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.

Crucially, by using scale-mixtures instead of the spike and slab Fadeout does not require score-function estimators based on likelihood. This will make it compatible with the algorithm for stochastic variational inference in undirected graphical models presented in section 3.

Estimators for the centered posterior. Fadeout optimizes a mean-field Gaussian variational distribution over the noncentered parameters $q(\tilde{\theta}, \log \sigma)$. Because of the mean-field assumption, the posterior mean over the centered parameterization 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 regimes, respectively.

“Weak data” and thus noncentered parameterizations may seem antithetical to modern contexts of machine learning, but it is important to remember that a sufficiently large and expressive model can make most data weak. Modeling scenarios where hierarchical priors give interesting results are usually modeling scenarios in which the hierarchical prior has not been overwhelmed by the data. Hence it would seem noncentered parameterizations could be greatly beneficial in maximizing the utility of variational mean-field approximations.

3github.com/HIPS/autograd
4estimators of log-partition functions are extremely difficult to make unbiased with a short amount of computation
Bayesian Sparsity for Intractable Distributions

The term $\frac{1}{2}e^{2s_{log} \sigma}$ 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.

2.2. Connection to Dropout

Dropout regularizes inference by randomly perturbing hidden units in a directed network with Bernoulli or Gaussian noise (Srivastava et al., 2014). Subsequent to its widespread application for neural networks, it has been recognized as accidental variational inference under at least two different schemes (Gal & Ghahramani, 2015; Kingma et al., 2015). Here we find a similar case of an ‘automatic Dropout’ arising in variational inference by Fadeout. To see this, we focus on a single “unscaled” parameter $\tilde{\theta}$. If we take the uncertainty in $\tilde{\theta}$ as low and clamp the other variational parameters, we recover the gradient estimator

$$z \sim \mathcal{N}(0, I_{|\theta|})$$

$$\sigma \leftarrow \exp \left\{ \mu_{log} \sigma + s_{log} \sigma \right\} \odot z$$

$$\tilde{\theta} \leftarrow \sigma$$

$$\nabla_{\mu_{\mu}} \mathcal{L} \leftarrow \sigma \odot \nabla_{\sigma} \log p(x|\theta) - \tilde{\theta}$$

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, the correlated funnel of Fadeout injects scale uncertainty rather than the discrete Bernoulli inclusion & exclusion of Dropout\(^5\). 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).

3. Persistent Stochastic Variational Inference

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

$$p(x|\theta) = \frac{1}{Z(\theta)} \exp \left\{ \sum_i \theta_i f_i(x) \right\}$$

where $i$ indexes a set of features $\{f_i(x)\}$ that typically depend on subsets of the variables (Koller & Friedman, 2009). Given data $D = \{x^{(1)}, \ldots, x^{(N)}\}$ the gradient of the scaled log likelihood is

$$\frac{\partial}{\partial \theta_i} \frac{1}{N} \log p(D|\theta) = \mathbb{E}_D [f_i(x)] - \mathbb{E}_{p(x|\theta)} [f_i(x)]$$

where the first term is an average of feature $f_i(x)$ over the data while the second term is the average of feature $f_i(x)$ over the distribution defined by $\theta$. Computing these averages is intractable for all but the smallest systems, and so typically some form of sampling must be done every time $\theta$ is modified.

Doubly intractable ELBO. Variational inference selects the parameters of a variational distribution $q(\theta|\phi)$ by maximizing the Evidence Lower Bounded $\mathcal{L}(\phi)$:

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

Optimization requires $\nabla_{\phi} \mathcal{L}(\phi)$. Stochastic variational inference (Hoffman et al., 2013) proposes a stochastic approach where gradients are estimated with samples from $q(\theta|\phi)$ and the score-function estimator (Ranganath et al., 2013)

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

Naively substituting (5) in (8) nests an intractable log partition function within the average over $q(\theta|\phi)$. Estimators of $\log Z(\theta)$ are notorious for requiring extensive sampling to reach unbiasedness, suggesting this will be an untenable route to stochastic variational inference in undirected models. Fortunately, the ‘reparameterization trick’ (Kingma

Algorithm 1 $\nabla \text{ELBO}$ for single-variable sparsity

Require: Global parameters $\{\mu_{\mu}, s_{\mu}\}$
Require: Local parameters $\{\mu_{\phi}, \mu_{log} \sigma, \tilde{s}_{\phi}, s_{log} \sigma\}$
Require: Hyperprior gradient $\nabla_{log} \sigma, \tau \log p(\log \sigma, \tau)$
Require: Likelihood gradient $\nabla_{\phi} p(x|\theta)$

Sample from variational distribution

$$z_1 \sim \mathcal{N}(0, I_{|\tau|})$$

$$z_2 \sim \mathcal{N}(0, I_{|\tilde{\mu}|})$$

$$z_3 \sim \mathcal{N}(0, I_{|\mu_{log} \sigma}|)$$

$$\tau \leftarrow \mu_{\tau} + \exp\{s_{\tau}\} \odot z_1$$

$$\tilde{\mu} \leftarrow \mu_{\tilde{\mu}} + \exp\{s_{\tilde{\mu}}\} \odot z_2$$

$$\sigma \leftarrow \exp\{s_{log} \sigma\} \odot z_3$$

$$\theta \leftarrow \tilde{\theta} \odot \sigma$$

Centered global parameters

$$\nabla_{\mu_{\mu}} \mathcal{L} \leftarrow \nabla_\tau \log p(\log \sigma, \tau)$$

$$\nabla_{s_{\mu}} \mathcal{L} \leftarrow \exp\{s_{\tau}\} \odot z_1 \odot \nabla_{\mu_{\mu}} \mathcal{L} + 1$$

Noncentered local parameters

$$\nabla_{\mu_{\phi}} \mathcal{L} \leftarrow \sigma \odot \nabla_{\tilde{\mu}} \log p(x|\theta) - \tilde{\theta}$$

$$\nabla_{s_{log} \sigma} \mathcal{L} \leftarrow \theta \odot \nabla_{\tilde{\mu}} \log p(\theta) + \nabla_{log} \sigma \log p(\log \sigma, \tau)$$

$$\nabla_{\tilde{s}_{\phi}} \mathcal{L} \leftarrow \exp\{s_{\tilde{\mu}}\} \odot z_3 \odot \nabla_{\tilde{\mu}} \mathcal{L} + 1$$

$$\nabla_{log} \sigma \mathcal{L} \leftarrow \exp\{s_{log} \sigma\} \odot z_3 \odot \nabla_{\log} \sigma \mathcal{L} + 1$$

\(^5\)The connection to Dropout in directed nets can be developed with group-sparse priors over all outgoing weights from each unit.
and instead rely on its gradient (6), which can be efficiently estimated using persistent Markov chains.

**Persistent Markov chains.** Stochastic maximum likelihood for undirected models uses sampling of (6) as an estimator for the true gradient (Murphy, 2012). It was shown in (Younes, 1989) that this procedure can be made considerably more efficient if the Markov chains used for sampling are persistent, where each chain at the current iteration is initialized with its final state from the previous iteration. This ensures that the Markov chains are near their equilibrium distributions, and, after being subject to small changes in the parameters $\theta$ during learning, can quickly reequilibrate. Neal used persistent Markov chains for learning sigmoid belief nets (Neal, 1992), which helped inspire the rediscovered application to undirected models by Tieleman (Tieleman, 2008) with Persistent Contrastive Divergence (PCD).

We propose the use of persistent Markov chains for estimating $\nabla \text{ELBO}$ in undirected models, and refer to this as Persistent Stochastic Variational Inference (PSVI). Following the notation of PCD-$k$, PSVI-$k$ refers stochastic variational inference with gradients estimated by persistent Markov Chains with $k$ sweeps of Gibbs sampling between iterations. This approach is generally compatible with any estimators of the ELBO based on the gradient of the log likelihood, several examples of which are explained in (Kingma & Welling, 2013; Rezende et al., 2014; Titsias & Lázaro-Gredilla, 2014).

### 4. Experiments

#### 4.1. Ising model

For single-variable sparsity, we focused on a prototypical undirected graphical model, the Ising model. It can be seen as a special, fully observed case of the Boltzmann machine, and is typically parameterized with signed spins $x \in \{-1,1\}^D$ and a likelihood given by

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

(11)

The Ising model has been a consistent benchmark for various approximate methods introduced for inference of undirected graphical models (Sohl-Dickstein et al., 2011; Aurell & Ekeberg, 2012). In addition to its original role in physics as a prototypical model for phase transitions and the quenched disorder in spin glasses (Nishimori, 2001), it also finds recent application in biology, as a model for the collective firing patterns of neural spike trains (Schneidman et al., 2006; Shlens et al., 2006) and the mutational constraints on the evolution of HIV (Ferguson et al., 2013). Many of these systems are defined by a sparse set of underlying interactions and thus favorable settings for deploying sparsity-promoting priors.

We generated three synthetic systems. The first was a cu-
bic ferromagnet, where neighboring spins $x_i, x_j$ have a favorable interaction of $J = 0.2$ based on a $4 \times 4 \times 4$ periodic lattice. This coupling strength equates to being slightly above the critical temperature, meaning the system will be highly correlated despite the underlying interactions being only nearest-neighbor. The second and third systems were diluted Sherrington-Kirkpatrick spin glasses (Sherrington & Kirkpatrick, 1975) at two levels of dilution over 100 spins. The couplings in these models are defined by Erdős-Rényi random graphs (Erdős & Rényi, 1960) with non-zero edge weights distributed as $J_{ij} \sim N\left(0, \frac{1}{k_i}\right)$ where $(k)$ is the average degree. The second system was sparse with $(k) = 2$ while the third system was dense with $(k) = 10$. We used Swendsen-Wang sampling (Swendsen & Wang, 1987) to generate either 500 or 1000 sequences and confirmed that there was no significant autocorrelation was visible in any of these samples.

Across all of the systems, we tested both $L_1$-regularized approaches for inference as well as variational approaches based on PSVI (section 3). The $L_1$ regularized approaches included pseudolikelihood, (PL) (Aurell & Ekeberg, 2012), minimum probability flow (MPF) (Sohl-Dickstein et al., 2011), and Persistent Contrastive Divergence (PCD) (Tieleman, 2008). The variational approaches were based on either a (collapsed) Laplacian prior, a centered Horseshoe prior, or a noncentered Horseshoe prior (Fadeout). In all of the variational problems, we used PSVI-3 with 100 persistent Markov chains and performed stochastic gradient descent using Adam (Kingma & Ba, 2014) with default momentum and a learning rate that linearly decayed from 0.01 to 0 over $5 \times 10^4$ iterations. We place a separate Horseshoe prior on the fields and couplings, giving (centered) the hierarchy

\[
s_h \sim C^+(0, 1) \quad s_{ij} \sim C^+(0, 1) \\
\sigma_i \sim C^+(0, s_h) \quad \sigma_{ij} \sim C^+(0, s_{ij}) \\
h_i \sim \mathcal{N}(0, \sigma_i^2) \quad J_{ij} \sim \mathcal{N}(0, \sigma_{ij}^2)
\]

where $C^+(0, 1)$ is the Half-Cauchy distribution. Figure 3 shows the factor graphs for both the centered and noncentered parameterizations of this hierarchy for a small model with 5 spins.

To give $L_1$ regularization a fair chance, we selected the hyperparameter $\lambda_1$ using 10-fold cross-validation over 10 logarithmically spaced values on the interval $[0.01, 1]$. We performed $L_1$ regularization of the deterministic objectives using optimizers from (Schmidt, 2010), and set the corresponding $L_1$ hyperparameters for PSVI-Laplacian and PCD-$L_1$ with the optimal cross-validated $\lambda$ under pseudo-likelihood.

We found that a noncentered Horseshoe prior (Fadeout) inferred with PSVI produced systematically lower reconstruction error of the couplings $J$ on both of the sparse systems (Table 2). This aligns with our understanding that no prior can be the best in all regimes of coupling, and that the important quality for an approximate method is to behave as intended. When we look at the mean-field approximation of the centered Horseshoe (Figure 4), we see that it fails to attain sparsity unlike the noncentered Horseshoe.

### 4.2. Potts model

The strongest advantages of hierarchical, sparsity-promoting priors are realized with group-sparsity. In the scale-mixtures of normals, this is accomplished by jointly integrating over a common mixing density, yielding a high-dimensional funnel that only amplifies the failures of mean-field inference of the centered parameters (section 2). We investigate the usefulness of group sparsity with the Potts model, a close cousin of the Ising model with spins that are categorical rather than binary. The factor graph is the same (Figure 3), except now $x \in \{1, \ldots, q\}^D$ and each $J_{ij}$ is a $q \times q$ matrix.

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

The Potts model has recently generated considerable excitement in biology, where it has been used to derive the 3D-structures of biological molecules solely from patterns

---

| Sample size | $4^3$ ferromagnet | Sparse SK spin glass | Dense SK spin glass |
|-------------|--------------------|----------------------|---------------------|
|             | 500    | 1000    | 500    | 1000    | 500    | 1000    |
| Naive mean-field | 0.086 | 0.053 | 0.916 | 0.301 | 1.504 | 0.3899 |
| PL + 10xCV $L_1$ | 0.031 | 0.024 | 0.037 | 0.034 | 0.184 | 0.1404 |
| MPF + 10xCV $L_1$ | 0.032 | 0.022 | 0.026 | 0.023 | 0.132 | 0.1410 |
| PCD-$3 + L_1$ (PL) | 0.031 | 0.022 | 0.047 | 0.044 | 0.155 | 0.1108 |
| PSVI-3, Laplacian (PL) | 0.037 | 0.027 | 0.049 | 0.044 | 0.228 | 0.1343 |
| PSVI-3, Horseshoe | 0.034 | 0.024 | 0.027 | 0.018 | 0.300 | 0.1872 |
| Fadeout PSVI-3, Horseshoe | **0.030** | **0.017** | **0.023** | **0.012** | 0.199 | 0.1357 |

$\text{Using } \lambda_{\text{Laplacean}} = \frac{1}{\pi \lambda_1}$
of correlated mutations in the sequences that encode them (Marks et al., 2011; Morcos et al., 2011). For proteins, each $J_{ij}$ is a $20 \times 20$ matrix encoding preferred combinations for amino acids. This is naturally a setting in which we wish to deploy a hierarchical, group-sparse prior over each $J_{ij}$ and will also be the motivation for our simulated experiment.

We constructed a synthetic Potts spin glass with a topology inspired by biological macromolecules. After forming a contact topology from a random packed polymer, we generated synthetic group-Student-t distributed sitewise biases and Gaussian distributed coupling parameters over an alphabet of size 20 to mirror the strong site-bias and weak-coupling regime of proteins. Since this system is highly frustrated, we thinned $2 \times 10^6$ sweeps of Gibbs sampling to 2000 sequences that exhibited no autocorrelation.

Given 400 of the 2000 synthetic sequences, we inferred $L_2$ and group $L_1$-regularized MAP estimates under a pseudolikelihood approximation with 5-fold cross validation to choose hyperparameters from 6 values in the range $\{0.3, 1.0, 3.0, 10.0, 30.0, 100.0\}$. We also ran PSVI-10 with 40 persistent Markov chains and 5000 iterations of stochastic gradient descent with Adam$^8$ (Kingma & Ba, 2014). Fadeout with a noncentered Horseshoe yielded more accurate (Figure 5, bottom left), less shrunk (Figure 5, bottom right) estimates of interactions, that were more predictive of the 1600 remaining test sequences (Table 3). We note that the current standards of the field are based on $L_2$ and Group $L_1$ regularized pseudolikelihood (Balakrishnan et al., 2011; Ekeberg et al., 2013).

5. Related work

We are aware of at least one previous work in which noncentered reparameterizations of models were used to improve variational inference. In the context of Generalized Linear Mixed Models, (Tan et al., 2013) performed a thorough comparative analysis of centered, noncentered, and partially noncentered parameterizations. Inspired by the interweaving approach of (Yu & Meng, 2011) they found that a variational distribution that interpolates between centered and noncentered parameterizations was the most flexible. While we committed to a noncentered parameterization for simplicity, this type of hybrid approach suggests a fruitful direction for future research.

---

We find this effective sample size to mirror natural protein families (unpublished)

Table 3. Average log-pseudolikelihood for test sequences.

| Method               | Test $-p(\mathcal{D}|h, J)$ | Runtime (s) |
|----------------------|-----------------------------|-------------|
| 5xCV PL $L_2$        | 67.3                        | 375         |
| 5xCV PL Group $L_1$  | 59.6                        | 303         |
| Noncentered Horseshoe| 54.2                        | 585         |

$^8\alpha = 0.01, \beta_1 = 0.9, \beta_2 = 0.999$, no decay
5.1. Black box variational inference

One strategy for improving variational inference is by introducing correlations in the variational distribution through transformations. (Rezende & Mohamed, 2015; Tran et al., 2015) demonstrate how it is possible to backpropagate through and thus learn transformations that capture the geometry of complex posteriors. Noncentered parameterizations of models may be complementary to these approaches by enabling more efficient representations of correlations between parameters and hyperparameters.

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 (Jaynes, 1957). We remind the reader that these methods for inference can be equivalently formulated as maximum likelihood in an exponential family (MacKay, 2003). Consequently, \( L_1 \) regularized-maximum entropy modeling (MaxEnt) is equivalent to the disfavored “integrate-out” approach\(^9\) to inference in hierarchical models explained in (MacKay, 1996). The result is that MAP estimates of maximum entropy models are prone to be dramatically biased with respect to the complexity of the distributions that they will recover if we cannot measure the constraints precisely (Macke et al., 2011).

How can we manage bias in MaxEnt? One hint arises in a seemingly simple problem: estimating the entropy of a discrete distribution. Here, tremendous progress was made by introducing hierarchical priors constructed to be less biased with respect to model complexity (Nemenman et al., 2001). This approach of using hierarchical priors proved to be similarly important for estimates of mutual information (Archer et al., 2013). Here we have introduced two tools to extend the reasoning behind these estimators of information quantities to derive estimators for the model parameters themselves. We learn not just a point estimate but a correlated variational distribution that captures uncertainty about the scale of the constraints in our system. One important area of future research may be to formulate hierarchical priors that are less biased with respect to model complexity, which will likely benefit from the large body of research on spin glasses and the entropic consequences of random distributions of couplings.

6. Conclusion

We introduced a variational approach designed to match the geometry of hierarchical, sparsity-promoting priors. We extended stochastic variational inference to work with undirected graphical models and found that, when combined, these two methods give substantially improved inferences of undirected graphical models.

### Acknowledgements

We thank David Duvenaud, Finale Doshi-Velez, Miriam Huntley, members of the Marks lab, and Chris Sander for helpful comments and discussions. JBI was supported by a NSF Graduate Research Fellowship 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.
- Balakrishnan, Sivaraman, Kamisetty, Hetunandan, Carbonell, Jaime G, Lee, Su-In, and Langmead, Christopher James. Learning generative models for protein fold families. *Proteins: Structure, Function, and Bioinformatics*, 79(4):1061–1078, 2011.
- 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. *arXiv preprint arXiv:1408.2047*, 2014.
- Ekeberg, Magnus, Lökvist, Cecilia, Lan, Yueheng, Weigt, Martin, and Aurell, Erik. Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Physical Review E*, 87(1):012707, 2013.
- Engelhardt, Barbara E and Adams, Ryan P. Bayesian structured sparsity from gaussian fields. *arXiv preprint arXiv:1407.2235*, 2014.
Bayesian Sparsity for Intractable Distributions

Erdős, Paul and Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci.*, 5:17–61, 1960.

Ferguson, Andrew L., Mann, Jaclyn K, Omarjee, Saleha, Ndung’u, Thumi, Walker, Bruce D, and Chakraborty, Arup K. Translating hiv sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity*, 38(3):606–617, 2013.

Gal, Yarin and Ghahramani, Zoubin. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *arXiv preprint arXiv:1506.02142*, 2015.

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.

Jaynes, Edwin T. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

Kingma, Diederik and Ba, Jimmy. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, Diederik P and Welling, Max. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kingma, Diederik P, Salimans, Tim, and Welling, Max. Variational dropout and the local reparameterization trick. *arXiv preprint arXiv:1506.02557*, 2015.

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

Kucukelbir, Alp, Ranganath, Rajesh, Gelman, Andrew, and Blei, David. Automatic variational inference in stan. In *Advances in Neural Information Processing Systems*, pp. 568–576, 2015.

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

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

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

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.

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, Heller, Katherine, and Ghahramani, Zoubin. Bayesian and l1 approaches to sparse unsupervised learning. *arXiv preprint arXiv:1106.1157*, 2011.

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. Mcmc for doubly-intractable distributions. *arXiv preprint arXiv:1206.6848*, 2012.

Neal, Radford M. Connectionist learning of belief networks. *Artificial intelligence*, 56(1):71–113, 1992.

Nemenman, Ilya, Shafee, Fariel, and Bialek, William. Entropy and inference, revisited. *arXiv preprint physics/0108025*, 2001.

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.

Polson, Nicholas G, Scott, James G, and Windle, Jesse. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349, 2013.

Ranganath, Rajesh, Gerrish, Sean, and Blei, David M. Black box variational inference. *arXiv preprint arXiv:1401.0118*, 2013.
Bayesian Sparsity for Intractable Distributions

Rasmussen, Carl Edward and Ghahramani, Zoubin. Occam’s razor. *Advances in neural information processing systems*, pp. 294–300, 2001.

Rezende, Danilo Jimenez and Mohamed, Shakir. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

Rezende, Danilo Jimenez, Mohamed, Shakir, and Wierstra, Daan. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 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.

Sherrington, David and Kirkpatrick, Scott. Solvable model of a spin-glass. *Physical review letters*, 35(26):1792, 1975.

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.

Tan, Linda SL, Nott, David J, et al. Variational inference for generalized linear mixed models using partially noncentered parametrizations. *Statistical Science*, 28(2):168–188, 2013.

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. Variational gaussian process. *arXiv preprint arXiv:1511.06499*, 2015.

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

Yu, Yaming and Meng, Xiao-Li. To center or not to center: That is not the question—an ancillarity–sufficiency interweaving strategy (asis) for boosting mcmc efficiency. *Journal of Computational and Graphical Statistics*, 20(3):531–570, 2011.