Variational Inference for Sparse and Undirected Models: Appendix

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

1. Appendix I: PVI algorithm

See Algorithm 1.

2. Appendix II: Experiments

2.1. Spin Models

We generated two synthetic systems. The first system was ferromagnetic (all $J \geq 0$) with 64 spins, where neighboring spins x_i, x_j have a nonzero interaction of $J_{ij} = 0.2$ if adjacent 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 system was a diluted Sherrington-Kirkpatrick spin glass (Sherrington & Kirkpatrick, 1975; Aurell & Ekeberg, 2012) with 100 spins. The couplings in this model were defined by Erdős-Renyi random graphs (Erdős & Rényi, 1960) with non-zero edge weights distributed as $J_{ij} \sim \mathcal{N}\left(0, \frac{1}{Np}\right)$ where Np is the average degree. We generated 5 random systems where the average degree was $Np = 100(0.02) = 2$. Across all of the systems, we used Swendsen-Wang sampling (Swendsen & Wang, 1987) to sample synthetic data and checked that the sampling was sufficient to eliminate autocorrelation in the data.

For inference, we tested both L_1-regularized deterministic approaches as well as a variational approach based on Persistent VI. 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). Additionally, we tested the proposed alternative regularization method of Pseudolikelihood Decimation (Decelle & Ricci-Tersenghi, 2014).

For L_1 regularized Pseudolikelihood and Minimum Probability Flow, we selected the hyperparameter λ using 10-fold cross-validation over 10 logarithmically spaced values on the interval $[0.01, 10]$. We performed L_1 regularization of the deterministic objectives using optimizers from (Schmidt, 2010), and chose the corresponding L_1 hyperparameter for PCD + L_1 based on the optimal cross-validated value of λ that was selected for L_1-regularized Pseudolikelihood.

For the hierarchical model inferred with Persisent VI, we placed a separate noncentered Horseshoe prior over the fields and couplings, in accodance with the (centered) hierarchy

$$\begin{aligned} s_h &\sim \mathrm{C}^+(0,1), & s_J &\sim \mathrm{C}^+(0,1), \\ \sigma_i &\sim \mathrm{C}^+(0,s_h), & \sigma_{ij} &\sim \mathrm{C}^+(0,s_J), \\ h_i &\sim \mathcal{N}(0,\sigma_i^2), & J_{ij} &\sim \mathcal{N}(0,\sigma_{ij}^2). \end{aligned}$$

where $\mathrm{C}^+(0,1)$ is the standard Half-Cauchy distribution. We then used PVI-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 5×10^4 iterations.

2.2. Synthetic Protein Data

We constructed a synthetic Potts spin glass with sparse interactions chosen to reflect an underlying 3D structure. After forming a contact topology from a random packed polymer, we generated synthetic group-Student-t distributed sitewise bias vectors $\mathbf{h}_i$ (each 20×1) and Gaussian distributed coupling matrices $\mathbf{J}_{ij}$ (each 20×20) to mirror the strong site-bias and weak-coupling regime of proteins. Since this system is highly frustrated, we thinned 2×10^6 sweeps of Gibbs sampling to 2000 sequences that exhibited no autocorrelation.

Given 400 of the 2000 synthetic sequences[1], 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 PVI-10 with 40 persistent Markov chains and 5000 iterations of stochastic gradient descent with Adam[2] (Kingma & Ba, 2014). We note that the current standards of the field are based on L_2

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

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

[2] $\alpha = 0.01, \beta_1 = 0.9, \beta_2 = 0.999$, no decay

Algorithm 1 Persistent Variational Inference (PVI-n) with Gaussian $q(\theta|\phi)$

```
Require: Model. Undirected $p(\mathbf{x}|\boldsymbol{\theta})$ defined by $k$ features $\{f_i(\mathbf{x})\}_{\mathbf{i=1}}^{\mathbf{k}}$ on $\mathbf{x} \in \{1, \ldots, q\}^D$
Require: Data. Expectations of the features $\{\mathbb{E}_{\mathcal{D}}[f_i(\mathbf{x})]\}_{i=1}^{k}$ and sample size $N$
Require: Prior. Prior gradient $\nabla \log P(\boldsymbol{\theta})$
Require: Number of Gibbs sweeps $n$, Markov Chains $M$, variational samples $Q$
Require: Initial variational parameters $\boldsymbol{\mu}_0, \log \boldsymbol{\sigma}_0$ (e.g. $\{0, -3\}$)
// Initialize parameters and Markov chains $\tilde{x}$
$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu}_0, \log \boldsymbol{\sigma} \leftarrow \log \boldsymbol{\sigma}_0$
$\tilde{\mathbf{x}}^{(1:M)} \leftarrow \mathrm{RandInt}(1, q)$
$t \leftarrow 0$
while not converged do
  // Estimate $\nabla\mathrm{ELBO}$ with $Q$ samples from the variational distribution
  $\nabla_{\boldsymbol{\mu}}\mathcal{L} \leftarrow 0$, $\nabla_{\log \boldsymbol{\sigma}}\mathcal{L} \leftarrow 0$
  for $s = 1 \ldots Q$ do
    $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I_{|\boldsymbol{\mu}|})$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$
    // Estimate model-dependent expectations $\mathbf{E}$, where $E_i = \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})}[f_i(\mathbf{x})]$
    $\mathbf{E} \leftarrow \mathbf{0}$
    for $m = 1 \ldots M$ do
      for $j = 1 \ldots n$ do
        $\tilde{\mathbf{x}}^{(m)} \leftarrow \mathrm{GibbsSweep}(p(\mathbf{x}|\boldsymbol{\theta}), \tilde{\mathbf{x}}^{(m)})$
        $\mathbf{E} \leftarrow \mathbf{E} + \frac{1}{Mn}\{f_i(\tilde{\mathbf{x}}^{(m)})\}_{i=1}^{k}$
      end for
    end for
    // Compute stochastic gradient
    $\mathbf{G} \leftarrow N(\mathbb{E}_{\mathcal{D}}[f_i(\mathbf{x})] - \mathbf{E}) + \nabla \log P(\boldsymbol{\theta})$
    $\nabla_{\boldsymbol{\mu}}\mathcal{L} \leftarrow \nabla_{\boldsymbol{\mu}}\mathcal{L} + \frac{1}{Q}\mathbf{G}$
    $\nabla_{\log \boldsymbol{\sigma}}\mathcal{L} \leftarrow \nabla_{\log \boldsymbol{\sigma}}\mathcal{L} + \frac{1}{Q}\left(\mathbf{G} \odot (\boldsymbol{\theta} - \boldsymbol{\mu}) + 1\right)$
  end for
  // Update parameters with Robbins-Monro sequence $\{\rho_t\}$
  $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} + \rho_t \nabla_{\boldsymbol{\mu}}\mathcal{L}$
  $\log \boldsymbol{\sigma} \leftarrow \log \boldsymbol{\sigma} + \rho_t \nabla_{\log \boldsymbol{\sigma}}$
  $t \leftarrow t + 1$
end while
```

and Group L_1 regularized Pseudolikelihood (Balakrishnan et al., 2011; Ekeberg et al., 2013).

2.3. Real Protein Data

2.3.1. Sample reweighting

Natural protein sequences share a common evolutionary history that introduces significant redundancy and correlation between related sequences. Treating them as independent data is biased by both the overrepresentation of certain sequences due to the evolutionary process (phylogeny) or the human sampling process (biased sequencing of particular species). Thus, we follow a standard practice of correcting the overrepresentation of sequences by a sample-reweighting approach (Ekeberg et al., 2013).

Sequence reweighting. If we were to treat all data as independent, then every sample would receive unit weight in the log likelihood. To correct for the over and underrepresentation of certain sequences, we estimate relative sequence weights using a common inverse neighborhood density based approach from the field (Ekeberg et al., 2013). We set the relative weight of each sequence proportional to the inverse number of neighboring sequences that differ by a normalized Hamming distance of less than θ. We use the established value of $\theta = 0.2$.

Effective sample size estimation. We propose a new definition for an effective sample size N_{eff} of correlated discrete data and derive an algorithm for estimating it from count data. The estimator is based on the assumption that in limited data regimes for sparsely coupled systems, the sample Mutual Information between random variables is dominated by random, coincidental correlations rather than actual correlations due to underlying interactions. This is consistent with classic results on the bias of information quantities in limited data regimes known as "Miller Maddow bias" (Miller, 1955; Paninski, 2003). If we can define a null model for how such coincidental correlations would arise for a given random sample of size N, then we define N_{eff} as the sample size that matches the expected null MI to the observed MI.

$$\mathbb{E}_{i,j}\left[\text{MI}_{null}|N_{eff}\right] = \mathbb{E}_{i,j}\left[\text{MI}_{data}\right] \tag{1}$$

The expectation on the right is given by the average sample Mutual Information in the data, while the expectation on the left will be specific to a null model for Mutual Information $\mathbb{E}_{i,j}\left[MI_{null}|N\right]$. Given a noisy estimator for $\mathbb{E}_{i,j}\left[MI_{null}|N_{eff}\right]$, we solve for N_{eff} by matching the expectations with Robbins-Monro stochastic optimization.

To define the null model of mutual information $\mathbb{E}_{i,j}\left[MI_{null}|N\right]$ we treat every variable as independent categorical counts that were drawn from a Dirichlet-Multinomial hierarchy with a log-uniform hyperprior over the (symmetric) concentration parameter α.

Given observed frequencies $\mathbf{f}_i$ and $\mathbf{f}_j$ for letters x_i and x_j together with a candidate sample size N, we (i) use Bayes' theorem to sample underlying distributions $\mathbf{p}_i$, $\mathbf{p}_j$ that produced the observed frequencies, (ii) generate N samples from the null joint distribution $\mathbf{p}_i\mathbf{p}_j^T$, and (iii) compute the sample Mutual Information of this synthetic count data (Algorithm 2).

We also experimented with using both MAP and posterior mean estimators as plugin approximations $\hat{\mathbf{p}}_i$, $\hat{\mathbf{p}}_j$ for the latent distributions, but found that each of these were biased estimators of the true sample size in simulation. Posterior mode estimates generally underestimated the null entropy ($\hat{\mathbf{p}}_i$ too rough) while the posterior mean overestimated the entropy ($\hat{\mathbf{p}}_i$ too smooth). It seems reasonable that this would be the behavior of point estimates that do not account for the uncertainty in the null distributions that is signaled by the roughness of the frequency data.

We note that this estimator will become invalid as the data become strong, since the assumption that Mutual Information is dominated by sampling noise will break down. However, for the real protein data that we examined, we found that this approach for effective sample size correction was critical for Bayesian methods such as Peristent VI to be able to set the top level hyperparameters (the sparsity levels) from the data.

Algorithm 2 Sample the null mutual information as a function of sample size $\mathbb{E}_{i,j}\left[MI_{null}|N\right]$

```
Require: Sample size $N$
Require: Observed frequencies $\mathbf{f}_i$, $\mathbf{f}_j$
Sample positions $i \in [L]$, $j \in [L] \setminus i$
Set count data $\mathbf{C}_i \leftarrow N\mathbf{f}_i$, $\mathbf{C}_j \leftarrow N\mathbf{f}_j$
Sample concentration parameter $\alpha_i|\mathbf{C}_i$, $\alpha_j|\mathbf{C}_j$ with numerical CDF
Sample null distributions $\mathbf{p}_i|\mathbf{C}_i, \alpha_i$, $\mathbf{p}_j|\mathbf{C}_j, \alpha_j$ from Dirichlet
Sample joint count data $\mathbf{M}(x_i, x_j)$ from categorical joint distribution $\mathbf{p}_i\mathbf{p}_j^T$
Compute sample frequencies $\mathbf{f} = \frac{1}{N}\mathbf{M}(x_i, x_j)$, $\mathbf{f}_i = \frac{1}{N}\sum_{x_j}\mathbf{M}(x_i, x_j)$, $\mathbf{f}_j = \frac{1}{N}\sum_{x_i}\mathbf{M}(x_i, x_j)$
Compute sample Mutual Information $MI = \sum_{x_i,x_j}\mathbf{f}(x_i, x_j)\log\frac{\mathbf{f}(x_i,x_j)}{\mathbf{f}_i(x_i)\mathbf{f}_j(x_j)}$
```

2.3.2. Inference and results

Alignment Our sequence alignment was based on the Pfam 27.0 family PF00018, which we subsequently processed to remove all sequences with more than 25% gaps.

Indels Natural sequences contain insertions and deletions that are coded by 'gaps' in alignments. We treated these as a 21st character (in addition to amino acids) and fit a $q = 21$ state Potts model. We acknowledge that, while this may be standard practice in the field, it is a strong independence approximation because all of the gaps in deletions are perfectly correlated.

Inference We used 10,000 iterations of PVI-10 with 10 variational samples per iteration and 40 persistent Gibbs chains.

Comparison to 3D structure We collected about 260 3D structures of SH3 domains referenced on PF00018 (Pfam 27.0) and computed minimum atom distances between all positions in the Pfam alignment. For each pair i, j, we used the median of distances across all structures to summarize the "typical" minimum atom distance between i and j.

References

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

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.

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.

Ekeberg, Magnus, Lövkvist, 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.

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

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

Miller, George A. Note on the bias of information estimates. *Information theory in psychology: Problems and methods*, 2(95): 100, 1955.

Paninski, Liam. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003.

Schmidt, Mark. *Graphical model structure learning with l1-regularization*. PhD thesis, UNIVERSITY OF BRITISH COLUMBIA (Vancouver, 2010.

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

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.

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

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.